No-code machine learning in radiology: implementation and validation of a platform that allows clinicians to train their own models

Daniel C. Elton[a,*], Giridhar Dasegowda[b,*], James Y. Sato[a,*], Emiliano G. Frias[b], Christopher P. Bridge[a], Artem B. Mamonov[a], Mark Walters[a], Martynas Ziemelis[a], Thomas J. Schultz[a], Bernardo C. Bizzo[a], Keith J. Dreyer[a] and Mannudeep K. Kalra[b]

[a]*Mass General Brigham AI, Boston 02108, MA, USA*
[b]*Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston 02114, MA, USA*

ARTICLE INFO

Keywords:
AutoML
no-code ML
machine learning
radiology
DevOps

ABSTRACT

Machine learning models can assist clinicians and researchers in many tasks within radiology such as diagnosis, triage, segmentation/measurement, and quality assurance. To better leverage machine learning we have developed a platform that allows users to label data and train models without requiring any programming knowledge. The technology stack consists of a TypeScript web application running on .NET for user interaction, Python, PyTorch, and MONAI for machine learning, DICOM WADO-RS to retrieve data from clinical systems, and Docker for model management. As a first trial of the system, researchers used it to train a model for clavicle fracture detection as part of an IRB-approved retrospective study. The researchers labeled 4,135 clavicle radiographs from 2,039 patients across 13 sites. The platform automatically split the data into training, validation, and test sets and trained a model until the validation loss plateaued. The system then returned a receiver operating characteristic curve, AUC, F1, and other metrics. The resulting model identifies clavicle fractures with 90% sensitivity, 87% specificity, and 88% accuracy with an AUC of 0.95. This model performance is equivalent to or better than similar models reported in the literature. More recently, our system was used to train a model to identify if ultrasound frames that contain personally identifiable information (PII). After validation, the model was used to help de-identify a large dataset that was to be used for research. This first-of-its-kind system streamlines model development and deployment and opens up an exciting new pathway for the use of AI within healthcare.

1. Introduction

The potential use-cases where artificial intelligence (AI) models can bring value to healthcare continues to grow. As of October 2023, the United States Food and Drug Administration (FDA) has has approved 692 AI algorithms.[1] However, this pace of AI adoption within healthcare has been slow.[2] One cause of this is that hospitals are cost strapped and still reeling from pandemic, and a tiny fraction of FDA-approved AI devices are covered by insurance.[3] A perhaps even bigger and more serious issue is that external validations of AI algorithms often show substantial drops in performance compared to what was originally reported in the FDA submission.[4, 5, 6] This degradation is understood to be caused by "distribution shift".[7] Distribution shift may be caused by differences in scanner type, image acquisition protocols, or patient demographics. While multiple strategies and techniques have been proposed to address the distribution shift issue, the problem remains largely unsolved.[8]

Developing models in-house is an attractive proposition for hospitals as they can use their own datasets for training, which helps ameliorate the issue of distribution shift. The current push towards internalizing AI model development is coming in time of maturing ML platforms. Advances in graphical/tensor processing units (GPUs/TPUs) alongside improved cloud computing infrastructure have reduced the cost needed to train models. Mature software libraries such as the Medical Open Network for Artificial Intelligence (MONAI) and PyTorch make it easier than ever to train machine learning (ML) models on medical imaging data. Even so, code has to be written and choices have to be made regarding data preprocessing, data augmentation, neural network model architecture, model hyperparameter settings, batch size, stopping criteria, and validation criteria. Thus, the process of training ML models still requires the involvement of specialized personnel with programming expertise such as postdoctoral researchers, data scientists, or machine learning engineers. With industry salaries increasing relative to what healthcare systems can pay, finding staff to fill such roles is becoming more and more of a challenge.[2]

Several no-code platforms for training ML models exist from companies like Amazon, Apple, Calrifai, Google, and Microsoft.[9] Some of these platforms have been tested on publicly available medical imaging datasets.[9] However, to our knowledge none of these platforms support DICOM natively and they suffer from a lack of integration with hospital IT infrastructure. To use such services images must be pulled down, converted to image formats, and uploaded. Even after a model has been developed it then has to be packaged and integrated with clinical systems. It has been our observation that the entire process of training, validating,

*Equal contribution, sharing first author status.
delton@mgh.harvard.edu (Daniel C. Elton); mkalra@mgh.harvard.edu (Mannudeep K. Kalra)
ORCID(s): 0000-0002-8323-4616 (Daniel C. Elton); 0000-0001-9246-8265 (Giridhar Dasegowda); 0000-0003-1997-676X (James Y. Sato); 0000-0001-8724-0143 (Emiliano G. Frias); 0000-0002-2242-351X (Christopher P. Bridge); 0000-0002-9686-6751 (Bernardo C. Bizzo); 0000-0001-9938-7476 (Mannudeep K. Kalra)

Figure 1: Platform overview.

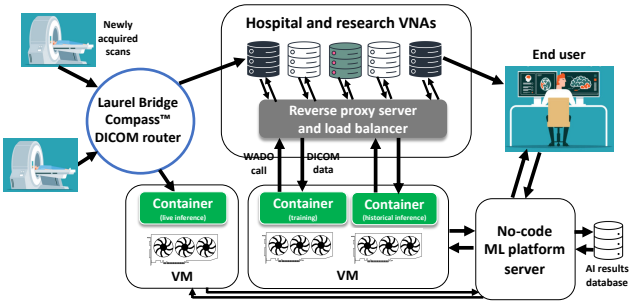


Figure 2: The interface for data labeling.

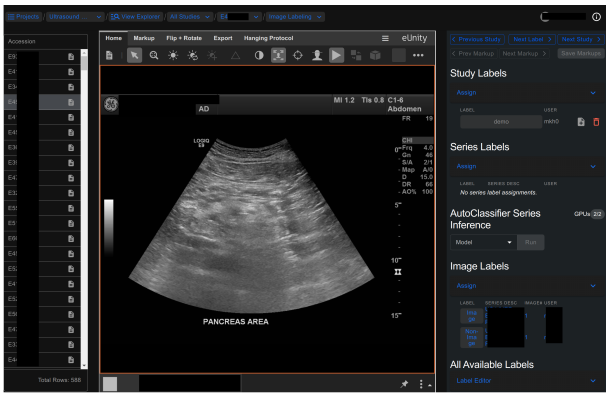
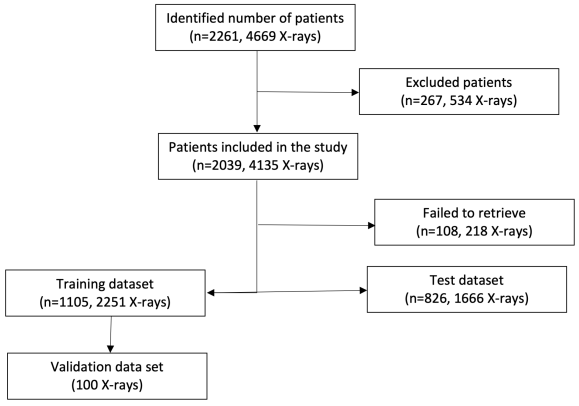


Figure 3: Flowchart of the inclusion and exclusion criteria for the training, validation, and test datasets for the pilot study.



and deploying a ML model within healthcare is very time-consuming, involving coordination between many people and a complicated juggling of code and data across multiple systems.[3]. Lack of a standard process means that there are risks to patient data privacy and data security.

Research in autonomous machine learning (AutoML) has previously shown that many of the steps mentioned above can be automated.[10] No-code ML goes a step further than AutoML by providing a platform that empowers individual users with the ability to independently and iteratively train, validate, and deploy models without any programming required.[11] In this paper we present a novel no-code ML platform which allows physicians to create their own ML models for radiology. While our focus is on medical imaging (DICOM) data, the overall framework could be extended to other types of healthcare data such as electronic health records.

There are many reasons clinicians may want to develop models. Previously several of us trained an in-house model to identify improperly taken chest radiographs.[12] An early version of the ML software described below was used to train a classification model to classify mammography image view (CC vs MLO) and laterality (L vs R) to check for mistakes with the DICOM orientation tag (four classes total). A DenseNet121 model (8 M parameters) was trained on 1,789 mammography images downsampled to 128x128 pixels. The model achieved an average accuracy of 99.52% on a test set of 665 images. Based on that initial success with that codebase we then decided to develop it further into a no-code system so that users without coding experience could train similar models in the future.

The aforementioned model for mammography view classification is an example of a model for quality assurance. Models can also be implemented to assist in the areas of time-consuming, mundane tasks such as worklist prioritization/triage, the automation of time-consuming tasks such as the segmentation, and even to assist with diagnosis. Since our platform was developed in-house and the resulting models are only used internally, FDA approval is not required for such models to be used clinically. The lack of FDA approval for these models underscores the need for careful internal validation and oversight. Recently Panch et al. have argued that internal development of AI models is a promising pathway for AI use in healthcare.[13] As Panch et al. explain, hospitals have a strong obligation to set up their own validation and monitoring infrastructure regardless of whether models are developed internally or are FDA approved.

We tested our platform by providing the system to researchers at Massachusetts General Hospital. They ran a pilot study which used the system to train a clavicle fracture detection model for X-ray radiographs. Next we utilized the system to train a model that can identify ultrasound image frames that have baked-in personally identifiable information (PII). The resulting model was successfully used to remove images with PII in order to generate a fully de-identified dataset for research purposes.

2. System overview

Figure 1 shows an overview of our system. To ensure data privacy, the entire system is implemented on secured platforms within the hospital firewall. The data from the clinical vendor neutral archives (VNAs) may be de-identified and transferred to a research VNA. Alternatively images may be pulled directly from the clinical VNAs. During training and inference imaging data is not saved to disk. A load balancer insures that clinical operations are not effected by any data pulls.

Table 1
Summary of some previously published fracture detection models. (NS = Not specified)

Reference	Fracture	Sensitivity	Specificity	AUC	95% CI
Current study (patient level)	Clavicle	91%	94%	0.97	0.96-0.98
Guermazi et al., 2022 [14]	Clavicle	84%	83%	0.90	NS
Jones et al., 2020 [15]	Shoulder & Clavicle	90%	91%	0.96	0.79-0.96
Ma et al., 2021 [16]	20 fracture types	85%	97%	NS	NS
Dupuis et al., 2022 [17]	All fractures	96%	91%	NS	NS
Hayashi et al., 2022 [18]	All fractures	91%	90%	0.93	0.88-0.97
Ashkani-Esfahani et al., 2022 [19]	Ankle	99%	99%	0.99	NS
Raisuddin et al., 2021 [20]	Wrist	NS	NS	0.98	0.97-0.99
Yoon et al., 2021 [21]	Scaphoid	87%	92%	0.96	NS
Murata et al., 2020 [22]	Spine	85%	87%	0.91	0.96-1.00
Mawatari et al., 2020 [23]	hip	88%	72%	0.90	NS
Blüthgen et al., 2020 [24]	Distal radius	64%	60%	0.80	NS
Cheng et al., 2020 [25]	Hip fracture	98%	84%	NS	NS
Chung et al., 2018 [26]	Proximal humerus	99%	97%	0.97	0.96-0.97

The no-code platform graphical user interface (GUI) was written in TypeScript with a .NET core backend. Currently only classification models are supported. Users can use Nuance's mPower software to search for studies to use and then import a list of studies into the system. This includes the ability to search by age and sex. A special GUI page for for data labeling was developed using the eUNITY viewing software to view images.

Users can also view the radiology text reports associated with each study. Users can provide labels at the both the study and image level. Once a training dataset has been created, then a training run can be started. A special interface was developed for training and model management. Training occurs in Docker containers which are run on dedicated virtual machines (VMs) with GPUs (NVIDIA A100s or V100s).

The Docker container contains code which uses the PyTorch ML framework and some functions from MONAI. For inputs, the container takes in a configuration file in Java Script Object Notation (JSON) format and a .csv file with labels and optional custom splitting (into training, validation, and test datasets). During training the container queries the hospital VNAs using Web Access to DICOM Objects (WADO) to retrieve images. The images are fed into the training on the fly. The trained model is saved in a local output directory alongside training logs and the original configuration .json file.

Training options can be modified by editing the JSON config file (in the future a GUI for this may be provided). While nearly all options in the system are configurable, the default settings were carefully chosen so they should be suitable for most applications.

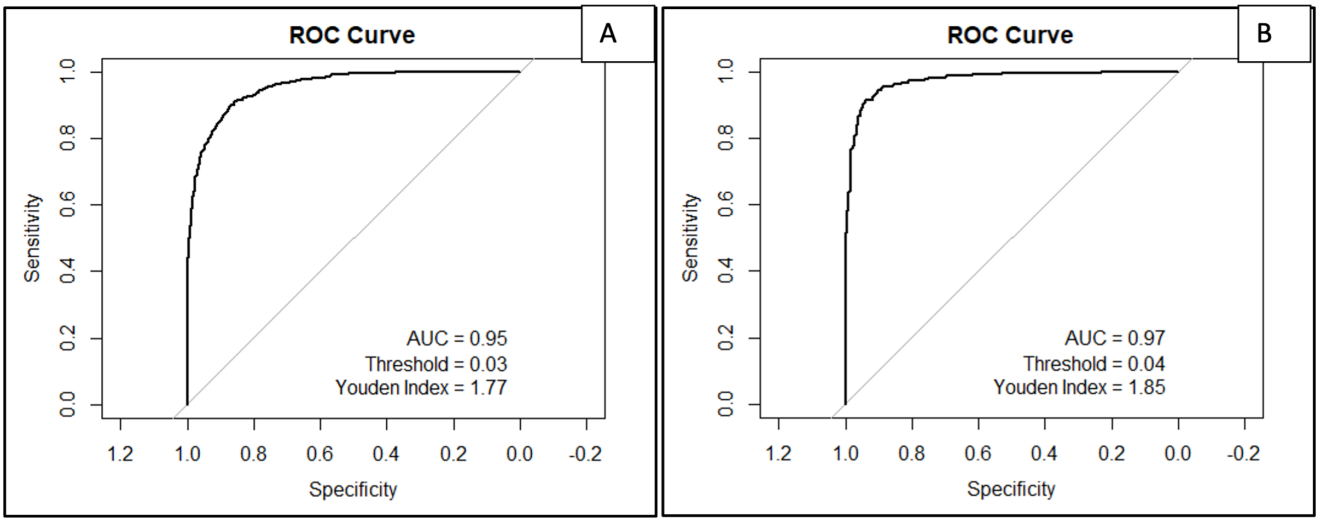
The default model is a pretrained DenseNet201 architecture (20.1 M parameters), but the system supports many models which are available from MONAI, such as EfficientNet models, which use less compute.[27, 28] The user must specify the number of classes and whether they are doing exclusive classification, non-exclusive classification, or regression.

By default both dropconnect[29] and dropout are used with a conservative rate of 0.2 and the following types of data augmentation are implemented - random rotations (-10 to +10 degrees), random zooms/crops (0.9 - 1.25x), random flipping, and random elastic deformations. The amount of elastic deformation is kept small as large deformations may not be suitable for all applications. Several options for data normalization are provided (clipping, rescaling, etc). Models are trained using the Adamax optimizer algorithm,[30] with weighted random sampling to improve performance in the case of unbalanced training labels. The default batch size is 6 and the default learning rate is set low at 0.0005 to avoid training instabilities. We use the well-known "reduce on plateau" learning rate schedule, which reduces the learning rate by a factor of 0.5 when the validation loss plateaus for 5 steps.

A key challenge in no-code ML is determining the criteria for stopping training. If the model is trained too long it will overfit, whereas if training is stopped too early the model's accuracy may not have reached the best possible value. Since the validation loss can be very noisy we smooth the validation loss using a moving average of width 5 iterations. We stop the training when the smoothed validation loss no longer decreases for 1500 iterations. The maximum number of epochs is set to 400.

The system contains several options for train-validation-test splitting. The default is to split the data by *PatientID*. The splitting can also be done by site (hospital) which is stored in the *IssuerOfPatientID* DICOM tag. Thus, data sources in the test set can be different from those in the training set, providing a form of "external" validation. Once a model has been trained, it is run on the test set. The system returns validation statistics, ROC curves, and a confusion matrix (see fig. 3.3. The Youden index is used to determine the optimal threshold for classification during inference.[31]

Figure 4: ROC curves and AUCs for X-ray level (A) and patient level (B) clavicle fracture detection.



After a model has been trained and tested, new containers for inference may be spun up. Inference can be done on historical data listed in a .csv or live-routing may be configured. During live-routing the newly acquired scans that match set critiera are routed to the container. Currently live routing must be manually configured using a custom C# script implemented in our Laurel Bridge Compass™ DICOM router. Any potential clinical use of models developed with our framework requires approval by our AI governance board and extensive period of validation with live data.

3. Pilot user study

3.1. Study design

Our institution review board (IRB) waived the written consent requirement for our retrospective, Health Insurance Portability and Accountability Act (HIPAA) compliant pilot study. The study details provided below are presented in conformance with the Checklist for Artificial Intelligence in Medical Imaging (CLAIM).[32]

The study dataset was comprised of clavicle radiographs from 2,039 adult patients (age > 18 years) sourced from 13 sites within our network including 10 hospitals, one urgent care center, and two quaternary care centers. To identify eligible radiographs for our study we used Nuance mPower Clinical Analytics Search (Microsoft Inc.), a cloud-based, commercial radiology reports search engine that integrates radiology reports data from the sites included in our study. The search key terms for identifying consecutive radiology reports and radiographs with and without clavicle fractures were “acute fracture” OR “no fracture” OR “displaced fracture” AND “clavicle X-ray”. The search was limited to clavicle radiographs performed between January 2016 – December 2022. Both right and left clavicle radiographs were included in the study. A post-doctoral radiology research fellow (2 years of experience) reviewed all radiology reports and radiographs to exclude clavicle radiographs with incomplete anatomic coverage of clavicles, metal-related artifact, prosthesis, or evidence of open reduction with internal fixation ($N = 267$). Non-clavicle radiographs (such as shoulder and chest radiographs) with or without clavicle fractures were not included in the study. A flowchart of the inclusion and exclusion with training and test data has been represented in fig. 1.

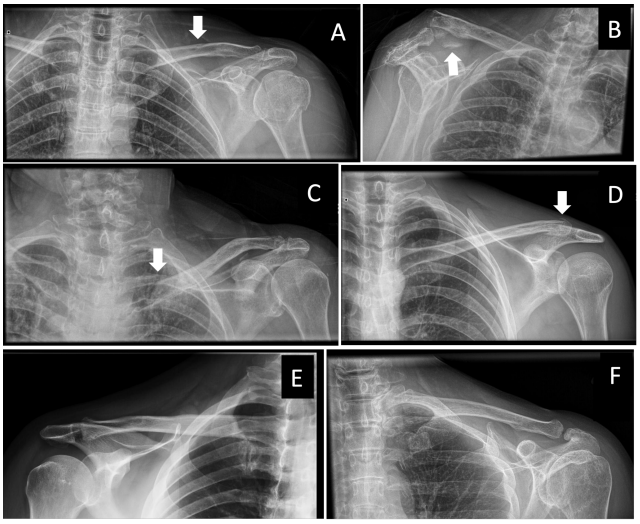
We exported radiology reports of the eligible radiographs from the radiology report search engine with the following data elements: radiology findings text, radiology impression text, date of examination, name of the radiographic procedure, site of radiographic acquisition, as well as patients’ age and gender. We reviewed the radiology reports and recorded the details of the presence of fracture to establish the ground truth. For this initial pilot study the data labeling was not done with the platform’s GUI. Instead, the data labels were uploaded via a .csv file into the system.

To avoid selection bias, all consecutive clavicle radiographs were included regardless of patients’ age, race, and sex as well as radiographic equipment and site of acquisition. Although a power analysis was not performed to determine adequate sample size for our test set, our sample size was larger than most prior publications in this domain. Clavicle radiographs from three sites were marked for inclusion in an external test dataset. The remaining radiographs were used for training and validation.

3.2. Dataset and model

A total of 2,151 clavicle radiographs were used in the training dataset, and 100 radiographs from were automatically split to comprise the validation dataset. The default settings described in section 2 were used. All radiographs were resampled to a size of 512x512 and normalized to have a mean of zero and standard deviation of one. After model training and testing the platform returned statistics for the performance of the AI model. The mean age (± standard deviation) of 2039 adult patients included in our study was 52 years. There were 1022 female patients and 1017 males. Site-wise distribution of patients was Site 1

Figure 5: Model performance examples on frontal projection radiographs of clavicles with AI-detected (true positive: A, B), AI-missed (false negative: C, D), and AI-false positive (E, F) clavicle fractures

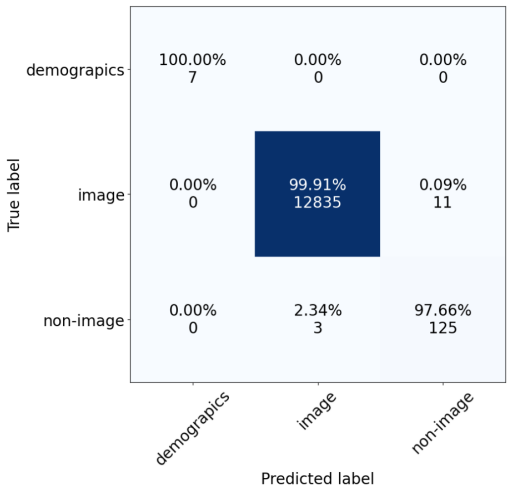


(N=621 patients), Site 2 (N=651), Site 3 (N=166), Site 4 (N=40), Site 5 (N=223), Site 6 (N=167), Site 7 (N=43), Site 8 (N=11), Site 9 (N=6), Site 10 (N=36), Site 11 (N=37), Site 12 (N=7), and Site 13 (N=31). Of the 2039 x-rays, there were 1225 radiographs of the right clavicle and 814 left clavicle radiographs. Most patients were either outpatients (N=1066) or in the emergency department (N=828), with only 145 inpatients. Radiographs of 108 patients who could not be automatically deidentified were excluded. 1,666 radiographs were used for model testing (772 radiographs with clavicle fracture and 894 radiographs without clavicle fracture) from two sites that did not contribute to the training datasets. The test set included 826 patients total.

3.3. Results

The model trained until it achieved 89% accuracy on the validation set. On the test set the model classified individual X-ray images with clavicle fractures with 90% sensitivity, 87% specificity, 88% accuracy, 0.86 F1, and an area under the reciever operating characteristic curve (AUC) of 0.95. The model had 91% sensitivity, 94% specificity, 93% accuracy, 0.91 F1, and an AUC of 0.97 (95% CI 0.96-0.98) when classifying at the patient level (since some patients had multiple images in their study, we average the softmax predictions across the images for each patient and use that as a 'patient level' classification). Receiver operating characteristic (ROC) curves are shown in fig. 3. Comparison with some previous models for fracture detection is provided in table 1. Additional statistical analyses were performed with SPSS. There was no significant difference in model performance in male or female patients and among patients in different locations at the time of their radiography ($p > 0.05$). For analysis, AI outputs were classified as true positive, true negative, false positive, and false negative. False positive findings were noted in X-rays when the clavicle had degenerative changes, skin folds, artifacts, and foreign bodies such as post catheter overlying the clavicle. False negative findings were present in both displaced and non-displaced clavicle fractures. Examples of true positive, true negative, false positive, and false negative outputs are shown in fig. 3.2.

Figure 6: Confusion matrix on a test set of ultrasound images.



4. First use-case: identifying PII in ultrasound images

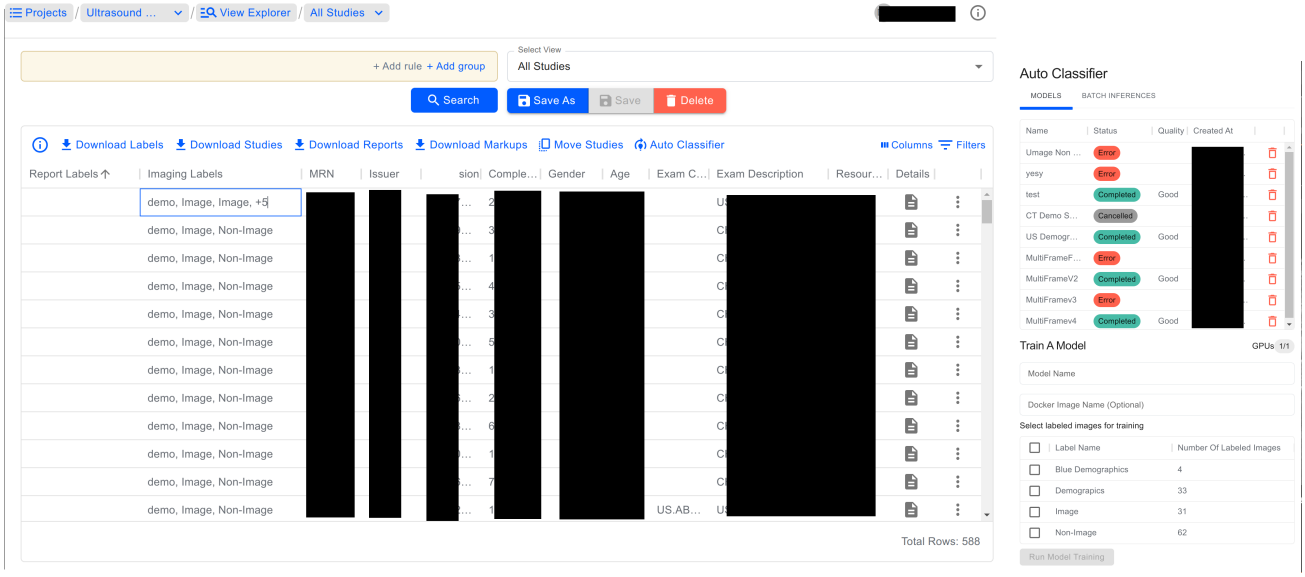
The first use-case of the fully developed system was to train a model to classify ultrasound images as to whether they were "normal" images, "non-images", or images with burned in PII. Examples of these image types are shown in figure 4. Ultrasound images with burned in PII are rare but must be removed when creating fully de-identified datatsets for research purposes. Previously this time-consuming process was done manually.

To properly test the system, the model development was done by a program manager with no technical coding knowledge. Using the system a training set had 421 images (including multiframe) were labeled. There were 62 "non-image" images, and 37 images with PII. While these numbers are small, the classification task is relatively easy due to the large differences between the images types. The "normal" images were DICOM MultiFrame images (sometimes called "Enhanced DICOM"). The multiframe images contained around 20-30 frames each. Since the system trains on all frames, the total number of "normal" images was around 9,500.

The test set consisted of 12,981 ultrasound instances, many of which were multiframe. Manually labeling the test set with our GUI was found to be time consuming, so it was done by downloading all of the data and viewing image thumbnails. This enabled the dataset to be labeled within hours rather than days. The confusion matrix is shown in fig. 3.3. The overall accuracy was 0.9989. While the number of demographics pages in the test set is small, these initial

Figure 7: Examples of a standard ultrasound image frame (left), an ultrasound image which is actually a screenshot with PII (middle), and an ultrasound image that is a "non-image" settings page. (PII has been redacted in black).

Figure 8: Screenshots of the data list (left) and ML model management pane (right).



results were encouraging enough that the model was used to help de-identify a separate dataset for research purposes.

5. Discussion

Our study demonstrates that an internally developed no-code ML platform give non-coding clinicians and personnel the ability to develop AI classification models in a fast and efficient manner. The performance of the clavicle fracture classification model trained with our system is similar or better than models reported in the literature (see table 1.[33, 34] For example, Guermazi et al. reported their AI model for identifying shoulder and clavicle fractures had a performance of 84% sensitivity, 83% specificity, and 0.90 AUC (95% CI 0.79-0.96).[14] Jones et al. reported on a deep learning system for identifying clavicle fractures, using an ensemble of 10 convolutional networks, which obtained 90% sensitivity, 91% specificity, and an AUC of 0.96.[15]

The clavicle detection model pilot study has some notable limitations. First, there was an asymmetric distribution of radiographs across different institutions and between the radiographs with and without clavicle fractures. Second, all clavicle radiographs belonged to a common healthcare system in the same geographic location in the Northeast part of the United States. Third, we did not assess variations in the model performance across patients' size, racial, or ethnic groups. Fourth, as stated above, the successful creation of a classification model for clavicle fractures does not imply that the same no-code ML platform will be successful at building sophisticated models or those for cross-sectional imaging modalities. Furthermore, model performance on different radiography techniques (i.e. computerized versus digital radiography) was not assessed; however, considering variations in equipment across the 14 hospitals, the model was trained and tested on radiographs from several different vendors. Likewise, due to the exclusion of radiographs with incomplete anatomic coverage, artifacts, and prior open reduction and internal fixation, we cannot comment on the model performance on such radiographs. Finally, we also did not assess the model's performance in non-displaced versus displaced fractures and for patients with chronic or non-healed clavicle fractures. Despite these limitations, we believe this model could improve the diagnostic accuracy of fracture detection, especially in clavicle fractures without

displacement that often pose a challenge to interpreting physicians.

There are several avenues open to improving this system. One thing that still needs to be implemented is better support for cross validation. Currently cross-validation must be done by providing a "fold" parameter in the config file. Then the training runs for each fold have to be initiated manually, and the metrics for each fold have to be averaged manually as well. This process could be automated on the web platform side. Next, the system could be expanded to work with 3D modalities such as CT or MRI. Since we have already implemented training on multiframe images, this should be easy to implement. A more substantial upgrade would be to extend the system beyond classification to enable segmentation and bounding box detection models to be created. We have already extensively tested the Redbrick AI (www.redbrickai.com) data labeling platform and integrated it with our VNA. Redbrick AI provides an application programming interface (API) which allows cohorts to be sent the platform. On the platform users can create segmentation, polygon, and bounding box labels which can then be retrieved via API. The image labeling platform could also be updated to make labeling faster and easier in some situations, for instance by implementing thumbnail views. Finally, in the future this system could be further developed to allow users to fine-tune multimodal transformer foundation models similar to GPT-4.

In conclusion, our no-code ML platform simplifies and expedites the development of AI models for medical imaging. It is our goal to make this system available to scientists, physicians, and engineers within our healthcare system so they can train models to assist in their everyday work.

References

[1] FDA. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices, 2023.

[2] Matthew J. Leming, Esther E. Bron, Rose Bruffaerts, Yangming Ou, Juan Eugenio Iglesias, Randy L. Gollub, and Hyungsoon Im. Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting. *npj Digital Medicine*, 6(1), July 2023. ISSN 2398-6352. URL http://dx.doi.org/10.1038/s41746-023-00868-x.

[3] Bernardo C. Bizzo, Giridhar Dasegowda, Christopher Bridge, Benjamin Miller, James M. Hillis, Mannudeep K. Kalra, Kimberly Durniak, Markus Stout, Thomas Schultz, Tarik Alkasab, and Keith J. Dreyer. Addressing the challenges of implementing artificial intelligence tools in clinical practice: Principles from experience. *Journal of the American College of Radiology*, 20(3):352–360, March 2023. ISSN 1546-1440. URL http://dx.doi.org/10.1016/j.jacr.2023.01.002.

[4] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, June 2021. ISSN 2168-6106. URL http://dx.doi.org/10.1001/jamainternmed.2021.2626.

[5] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20. ACM, April 2020. URL http://dx.doi.org/10.1145/3313831.3376718.

[6] Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32), August 2022. ISSN 2375-2548. URL http://dx.doi.org/10.1126/sciadv.abq6147.

[7] Vallijah Subasri, Amrit Krishnan, Azra Dhalla, Deval Pandya, David Malkin, Fahad Razak, Amol A. Verma, Anna Goldenberg, and Elham Dolatabadi. Diagnosing and remediating harmful data shifts for the responsible deployment of clinical AI models. *medRxiv*, 2023. doi: 10.1101/2023.03.26.23286718. URL https://www.medrxiv.org/content/early/2023/05/04/2023.03.26.23286718.

[8] Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In Marzyeh Ghassemi, Tristan Naumann, and Emma Pierson, editors, *ACM CHIL '21: ACM Conference on Health, Inference, and Learning, Virtual Event, USA, April 8-9, 2021*, pages 279–290. ACM, 2021. URL https://doi.org/10.1145/3450439.3451878.

[9] Samantha M. Santomartino, Nima Hafezi-Nejad, Vishwa S. Parekh, and Paul H. Yi. Performance and usability of code-free deep learning for chest radiograph classification, object detection, and segmentation. *Radiology: Artificial Intelligence*, 5(2), March 2023. ISSN 2638-6100. URL http://dx.doi.org/10.1148/ryai.220062.

[10] Shubhra Kanti Karmaker Santu, Md. Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. Automl to date and beyond: Challenges and opportunities. *ACM Computing Surveys*, 54(8):1–36, October 2021. ISSN 1557-7341. URL http://dx.doi.org/10.1145/3470918.

[11] Edward Korot, Zeyu Guan, Daniel Ferraz, Siegfried K. Wagner, Gongyu Zhang, Xiaoxuan Liu, Livia Faes, Nikolas Pontikos, Samuel G. Finlayson, Hagar Khalid, Gabriella Moraes, Konstantinos Balaskas, Alastair K. Denniston, and Pearse A. Keane. Code-free deep learning for multi-modality medical image classification. *Nature Machine Intelligence*, 3(4):288–298, March 2021. ISSN 2522-5839. URL http://dx.doi.org/10.1038/s42256-021-00305-2.

[12] Giridhar Dasegowda, Bernardo C. Bizzo, Reya V. Gupta, Parisa Kaviani, Shadi Ebrahimian, Debra Ricciardelli, Faezeh Abedi-Tari, Nir Neumark, Subba R. Digumarthy, Mannudeep K. Kalra, and Keith J. Dreyer. Radiologist-trained AI model for identifying suboptimal chest-radiographs. *Academic Radiology*, 30(12):2921–2930, December 2023. ISSN 1076-6332. URL http://dx.doi.org/10.1016/j.acra.2023.03.006.

[13] Trishan Panch, Erin Duralde, Heather Mattie, Gopal Kotecha, Leo Anthony Celi, Melanie Wright, and Felix Greaves. A distributed approach to the regulation of clinical ai. *PLOS Digital Health*, 1(5):e0000040, May 2022. ISSN 2767-3170. URL http://dx.doi.org/10.1371/journal.pdig.0000040.

[14] Ali Guermazi, Chadi Tannoury, Andrew J. Kompel, Akira M. Murakami, Alexis Ducarouge, André Gillibert, Xinning Li, Antoine Tournier, Youmna Lahoud, Mohamed Jarraya, Elise Lacave, Hamza Rahimi, Aloïs Pourchot, Robert L. Parisien, Alexander C. Merritt, Douglas Comeau, Nor-Eddine Regnard, and Daichi Hayashi. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*, 302(3):627–636, March 2022. ISSN 1527-1315. URL http://dx.doi.org/10.1148/radiol.210937.

[15] Rebecca M. Jones, Anuj Sharma, Robert Hotchkiss, John W. Sperling, Jackson Hamburger, Christian Ledig, Robert O'Toole, Michael

Gardner, Srivas Venkatesh, Matthew M. Roberts, Romain Sauvestre, Max Shatkhin, Anant Gupta, Sumit Chopra, Manickam Kumaravel, Aaron Daluiski, Will Plogger, Jason Nascone, Hollis G. Potter, and Robert V. Lindsey. Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. *npj Digital Medicine*, 3(1), October 2020. ISSN 2398-6352. URL http://dx.doi.org/10.1038/s41746-020-00352-w.

[16] Yangling Ma and Yixin Luo. Bone fracture detection through the two-stage system of crack-sensitive convolutional neural network. *Informatics in Medicine Unlocked*, 22:100452, 2021. ISSN 2352-9148. URL http://dx.doi.org/10.1016/j.imu.2020.100452.

[17] Michel Dupuis, Léo Delbos, Raphael Veil, and Catherine Adamsbaum. External validation of a commercially available deep learning algorithm for fracture detection in children. *Diagnostic and Interventional Imaging*, 103(3):151–159, March 2022. ISSN 2211-5684. URL http://dx.doi.org/10.1016/j.diii.2021.10.007.

[18] Daichi Hayashi, Andrew J. Kompel, Jeanne Ventre, Alexis Ducarouge, Toan Nguyen, Nor-Eddine Regnard, and Ali Guermazi. Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. *Skeletal Radiology*, 51(11):2129–2139, May 2022. ISSN 1432-2161. URL http://dx.doi.org/10.1007/s00256-022-04070-0.

[19] Soheil Ashkani-Esfahani, Reza Mojahed Yazdi, Rohan Bhimani, Gino M. Kerkhoffs, Mario Maas, Christopher W. DiGiovanni, Bart Lubberts, and Daniel Guss. Detection of ankle fractures using deep learning algorithms. *Foot and Ankle Surgery*, 28(8):1259–1265, December 2022. ISSN 1268-7731. URL http://dx.doi.org/10.1016/j.fas.2022.05.005.

[20] Abu Mohammed Raisuddin, Elias Vaattovaara, Mika Nevalainen, Marko Nikki, Elina Järvenpää, Kaisa Makkonen, Pekka Pinola, Tuula Palsio, Arttu Niemensivu, Osmo Tervonen, and Aleksei Tiulpin. Critical evaluation of deep neural networks for wrist fracture detection. *Scientific Reports*, 11(1), March 2021. ISSN 2045-2322. URL http://dx.doi.org/10.1038/s41598-021-85570-2.

[21] Alfred P. Yoon, Yi-Lun Lee, Robert L. Kane, Chang-Fu Kuo, Chihung Lin, and Kevin C. Chung. Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. *JAMA Network Open*, 4(5):e216096, May 2021. ISSN 2574-3805. URL http://dx.doi.org/10.1001/jamanetworkopen.2021.6096.

[22] Kazuma Murata, Kenji Endo, Takato Aihara, Hidekazu Suzuki, Yasunobu Sawaji, Yuji Matsuoka, Hirosuke Nishimura, Taichiro Takamatsu, Takamitsu Konishi, Asato Maekawa, Hideya Yamauchi, Kei Kanazawa, Hiroo Endo, Hanako Tsuji, Shigeru Inoue, Noritoshi Fukushima, Hiroyuki Kikuchi, Hiroki Sato, and Kengo Yamamoto. Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. *Scientific Reports*, 10(1), November 2020. ISSN 2045-2322. URL http://dx.doi.org/10.1038/s41598-020-76866-w.

[23] Tsubasa Mawatari, Yoshiko Hayashida, Shigehiko Katsuragawa, Yuta Yoshimatsu, Toshihiko Hamamura, Kenta Anai, Midori Ueno, Satoru Yamaga, Issei Ueda, Takashi Terasawa, Akitaka Fujisaki, Chihiro Chihara, Tomoyuki Miyagi, Takatoshi Aoki, and Yukunori Korogi. The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs. *European Journal of Radiology*, 130:109188, September 2020. ISSN 0720-048X. URL http://dx.doi.org/10.1016/j.ejrad.2020.109188.

[24] Christian Blüthgen, Anton S. Becker, Ilaria Vittoria de Martini, Andreas Meier, Katharina Martini, and Thomas Frauenfelder. Detection and localization of distal radius fractures: Deep learning system versus radiologists. *European Journal of Radiology*, 126:108925, May 2020. ISSN 0720-048X. URL http://dx.doi.org/10.1016/j.ejrad.2020.108925.

[25] Chi-Tung Cheng, Chih-Chi Chen, Fu-Jen Cheng, Huan-Wu Chen, Yi-Siang Su, Chun-Nan Yeh, I-Fang Chung, and Chien-Hung Liao. A human-algorithm integration system for hip fracture detection on plain radiography: System development and validation study. *JMIR Medical Informatics*, 8(11):e19416, November 2020. ISSN 2291-9694. URL http://dx.doi.org/10.2196/19416.

[26] Seok Won Chung, Seung Seog Han, Ji Whan Lee, Kyung-Soo Oh, Na Ra Kim, Jong Pil Yoon, Joon Yub Kim, Sung Hoon Moon, Jieun Kwon, Hyo-Jin Lee, Young-Min Noh, and Youngjun Kim. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthopaedica*, 89(4):468–473, March 2018. ISSN 1745-3682. URL http://dx.doi.org/10.1080/17453674.2018.1453714.

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. URL http://dx.doi.org/10.1109/CVPR.2017.243.

[28] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

[29] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using DropConnect. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1058–1066, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/wan13.html.

[30] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[31] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. ISSN 1097-0142. URL http://dx.doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

[32] John Mongan, Linda Moy, and Charles E. Kahn. Checklist for artificial intelligence in medical imaging (claim): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2):e200029, March 2020. ISSN 2638-6100. URL http://dx.doi.org/10.1148/ryai.2020200029.

[33] Robert Lindsey, Aaron Daluiski, Sumit Chopra, Alexander Lachapelle, Michael Mozer, Serge Sicular, Douglas Hanel, Michael Gardner, Anurag Gupta, Robert Hotchkiss, and Hollis Potter. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, 115(45):11591–11596, October 2018. ISSN 1091-6490. URL http://dx.doi.org/10.1073/pnas.1806905115.

[34] Guillaume Reichert, Ali Bellamine, Matthieu Fontaine, Beatrice Naipeanu, Adrien Altar, Elodie Mejean, Nicolas Javaud, and Nathalie Siauve. How can a deep learning algorithm improve fracture detection on x-rays in the emergency room? *Journal of Imaging*, 7(7):105, June 2021. ISSN 2313-433X. URL http://dx.doi.org/10.3390/jimaging7070105.