IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 August 2023, accepted 5 September 2023, date of publication 11 September 2023,
date of current version 15 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314380

RESEARCH ARTICLE

Radiology Decision Support System for Selecting Appropriate CT Imaging Titles Using Machine Learning Techniques Based on Electronic Medical Records

PEYMAN SHOKROLLAHI[1], JUAN M. ZAMBRANO CHAVES[1], JONATHAN P. H. LAM[1], AVISHKAR SHARMA[1], DEBASHISH PAL[2], NAEIM BAHRAMI[2], AKSHAY S. CHAUDHARI[1], AND ANDREAS M. LOENING[1]

[1]Department of Radiology, School of Medicine, Stanford University, Stanford, CA 94305, USA
[2]GE HealthCare, Sunnyvale, CA 94089, USA

Corresponding author: Peyman Shokrollahi (pshokrol@stanford.edu)

This work was supported in part by General Electric (GE) HealthCare.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) at Stanford University under Application No. 56914, and performed in line with the Declaration of Helsinki.

ABSTRACT Radiologists use an imaging order from the ordering physician, which includes a radiology title, to select the most suitable imaging protocol. Inappropriate radiology titles can disrupt protocol selection and result in mistaken or delayed diagnosis. The objective of this work is to develop an algorithm to predict correct radiology titles from incoming exam order data. The proposed instrument is an ensemble of five decision tree-based machine learning (ML) techniques (Light Gradient Boosting Machine, eXtreme Gradient Boosting Machine, Random Forest, Adaptive Boosting, and Random UnderSampling Boosting Model) trained to recommend radiology titles of computed tomography imaging examinations based on electronic medical records. Issues of imbalanced data and generalization were addressed. The tuned models were used to predict the top three radiology titles for the radiologist revision. The models were evaluated using a 10-fold cross-validation method, yielding an approximate average accuracy of 80.5% $\pm$ 2.02% and F1-score of 80.3% $\pm$ 1.67% for all models, while the ensemble classifier ($\sim$83% F1-score) outperformed individual models. An accumulated average accuracy of $\sim$92% was obtained for the top three predictions. ML techniques can predict radiology titles and identify highly important features. The proposed system can guide physicians toward selecting appropriate radiology titles and alert radiologists to inconsistencies between the radiology title in the exam order and the patient's underlying conditions, thereby improving imaging utility and increasing diagnostic accuracy, which favors better patient outcomes.

INDEX TERMS Artificial intelligence, machine learning, boosting, electronic medical records, protocols, computed tomography.

I. INTRODUCTION

Approximately 70 million individuals undergo computer tomography (CT) scans annually in the USA alone, and CT use is expected to continue to increase [1]. Artificial intelligence (AI) methods, including machine learning (ML) techniques, are a promising approach to improving the efficiency of workers within the workflow for CT scans. Thus far, researchers have focused mostly on the development of AI systems for downstream image interpretation tasks. There has been extensive research on tasks such as classifying images [2], identifying follow-up recommendations [3],

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu.

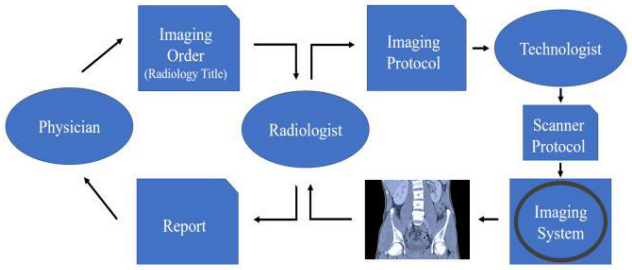


FIGURE 1. Radiology workflow – a referring physician makes an imaging order, the radiologist takes the imaging order and selects an appropriate protocol, the technologist performs the imaging acquisition through selecting appropriate machine-level parameter. The image is interpreted by a radiologist, and the results are reported to the ordering physician.

denoising and reconstructing images [4], detecting lesions and organs [5], and synthesizing super-resolution medical images [6]. Much of the work focused on developing deep learning (DL) AI algorithms for these tasks, including convolutional neural networks (CNNs) [2], 3D CNNs for MRI and CT images [5], recurrent neural networks [3], and generative adversarial networks (GANs) [4], particularly attention-based GANs [6]. Other research has focused on developing transfer learning algorithms for image analysis tasks, segmentation, object identification, disease categorization, and severity grading [7].

However, fewer studies have been performed on upstream radiology tasks such as title selection and protocol determination. We are not aware of any decision support system for radiology titles. There is scarce literature on the development of DL models for automating protocol selections using natural language processing (NLP). Trivedi et al. developed Watson DL to use free text to select contrast agents in MRI protocols [8], and Lee developed a CNN mechanism to determine routine or tumor MRI protocols [9]. However, these two works are limited to musculoskeletal MRI protocols. Few AI systems have been developed for protocol, contrast-agent, and machine-specification selection [10], [11], [12].

In a radiology workflow, a referring physician makes an imaging order, the radiologist translates this order to an appropriate imaging protocol, and a technologist tunes the machine based on the scanner protocol to perform the imaging acquisition. The radiologist interprets the images and reports the results to the ordering physician (Figure 1). The imaging request includes a physician-assigned title (referred to as the ''radiology title'' in this article) that indicates the type of exam to be performed on the patient. In tailoring the patients' needs to their care pathway, the goal is to design and deliver optimal outcomes and provide the best patient experience. There is substantial value at the beginning of the imaging chain when an imaging examination is ordered (i.e., assigning a radiology title) because this actionable information is translated by a radiologist to a more specific task of assigning a radiology protocol. The radiology protocol then determines how a technologist scans the patient. An inappropriate imaging examination may have no value for a referring physician to address patient needs if the radiology title is wrongly selected at the beginning. The appropriate selection of examination order can be achieved using a decision support system, through which referring physicians can select the radiology title and be made aware of the need for appropriate title assignments. Since ordering physicians are not always aware of the complexity of radiology protocol selections available to radiologists, the decision support system provides uniform information to radiologists to select the most suitable protocols, thereby optimizing protocol selections and the imaging workflow [13]. Consequently, many care providers across different departments can uniformly design and deliver optimal protocols for every circumstance [13].

Radiologists' CT protocol selections include selection of an anatomical region (e.g., abdomen) and focus within the region (e.g., liver) and whether exogenous contrast will be used (e.g., intravenous contrast) [9]. Radiology protocols consist of precise instructions for obtaining a desired set of medical images, and are used to translate referring physicians' orders into specific radiological imaging tasks.

However, because the primary role of radiologists is image interpretation, protocoling can be viewed as an interruptive task that reduces radiologists' workflow efficiency. Furthermore, it is time-consuming, cumbersome, and error-prone [11]. Protocol assignment takes approximately 6% of a radiologist's time [14].

Inappropriate CT protocol selection can result in missing diagnostic information, thereby jeopardizing patient health, delaying treatment, and increasing healthcare costs [11], [15]. Furthermore, although CT has advantages over other imaging modalities owing to its high spatial resolution and consistent quality, there remains a need to minimize exposure to potentially harmful radiation in CT scans [16]. Patients should be exposed to the lowest amount of radiation requisite for their specific clinical question. Conversely, optimal protocol assignment facilitates accurate diagnosis, reduces clinical uncertainty and follow-up examinations, thereby minimizing patient exposure to radiation and contrast agents. These benefits increase patient safety. Due to avoiding the repetition of the scan, it also reduces human workloads and infrastructure utilization costs [14], [17]. Much research has been done to develop DL algorithms for image-based radiologic diagnoses [18] as well as the acquisition, reconstruction, and interpretation of MRI data [19]. However, little has been done to model radiology protocols [9], [14], [15], [17]. To our knowledge, no ML/AI system has yet been developed for modeling radiology titles.

We developed a modeling system to predict radiology titles (e.g., CT abdomen triphasic liver with contrast, CT abdomen and pelvis with contrast) based on radiology exam orders from referring physicians (e.g., ''CT abdomen, with contrast, indication: hepatocellular carcinoma follow-up''). These radiology examination titles can be used by radiologists to aid appropriate protocol selection and to improve workflow

efficiency. Our model is implemented using data from the patient electronic medical records (EMRs).

One of the challenges in generating this model was accessing EMR data. EMRs are not readily accessible due to protections on patient privacy and personal information. The challenges involved in working with EMRs include heterogeneity, incompleteness, and imbalanced data. Even within a single institution, EMR data are often recorded differently, although they obey standard rules for being recorded. Since some diagnostic tests are dangerous or costly, the sequence of testing is incomplete in many cases. Thus, the various types of recorded data are imbalanced. Finally, methods for data analysis must be transparent and cannot use a black box technique. Given these constraints, there is increasing demand for a robust ML pipeline to address these issues [20]. Since inherently interpretable models have been strongly recommended for use in critical decision support systems such as healthcare [21], we used intuitive decision tree (DT)-based models with interpretable outputs. DTs combine simple classifiers (weak learners) and thus enable decisions to be made based on a classifier ensemble rather than single classifiers. DTs can be traversed by going through nodes owing to their tendency to be approximately balanced. Because each node requires checking the value of only a single feature, the overall prediction complexity is independent of the number of features. Thus, DTs can make predictions very fast, even with large training sets [22].

We incorporated boosting methods to support prediction of the most suitable titles [22]. Boosting is a process whereby simple classifiers are combined with weak learners to augment performance relative to a single classifier. A boosting ensemble of classifiers learns and combines many weak classifiers rather than learning a single robust classifier. It thus constitutes a robust complex classifier unto itself [23]. Generally, boosting algorithms train predictors sequentially, each trying to correct its predecessor [22]. They outperform in the processing of data with higher-order relationships. Boosting algorithms have been shown to surpass other ML models in several tasks, including in the emergency department triage [24].

Extreme gradient boosting (XGB) and light gradient boosting model (LGBM) algorithms are two efficient and optimized boosting methods. XGB is a highly effective scalable end-to-end tree-boost algorithm widely used in ML that can be employed effectively with sparse data [22], [25]. It uses a weighted quantile sketch to determine an efficient splitting point, effective memory usage, proper data compression to store the data efficiently, and selective sharing to make a scalable tree boosting system that is fast and efficient. XGB has been shown to be an effective classification and prediction algorithm in diverse applications, including predicting chronic obstructive pulmonary disease [26], improving the efficiency of cataract management [27], predicting side effects of analgesics for osteoarthritis patients [28].

The LGBM algorithm also addresses efficiency and scalability issues, especially for data with high feature dimensionality and large datasets. Compared to XGB that addresses computation time issues, its requirement for scanning all records to determine all possible split points is time-consuming. The LGBM algorithm addressed this issue by excluding a portion of data information, which has small-gradient information by using Gradient-based One-Side Sampling (GOSS) algorithm, and then using the remainder of information to estimate information gains. Thus, the LGBM has become a fast and efficient algorithm for ML modeling, with the tradeoff that its data exclusion makes it prone to overfitting [29]. LGBM and XGB have been used in developing decision support systems for orthodontic applications [30]. RF and XGB were used to develop a decision support system for predicting COVID-19 mortality [31].

Adaptive Boosting (AdaBoost) is a well-established boosting algorithm that uses weighted versions of the same training dataset instead of randomly subsampling the training set. In this algorithm, a set of weak learners operates sequentially, using reweighted versions of the training set, with the weights depending on the accuracy of the previous classifiers. The training set is always the same at each iteration, with each training instance weighted according to its misclassification by the previous classifiers. Thus, the weak learner focuses on misclassified patterns by considering the previous weak classifiers [23]. Thus, it does not need a large dataset to be trained. AdaBoost has been used in a decision support system for chronic type 2 diabetes [32] and for detecting lung cancer [33].

The random undersampling (RUSBoost) model is another boosting algorithm used to address the issues of learning from skewed training data [34]. RUSBoost uses a random undersampling mechanism to decrease the majority class instances in each boosting round [35]. RUSBoost was also used to develop a decision support system for early sepsis prediction [36] and stroke alert [37].

The random forest (RF) is a nonlinear classification method that builds a DT ensemble [38]. It uses a combination of DT predictors wherein each tree depends on independently sampled values of a random vector [39]. RF is an appropriate model for high-dimension data, data with missing values, data composed of various data types (e.g., numerical and categorical) [40]. Its ensemble strategies and random sampling help it to overcome overfitting issues [40]. Random forest has been used to develop several decision support systems for critical care [41] in predicting disease survivability [42] and heart arrhythmia [43].

Voting classifiers can be used to enhance modeling performance. They train other base-learner algorithms (including boosting and RF models), aggregate the predictions of each constituent classifier, and use a soft-voting mechanism to predict the highest-probability class, averaged over all the individual classifiers [22]. Soft voting usually results in higher performance than hard voting because it gives more

weight to highly confident votes, i.e., in hard voting, the majority wins by voting every individual classifier to a class, and in soft voting, the prediction of every individual classifier as a probability determines the target class [22].

Herein, we describe a newly developed decision support system designed to predict radiology titles. The developed algorithm is an ensemble of five DT-based techniques trained to recommend CT radiology report titles based on EMR data. The ensemble mechanism consists of a voting classifier inclusive of LGBM, XGB, RF, AdaBoost, and RUSBoost models.

The following contributions were made in developing the decision support system for CT radiology title:

- Determining the preprocessing steps required for preparing the EMR data
- Addressing the missing value issues in EMRs
- Identifying the most suitable classification models
- Integrating all five models into one via a voting mechanism
- Addressing imbalanced data issues
- Determining important features
- Overcoming the issue of generalization (i.e., adaptation to new cases)

In this article, we introduce the design and implementation of the pipeline, demonstrate the evaluation methods and results obtained from each model, and establish a prediction mechanism for radiologist revisions.

II. METHODS

A. DATA COLLECTION

We gathered EMR data for patients who underwent abdominal CT scans in the Stanford healthcare system with Institutional Review Board approval (Approving Institution: Stanford University, IRB Board Protocol Number: 56914, Date of Approval: 10/26/2020). The extracted data set was anonymized, and included data regarding patient demographics, admission, order, procedure, radiology, diagnostic, and laboratory information were obtained from hospitals, clinics, and ancillary services as summarized in Appendix A, Table 3. The radiology titles were pulled from the radiology reports stored at the data repository of our institution.

B. PIPELINE STRUCTURE

Our pipeline consists of 5 parts (Figure 2).

1) DATA CLEANING AND PREPROCESSING

Only EMR data for patients who underwent abdominal region CT scans at 18–90 years old (N = 53,345) obtained between May 2017 and October 2021 were retained, including accompanying data from scans of other regions. 'Reference only' radiology images (i.e., images scanned outside of our healthcare institution) were excluded. Order Type was limited to Imaging and the canceled orders were removed. The names of attributes in categorical descriptions (orders, procedures, and diagnoses) were standardized. For example, 'Event Date' and 'Recent Encounter Date' were renamed simply Date. The earliest of multiple recorded dates (e.g., Result Date, Order Date, Taken Date) was retained for the Date parameter. Diagnosis Type indicating the type of patient's problem (e.g., Primary, Chronic, or Hospital problem) was limited to Primary and Primary ED.

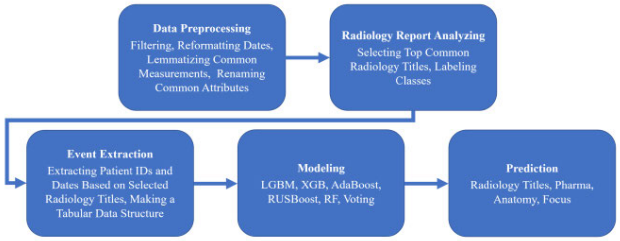


FIGURE 2. Pipeline structure – The pipeline consisted of data preprocessing to filter data, radiology report analysis to select top radiology titles, event extraction to extract patient data associated with selected radiology titles, modeling to train ML models, and predictions to provide top three radiology titles and automate the top one selected title above the determined thresholds.

Laboratory results related to kidney function (e.g., estimated glomerular filtration rate and creatinine) were selected and lemmatized to similar clinical measurements (e.g., 'urine creatinine' and 'creatinine, urine'). Any measurement values mixed with letters or signs were excluded. We retained smoking history in the form of nine subcategories such as every-day, some-day, regular, former, never, passive, unknown, heavy, and light smoker. We excluded alcohol use as a parameter because its reporting was not standardized.

2) RENAMING COMMON ATTRIBUTES

Attributes derived from various files and recorded under more than one name (e.g., Diagnosis Type and Diagnosis Description) were renamed to institute distinct names.

3) RADIOLOGY REPORT ANALYSIS

The radiology titles were obtained from radiology reports in our institution's database. They were ranked by frequency of occurrence. The top 15 radiology titles were selected (Figure 3) and the corresponding patient ID's and dates were retained.

4) SELECTING RECORDS AND MERGING FILES

Ultimately, data from 46,362 patients (multiple records per patient allowed) formed the input signals inclusive of 134,089 records of selected titles yielding feature dimensions of 30 attributes (Table 1). Figure 4(a) illustrates the Standards for Reporting of Diagnostic Accuracy (STARD) flow chart diagram that indicates the number of patients, number of records, and reasons for exclusion from original data to prepared data for our modeling.

The data size illustrated in Figure 4(b) indicates the number of patients and number of records in the initially collected cohort from the Radiology Information System (RIS, a subcomponent of the EMR) for the patients who

TABLE 1. Selected features.

File	FEATURE NAMES							
ADT	Event Type In	Patient Service	Patient Class	Patient Level Care	Department			
Demographics	Sex	Race	Ethnicity	Height	Weight	Smoking Hx†	Insurance name	Insurance Type
Labs	Creatinine, Serum	Creatinine, Urine	BUN/Creatinine Ratio	eGFR	eGFR (African American)			
Orders	Order Type Code	Order Type	Proc ID	Order Description	Quantity	Order Status Code		
Diagnoses	ICD10 Code	Diagnosis Description	Source					
Procedures	Code Type							
Radiology	Type							

Age

† Smoking Hx: smoking history.

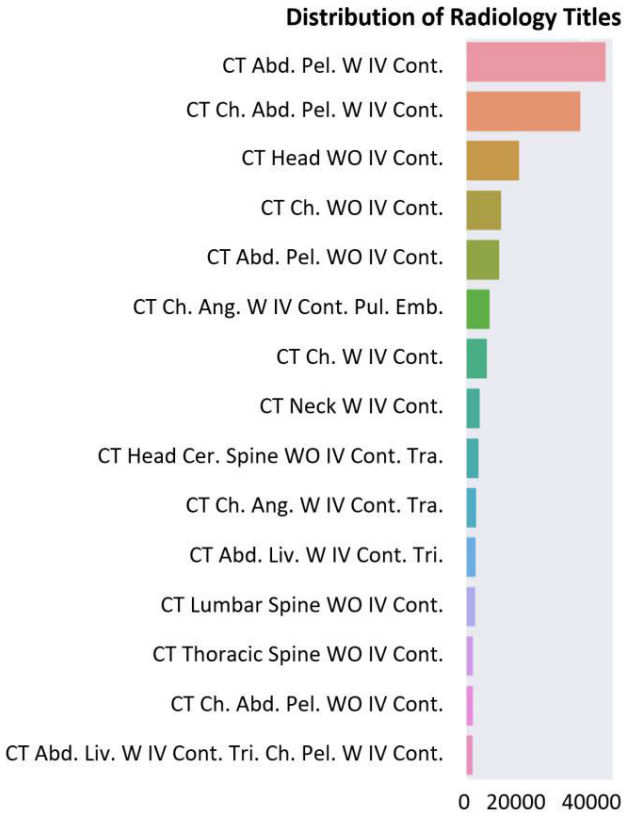


FIGURE 3. Distribution of radiology titles. The top 15 CT radiology titles were selected and labeled for supervised learning. The classes were labeled Class 0 to Class 14 from top to bottom on this distribution (Abdomen (Abd.), Angiography (Ang.), Chest (Ch.), Contrast (Cont.), Cervical (Cer.), Embolism (Emb.), Intravenous (IV), Liver (Liv.), Pelvis (Pel.), Pulmonary (Pul.), Triphasic (Tri.), Trauma (Tra.), With (W), and Without (WO)).

mainly undergone abdominal CT scans, and selected data refers to the data selected after selecting 15 top radiology titles.

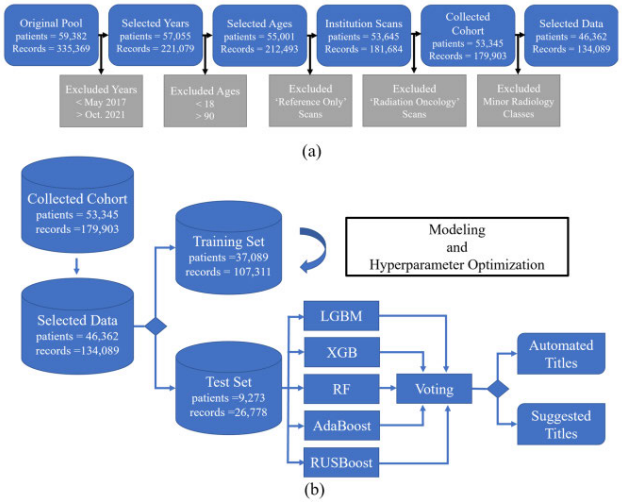


FIGURE 4. a) The Standards for Reporting of Diagnostic Accuracy STARD flowchart diagram – STARD indicates the number of patients and records in the original pool and exclusion criteria used to prepare the selected data for modeling, and b) data size and overall mechanism – Data size indicates the number of patients and number of records in the collected cohort from the Radiology Information System (RIS) for the patients who have undergone CT scans. Selected data refers to the data selected after selecting 15 top radiology titles, and the data was split by 80% and 20% by patients. Overall mechanism indicates modeling and hyperparameter optimization performed on training set. The test set is used to evaluate the models. Predicted titles with a probability above a certain threshold are automatically assigned; the rest will be provided to radiologists for their revisions.

5) MODELING

Numerical features (e.g., age, weight, and height) were scaled by the MaxAbsScaler method, which normalizes values by dividing them the maximum absolute value. Categorical data were transformed to numerical values between 0 and the number of subcategories for that feature with the label encoder in the Scikit Learn software package (0.24.2). We added 1 to

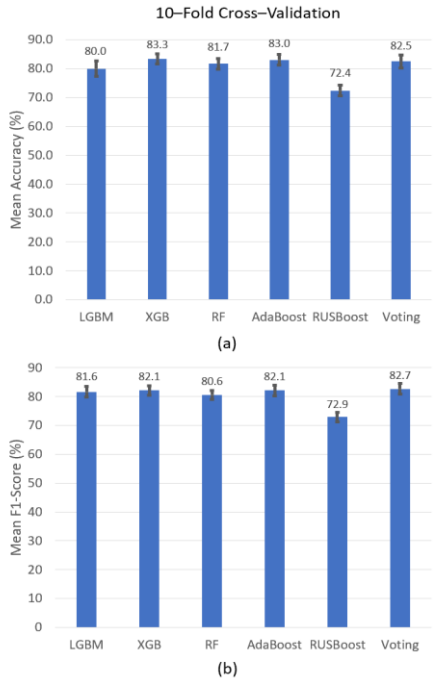


FIGURE 5. Cross-validation results for the entire dataset. The 10 – fold cross-validation method to evaluate model performance, represented by a) accuracy and b) F1-score.

all encoded values to avoid mixing encoded 0 values with missing values, which were filled with zeros.

The data were allocated to training (80%) and test (20%) sets randomly by patient. The entire dataset was split into training and test sets using a stratified sampling mechanism with a single number of split and random shuffle. The training set was then used for model training, and the test set was left intact for calculating the performance of the model predictions, as illustrated in Figure 4(b).

LGBM, XGB, RF, AdaBoost, and RUSBoost algorithms were used for supervised learning because of their ability to model non-linear relationships in EMR data [24]. Hyperparameters (maximum depth, learning rate, and number of leaves) were tuned with a grid search mechanism in a five-fold cross-validation method. The list of selected hyperparameters, their values, and the range of values for tuning the models are reported in Appendix B. To evaluate the generalization performance of the algorithms, a 10-fold cross-validation mechanism was performed on the entire dataset using accuracy and F1-score as the scoring function. In each cross-validation iteration, nine folds of data were used for training and one-fold for the test. Means and standard deviations are shown in Figure 5.

Summary plots of Shapley Additive Explanations (SHAP) values were developed to reveal the contribution of each feature to predictions and to indicate the relative importance of each feature [44].

We employed a voting classifier trained by LGBM, XGB, RF, AdaBoost, and RUSBoost classifiers. The voting classifier aggregated the predictions of each component classifier and predicted the most frequently voted class. It used soft voting to predict the class with the highest probability, averaged over all individual classifiers from the outputs of the models. Imbalance in the dataset was addressed by including class weights in the models.

Missing values were filled in based on prior patient data. The most time-proximal data source within a window of a year from the target abdominal CT was used. Otherwise, they were considered missing values.

6) PREDICTING

Finally, the tuned models were used to predict radiology titles and their probabilities. The pipeline produced a list of top-three title suggestions, based on F-scores, with their probabilities. In addition, a threshold value was determined to indicate a minimum probability for automating radiology title selection by the system.

III. RESULTS

Table 2 indicates the results of model evaluations, including accuracy, precision, recall, and F1-scores for all models on a held-out test set. The models were also evaluated using cross-validation, yielding an approximate average accuracy of 80.5% ± 2.02% and F1-score of 80.3% ± 1.67% for all models (Figure 5).

Procedure ID (Proc ID) (determined in the ordering process), Order Description, Patient Class (Pat Class), and ICD10 (International Classification of Diseases) Code, which were the most important features according to our SHAP plots, played key roles in modeling and correlated with routine title selections (Figure 6). The ICD10 codes were not stratified to avoid reducing the variability in these codes.

The top-three radiology titles produced, along with their probabilities, were used to select the most appropriate titles in the context of a clinical decision support system. These top-three most probable predicted titles and their probabilities were reported for radiologist review (Figure 7).

We obtained an accumulated accuracy of 91.2%, 93.4%, 92.6%, 93.2%, 90.1%, and 93.2% for the top-three predicted titles for LGBM, XGB, RF, AdaBoost, RUSBoost, and Voting, respectively (Figure 8).

A threshold value can be defined to indicate the titles that will require subsequent radiologist or technologist revision. The probabilities were analyzed for all models to determine the threshold value for automating the radiology

TABLE 2. Summary of model evaluations.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LGBM	79.2	83.7	79.2	80.0
XGB	81.3	79.4	81.3	78.0
RF	82.0	81.2	82.0	80.4
AdaBoost	72.2	75.3	72.2	69.2
RUSBoost	77.7	78.6	77.7	78.1
Voting	82.9	83.0	82.9	82.9

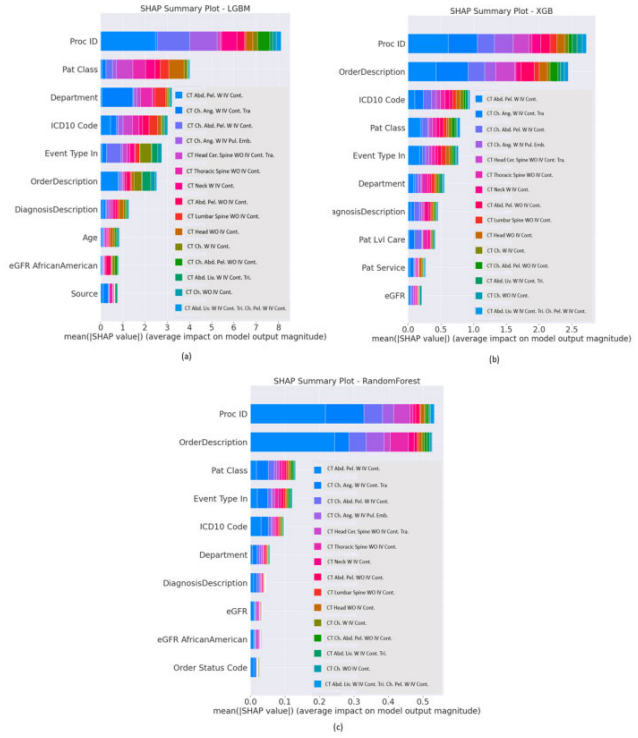


FIGURE 6. SHAP Summary plot. Summary plots of Shapley Additive Explanations (SHAP) values were developed for the LGBM, XGB, and RF models to explain the contribution of each feature to the prediction and indicate the relative importance of each feature. Proc ID = Ordering Procedure ID, Pat Class = Patient Class, ICD = International Classification of Diseases, Pat Lvl Care = patient level care; for description refer to Appendix A. (Abdomen (Abd.), Angiography (Ang.), Chest (Ch.), Contrast (Cont.), Cervical (Cer.), Embolism (Emb.), Intravenous (IV), Liver (Liv.), Pelvis (Pel.), Pulmonary (Pul.), Triphasic (Tri.), Trauma (Tra.), With (W), and Without (WO)).

Choice 1	Choice 2	Choice 3	None of the Above	Truth Label	Top-1 Correct	Top-3 Correct
CT ABDOMEN PELVIS W IV CONTRAST (0.86)	CT CHEST ABDOMEN PELVIS W IV CONTRAST (0.02)	CT ABDOMEN PELVIS WO IV CONTRAST (0.02)	0.10	CT ABDOMEN PELVIS W IV CONTRAST	Yes	Yes
CT NECK W IV CONTRAST (0.46)	CT HEAD WO IV CONTRAST (0.12)	CT ABDOMEN PELVIS W IV CONTRAST (0.07)	0.35	CT HEAD WO IV CONTRAST	No	Yes
CT HEAD WO IV CONTRAST (0.71)	CT ABDOMEN PELVIS W IV CONTRAST (0.06)	CT CHEST ANGIOGRAPHY W IV CONTRAST PULMONARY EMBOLISM (0.05)	0.18	CT CHEST ANGIOGRAPHY W IV CONTRAST PULMONARY EMBOLISM	No	Yes
CT HEAD CERVICAL SPINE WO IV CONTRAST TRAUMA (0.4)	CT CHEST ANGIOGRAPHY W IV CONTRAST TRAUMA (0.15)	CT ABDOMEN PELVIS W IV CONTRAST (0.11)	0.34	CT LUMBAR SPINE WO IV CONTRAST	No	No

FIGURE 7. Examples of voting classifier prediction results. The predicted titles and their probabilities were reported for radiologist review.

title selections. The Voting Classifier determines a threshold value of approximately 0.9 for the current data set. Although obtaining the threshold value that can be in clinical use requires further clinical evaluations, the obtained threshold just indicates the feasibility of the pipeline to segregate predicated titles to automated and revised cases. The receiver operating characteristic (ROC) curve was obtained using the One-vs-Rest method for all classes. Figure 9 illustrates the ROC curves of the top five classes and the macro and micro averaging plots of all 15 classes, where Class 0, CT Abdomen Pelvis w IV Contrast (AP w IV Cont.); Class 1, CT Chest

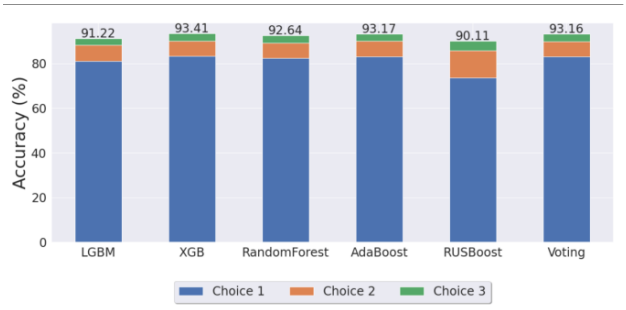


FIGURE 8. Top three predicted accuracies. Accumulated accuracies for the top three predicted titles were evaluated for all models.

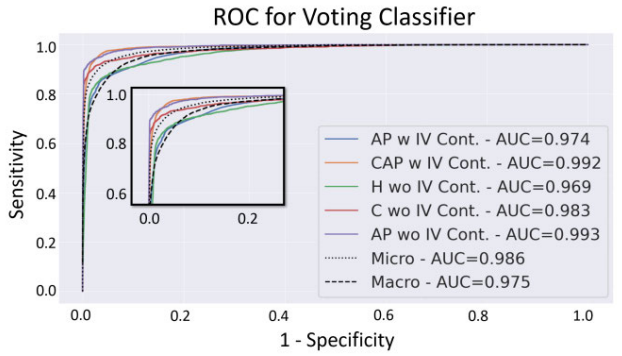


FIGURE 9. The characteristic (ROC) curve for the top five classes and the macro and micro averaging of all 15 classes using the One-vs-Rest method. Class 0 to 4 abbreviated using A for Abdominal, C for Chest, H for Head, P for Pelvis, H for Head, w for with, wo for without, and Contr. for Contrast. The inset illustrates the zoomed ROC curves.

Abdomen Pelvis w IV Contrast (CAP w IV Cont.); Class 2, CT Head wo IV Contrast (H wo IV Cont.); Class 3, CT Chest wo IV Contrast (C wo IV Contr.); and Class 4, CT Abdomen Pelvis wo IV Contrast (AP wo IV Cont.). The ROC curves' micro and macro averaging were calculated with the area under the curve (AUC) of 98.6% and 97.5%, respectively.

IV. DISCUSSION

Based on important feature plots, the role of features in each model was compared comprehensively to reveal the most important features [38]. The top four features were Proc ID, Order Description, Pat Class, and ICD10.

Due to variations among protocols, we initially limited our data pull to abdominal CT scans, which themselves already have substantial variation. At our institution, the abdomen is the most common region subjected to CT routinely. Abdominal CT is used to evaluate abdominal, flank, and pelvic pain, to detect masses and fluid collections, to identify sites of malignancy, inflammation, and infection, to reveal causes of bowel obstruction, weight loss, and fever, and to clarify laboratory or other imaging findings [45]. However, we expanded our database by including head CT and chest CT without IV contrast of the patients who had undergone abdominal CT, principally due to non-localized conditions, such as metastasis. Due to recording inconsistencies between the ICD10

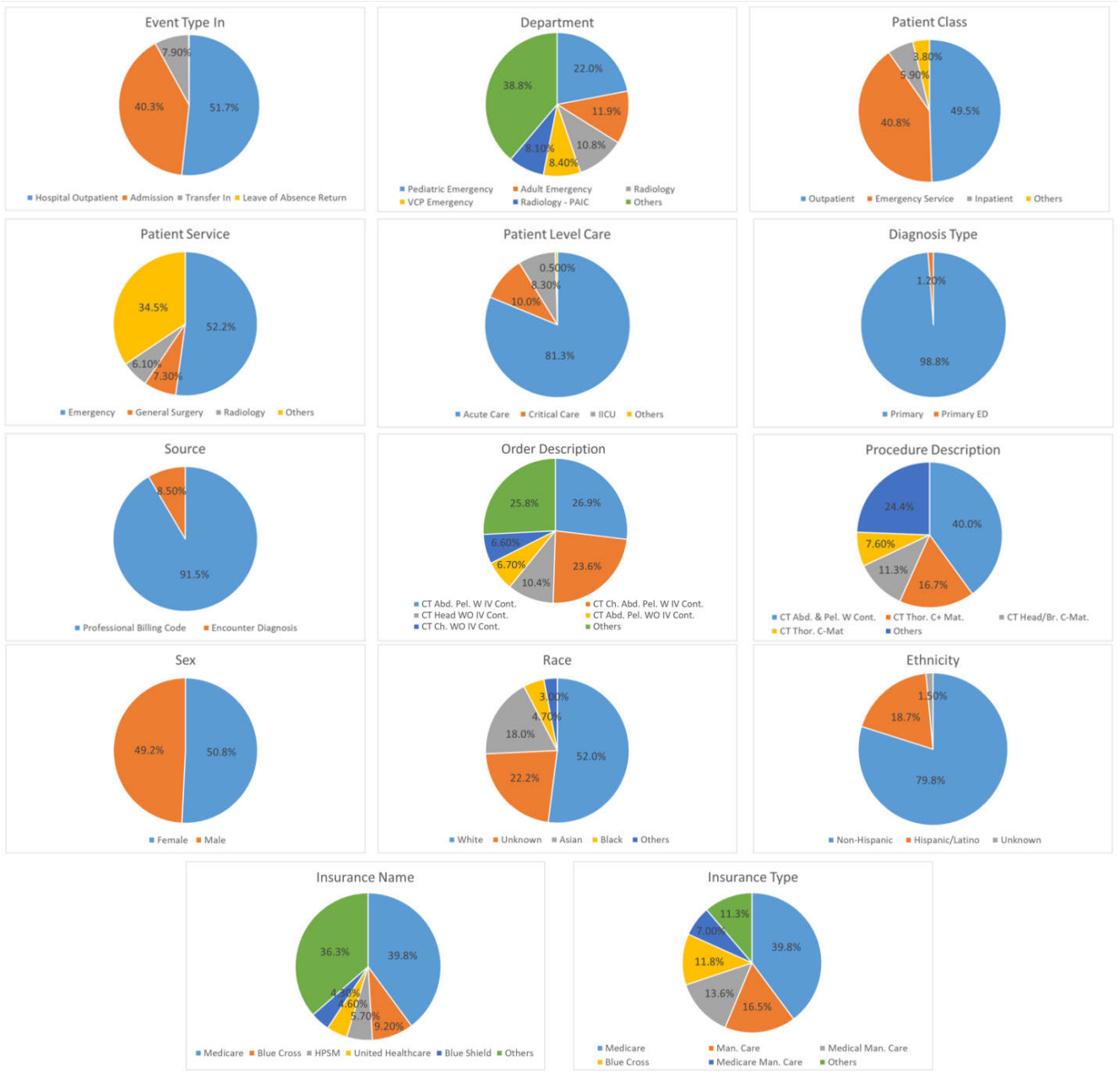


FIGURE 10. Pie charts of categorical variable distributions.

codes and their descriptions, as well as Ordering Procedure IDs and their descriptions, we kept the codes and their descriptions in the input signals. This is the reason for obtaining different roles between the codes and their descriptions in the SHAP summary plot.

Understanding the data distribution and maintaining it after splitting are critical because the dataset should be verified to be an appropriate proxy for the ability of interest for the modeling goal [46]. Appendix C includes a table indicating the distribution of numerical variables and pie charts illustrating the distribution of categorical variables.

Normalization of numerical variables and transformation of categorical variables were performed to standardize data before tuning the models. This process conforms the collected data to a standard scale and constrains each variable to a limited range, thereby allowing data recorded from different sources to be used together. Standardization also reduces model sensitivity to the scale of the input variables.

The patient information was not limited to radiology reports. In addition to radiology reports the EMRs also included admissions (ADT), procedures, labs, orders, and diagnoses.

TABLE 3. Summary of selected attributes.

File	Attribute	Description
ADT	Event type in	Type of admit/transfer event (e.g., Hospital outpatient; Transfer in; Admission).
	Patient service	Category value corresponding to the hospital service for the patient of the event record at the effective time (e.g., Radiology, Cardiology, Endoscopy, etc.).
	Patient class	String category value corresponding to the classification for the patient of the event record at the effective time (e.g., Outpatient, Inpatient, Radiation therapy series, etc.).
	Patient level care	String category value corresponding to the level of care for the patient of the event record at the effective time (e.g., Acute care, Critical care, Intermediate care, etc.).
Procedures	Code	CPT, ICD9CM, or ICD10PCS code.
	Type code	String, or number in some cases, indicating the procedure code type (e.g., 74160 for CPT; 45.23 for ICD9CM; and 07TP0ZZ for ICD10PCS).
	Description†	Categorical text describing the procedure code.
Labs	Results	Laboratory variable name acting as a grouper for similar terms. In general, List of Value (LOV) variables are reasonably clean but Epic permits value list customization and there may be historical data present that uses different strings for synonymous concepts, requiring some pre-analysis cleansing, namely unification of terms.
	Values	The original string containing the result. When the result is numeric, the numeric conversion appears in ord_num_value form (Lab Result Numeric). Owing to data cleansing, attributes had numeric identifiers (e.g., 42.75, 100).
	Units	Reference Range Units for LOV variables. Due to the potential for historical data with different strings for synonymous concepts, some pre-analysis cleansing may be required. We tested unit uniqueness and cleansed units for form variations (e.g., mg/dL, mL/min/1.73 m^2, etc.)
Orders	Order type	Descriptive string for order type (Imaging, Lab, ECG, etc.).
	Order type code	Numeric value indicating order type (1-, 2-, or 4-digit code).
	Description	A brief summary of the procedure order in text format (e.g., CT Abdomen with IV Contrast, XR Abdomen 1 View, MR Brain with and without Contrast).
	Order status code	Categorical (string) value indicating status [Pending (1), Sent (2), Results obtained (3), Cancelled (4), Completed (5), Holding for referral (6), Denied approval (7), Suspended (8), Discontinued (9), Verified (10), Dispensed (11), and Pending Verification (12)] We selected only completed (5)-status data.
	Proc ID	Internal identification for the procedure being ordered (4- or 6-digit number) duplicate of order status code was deleted from this row
	Quantity	The numerical code associated for one of the Order Status Code, <- this description doesn't appear to match "Quantity". Your list may be discombobulated here, please check
Diagnoses	Source	One of the following: Encounter, Problem, Professional Billing code, Admit diagnosis, or External injury code.
	Type	One of the following: present-on-admission, primary, chronic, hospital problem. There may be multiple types for a given diagnosis.
	ICD10 code	ICD-10 code for the patient's diagnosis. Where the procedure was coded originally with an ICD-9 code, there may be multiple ICD-9 codes given.
	Description	Text (categorical data) describing the ICD code (e.g., Abdominal distension).
Radiology	Type	Categorical (string) value indicating imaging type (e.g., CT, MRI, US, X-ray).
	Title	Categorical (string) value indicating the description of the performed imaging protocol (e.g., CT Abdomen with IV Contrast)

TABLE 4. Selected Hyperparameters.

Model	Hyperparameter	Selected Value	Tuning Range
LGBM	Feature Fraction by Node	0.8	
	Learning Rate	0.1	[0.01 – 0.5]
	Boosting Type	GBDT	[DART, GBDT]
	Objective	Multiclass	
	Metric	Multiclass Logarithmic Loss	
	Class Weight	Balanced	
	Maximum Depth	20	[10 – 30]
	Number of Leaves	50	[10 – 100]
	Maximum Bin	500	
	Lambda L1	0.01	
	Lambda L2	0.01	
	N Estimators	50	[10 – 80]
XGB	Importance Type	Gain	
	Learning Rate	0.01	[0.001 – 0.05]
	Missing	1	
	Gamma	3	[1 – 10]
	N Estimators		
	Random State	42	
	Regularization Lambda	5	
	Evaluation Metric	Multiclass Logarithmic Loss	
	Column Sample by Tree	0.3	
	Maximum Depth	25	[5 – 40]
	Seed	0	
	Class Weight	Balanced	
	Objective	Multiclass Softmax	
RF	Bootstrap	True	
	Criterion	Entropy	[Entropy, GINI, MSE]
	Minimum Impurity Decrease	0.0	
	Minimum Samples Split	2	
	Minimum Weight Fraction Leaf	0	
	N Estimators	100	[50 – 200]
	Maximum Features	Auto	[Auto, SQRT, Log2]
	Maximum Depth	25	[10 – 50]
	Minimum Samples Leaf	10	[5 – 20]
	OOB Score	False	
	Random State	42	
	Verbose	0	
	Warm Start	False	
	Class Weight	Balanced	
AdaBoost	N Estimators	50	
	Learning Rate	0.001	[0.0001 – 0.01]
	Algorithm	SAMME.R	[SAMME.R, SAMME]
	Class Weight	Balanced	
RUSBoost	N Estimators	50	
	Learning Rate	0.001	[0.0001 – 0.01]
	Algorithm	SAMME	[SAMME.R, SAMME]

To minimize the impact of random split on the data distribution, the entire dataset was split using a stratified sampling mechanism. Data were split into strata (i.e., subgroups) sharing common characteristics. This method keeps the distribution of target variables (i.e., radiology titles) equivalent across various splits and reduces the sampling variability [47].

Dimensionality reduction was not applied because the feature space was not too large. Dimensionality reduction methods such as principal component analysis, singular value decomposition, and linear discriminant analysis are recommended when dealing with high-dimensional data [48]. Since the ratio of feature space dimensions to number of datapoints (30/134,089) was low, applying dimensionality reduction methods may not notably change the results. Future studies that use higher-dimensional feature space (i.e., by including image and free-text information) will apply these methods.

TABLE 5. Distribution of numerical variables.

Feature	Count†	Mean	STD	Min	25%	50%	75%	Max
Age	134089	59.5	16.9	18.0	48.3	61.8	72.4	90.0
Height	124961	167	10.9	21.8	160	168	175	216
Weight	129513	75.4	21.3	1.00	60.3	72.6	86.6	535
Quantity	123269	5.00	0.07	2.00	5.00	5.00	5.00	5.00
BUN/Creatinine Ratio	2116	17.5	6.11	3.00	13.0	17.0	21.0	56.0
Creatinine Serum	25514	1.12	0.993	0.200	0.700	0.900	1.20	18.6
Creatinine Urine	15481	79.0	123	0.720	33.7	60.6	104	9360
eGFR (African American)	68503	91.3	35.3	1.00	68.0	95.0	116	334
eGFR	69950	79.5	30.2	1.00	60.0	83.0	100	289
Smoking Hx††	134081	4.32	1.39	0.00	4.00	5.00	5.00	10.0

† Except Count, all variables are reported to three significant digits
†† Smoking Hx: smoking history

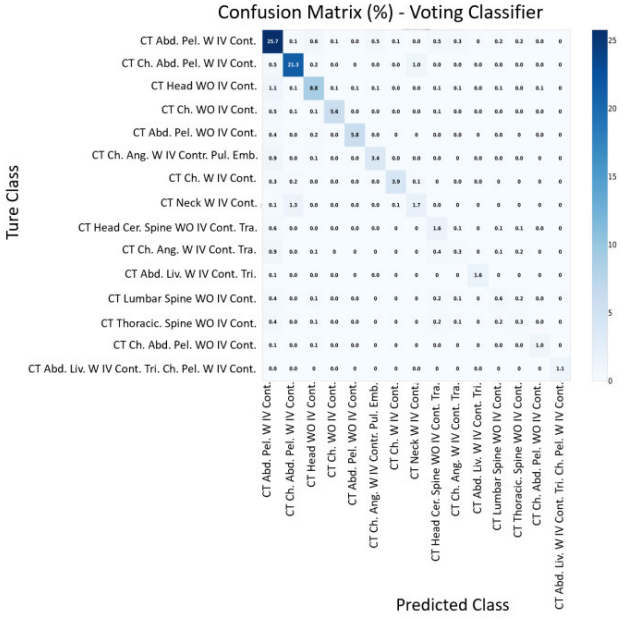


FIGURE 11. Confusion matrix in percentage for the voting classifier on the test set.

Two engineering challenges were addressed: overfitting and imbalanced data. For example, overfitting was present in XGB and RF models, and the training accuracy of 95.3% and 98.8% were obtained while their test accuracies were 55.1% and 53.2%, respectively. The issue of overfitting was overcome by expanding the cohort size and the feature space. To expand the cohort size, we included data for a wide age range (18, 90) and included all related scans, even beyond the abdomen. With a limited number of samples, rules can be too specific, leading to overfitting due to the induction rules that define minority concepts being much fewer or weaker than those of majority concepts [49]. Addressing overfitting by expanding the data such that other radiology titles were included (e.g., for head CT without contrast and chest CT without contrast) expanded the data beyond abdominal CT cases.

TABLE 6. Summary of missing imputations.

File	Attribute	Missing Percentage (%)			
		Initial Data	Imputed Using Previous Records	Imputed Just by MissForest	Imputed Using Previous Records & MissForest
ADT	Patient Level Care	87.4	51.0	87.4	51.0
	Patient Service	59.4	19.4	59.4	19.4
	Patient Class	0.844	0.0111	0.844	0.01
	Event Type In	0.00	0.00	0.00	0.00
	Department	0.00	0.00	0.00	0.00
Demographics	Height	9.30	9.30	0.00	0.00
	Weight	4.37	4.37	0.00	0.00
	Insurance name	7.88	7.88	7.88	7.88
	Insurance Type	7.88	7.88	7.88	7.88
	Race	0.00	0.00	0.00	0.00
	Smoking Hx†	0.00	0.00	0.00	0.00
	Ethnicity	0.00	0.00	0.00	0.00
	Sex	0.00	0.00	0.00	0.00
Labs	BUN/Creatinine Ratio	97.0	93.5	0.00	0.00
	Creatinine Serum	95.7	77.5	0.00	0.00
	Creatinine Urine	95.7	71.8	0.00	0.00
	eGFR (African American)	8.35	3.14	0.00	0.00
	eGFR	6.91	2.29	0.00	0.00
Orders	Order Status Code	3.23	0.435	3.23	0.435
	Quantity	3.23	0.435	0.00	0.00
	Proc ID	0.00	0.00	0.00	0.00
	Order Description	0.00	0.00	0.00	0.00
	Order Type	0.00	0.00	0.00	0.00
	Order Type Code	0.00	0.00	0.00	0.00
Diagnoses	ICD10 Code	0.0720	0.0000220	0.00720	0.0000220
	Diagnosis Description	0.00	0.00	0.00	0.00
	Source	0.00	0.00	0.00	0.00
Procedures	Code Type	0.00	0.00	0.00	0.00
Radiology Reports	Type	0.00	0.00	0.00	0.00

† Smoking Hx: smoking history

We studied the confusion matrices for all models using the test sets. Appendix D shows the confusion matrix in percentage for the voting classifier. Analysis of the confusion matrices derived on the test set using all models showed that most of the errors that occurred were in predicting all titles as the most common radiology title, CT Abdomen Pelvis with IV Contrast.

We assumed that overfitting was secondary to data imbalance, which can cause learning models to fail to generalize inductive rules. Thus, the imbalance issue was addressed by incorporating the reweighting schedule mechanism that includes sample weights in the models. With this mechanism,

the sample weights are changed to give more significance to minor class samples in the training data [50]. In addition, we exploited the class weight mechanism, in which the minority classes are penalized for misclassification more strongly than are the majority classes to balance the contribution of each class to the total loss [50]. This mechanism also reduces bias toward the majority classes and achieves well-balanced performance for all classes.

Boosting algorithms have the advantage of being highly interpretable. They make predictions efficiently because the performance of each model feature can be determined and thus linked to radiology images. For example, identifying the significance of the role of the Proc ID, Order Description, and ICD10 feature is reasonable because these features carry information about imaging order and selection practices. Although boosting is a resilient method for addressing overfitting, it can be highly sensitive to outliers due to each classifier trying to fix predecessor errors [22].

We evaluated two imputation mechanisms to address data missingness: filling with previous records and MissForest [51]. Table 6 in Appendix E shows the percentages of missing values in the initial data and after the use of each and both methods. The filling mechanism notably reduced the number of missing values in Patient Level Care, Patient Service, Patient Class, Order Status Code, and Quantity. MissForest played a role when the filling mechanism could not address the missingness, e.g., for demographic characteristics such as Height and Weight and laboratory results such as the BUN/Creatinine Ratio, Creatinine Serum, and Creatinine Urine.

Based on the model evaluations presented in Table 2, the highest accuracy score (82.9%) was obtained for the voting classifier on a held-out test set. A precision score of 79.5% was obtained on average. The LGBM and Voting classifiers obtained the highest precision (>83.0%), indicating acceptable performance in obtaining high true positive and low false positive rates. The voting classifier's highest recall score (82.9%) indicates that this model also outperformed in identifying true cases and has the lowest false negative rate. The highest F1-score (82.9%) was also obtained for the voting classifier. The results indicate that the voting classifier outperformed individual models, i.e., making the ensemble more effective than the individual ones.

The F1-score, used in the presently employed 10-fold cross-validation process, is an appropriate model performance metric for the evaluation of imbalanced data. It has been used widely to evaluate classifier performance when encountering a rare class [52]. Although the accuracy values were slightly larger than the F1-scores in most models, we considered the F1-score to be a more appropriate reflection of acceptable model performance because it is a "single-number evaluation metric" [53], which is a harmonic (i.e., weighted) mean of precision and recall used for model comparison [22]. Whereas the regular mean treats all values equally, the weighted mean assigns more weight to low frequencies. The F1-score for a classifier is high when both recall and precision are high [22]. We obtained the best F1-score for the voting classifier in the cross-validation evaluations. The F1-scores of $> 80\%$ indicate the acceptable performance of the models for 15 radiology titles. Considering similar systems and using an RF model, Brown and Marotta obtained 83% accuracy for only five brain MRI protocols [11]. It should be noted that the tail classes (>5) had less than approximately 800 instances, as this might be attributed to the scarcity of training data, specifically the limited number of instances available for the minor classes. Although the definition of an acceptable F1-score depends on the task and application, F1-scores > 0.8 are generally considered to be acceptable in the medical field, such as in applications for human activity recognition using deep learning [54], liver lesion detection using deep convolutional neural networks [55], and autism disorder identification using deep learning and support vector machine models [56].

Comparison of SHAP plots showing the contribution of each feature to the predictions and thus the relative importance of each feature indicated that soft voting performed well, presumably because it gives more weight to highly confident votes. The most important features were consistently detected by the models having SHAP plots. More accurate title prediction was obtained with the voting classifier (82.7%) than with any of the base classifiers within the ensemble (LGBM, 81.6%; XGB, 82.1%; RF, 80.6%; AdaBoost, 82.1%; RUSBoost, 72.9%). Obtaining 98.6% and 97.5% for the micro and macro averages of AUC indicates the well performance of the algorithm. While the AUC micro and macro averages (98.6% and 97.5%, respectively) are in agreement, the AUC micro average reflects the overall good performance of the algorithm, and the AUC macro average indicates that the algorithm performs well despite having imbalanced data. The AUC macro average is the preferred choice of evaluation in imbalanced multi-class settings [57] because the AUC macro average evaluates the AUCs independently for each class and then computes the overall average; thus, all classes are treated equally. However, the AUC micro average considers the contribution of each class in computing the average metric, and the metric may be affected by a large number of major classes.

The contribution of this paper includes the development of a preprocessing mechanism, selection of appropriate features, model design, and the optimization of model parameters to obtain the most suitable pipeline for selecting appropriate radiology titles. These radiology titles could be used to guide radiologists quickly to appropriate imaging protocols for incoming exam orders, as well as suggest appropriate reporting templates during image interpretation. In follow-up work, we will extend our algorithm to output appropriate imaging protocols for incoming orders. The developed algorithm presented here has the potential to reduce, to some extent, radiologist workloads, which would enable radiologists to focus their time on more critical tasks such as image

interpretation. Adoption of such algorithms can improve the efficiency of radiology service delivery and thus reduce healthcare costs. Such improvements have been estimated to have the ability to reduce radiology department costs by more than $30 million (US dollars) per year [58].

The collected and preprocessed patient EMRs were not limited to radiology reports. EMRs were collected from various sources such as admissions (ADT), procedures, labs, orders, diagnoses, and radiology reports. Although the present models were trained based on data obtained from a single institution, they can also be trained using other resources because recorded EMRs across the globe share the same syntax and semantic standards. To make EMRs interoperable and accessible, the recorded data follow certain rules to be accurate, consistent, and reproducible [59]. These standards include the Health Level Seven International (HL7) and Fast Healthcare Interoperability Resources (FHIR) specifications and Logical Observation Identifiers Names and Codes (LOINC). HL7 and FHIR standards ensure clinical data exchange. FHIR has been employed to standardize data formats, data elements, and application programming interface protocols for EMR exchange across various institutions [60], and LOINC has been used to standardize EMR terminologies for lab results, clinical measurements, documents, and surveys. In addition, radiology lexicon (RadLex) has been employed as a common lexicon across institutions. RadLex has standardized many radiological terms, including anatomy, diseases, and findings. The LOINC-RSNA Radiology Playbook has also been used across institutions as a consistent structure for radiology titles and imaging procedure names [61]. Thus, our proposed pipeline can be generalized and tested using EMR resources from other institutions in the US and around the globe. Similar data can be pulled from the resources of any institution, and the proposed pipeline can be used to fine-tune models for that particular institution.

Limitations of this work include using data from only one healthcare system. The radiology titles in each healthcare system can vary slightly, thus there will likely be a need to retrain the models for each institution to ensure medical safety [62]. This limitation can also be addressed by incorporating more data from other institutions in future work. An additional limitation of this work was the use of a supervised learning mechanism. Consequently, we assumed that all labeled data are accurately labeled while there might be errors or disagreements in assigning the radiology titles, especially in the labels of minor classes. To address this limitation on supervised learning, we used the voting classifier to make an ensemble mechanism. The ensemble mechanism combined the prediction of all five models, each with different strengths and weaknesses. In future work, we will use unsupervised learning to identify the outliers in each class. In addition, we will develop a feedback mechanism for radiologists to return any wrongly classified instances. We hope that radiologists will help identify issues underlying the assignment of radiology titles and resolve disagreements in labeling.

V. CONCLUSION

The present study (1) demonstrated the ability of a newly developed system employing ML techniques to predict radiology titles based on EMR data and (2) identified the most important features in this process. Use of this type of system can guide radiology-referring physicians toward correct radiology title selection and it has the capacity to alert radiologists and radiology technologists to inconsistencies between radiology titles in exam orders and patients' conditions. Potential radiologist time savings with this AI system could be redirected to focus on critical tasks requiring their expertise.

APPENDIX A

Relevant attributes were selected in consultation with three radiologists working at the authors' healthcare center. The final selected attributes are described in Table 3.

APPENDIX B

Selected hyperparameters and their values and tuning ranges are reported in Table 4.

APPENDIX C

Table 5 indicates the distribution of numerical variables. Fig. 10 shows the distribution of categorical variables using pie charts.

APPENDIX D

Fig. 11 shows the confusion matrix in percentage for the voting classifier using the held-out test set.

APPENDIX E

Table 6 shows the percentages of missing values for all selected features before and after the application of the two imputation mechanisms (filling with previous records and MissForest).

REFERENCES

[1] M. B. De Basea, J. A. Salotti, M. S. Pearce, J. Muchart, L. Riera, I. Barber, S. Pedraza, M. Pardina, A. Capdevila, A. Espinosa, and E. Cardis, "Trends and patterns in the use of computed tomography in children and young adults in Catalonia—Results from the EPI-CT study," *Pediatr. Radiol.*, vol. 46, no. 1, pp. 29–119, Jan. 2016.

[2] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018.

[3] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Deep learning for natural language processing in radiology—Fundamentals and a systematic review," *J. Amer. College Radiol.*, vol. 17, no. 5, pp. 639–648, May 2020.

[4] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Creating artificial images for radiology applications using generative adversarial networks (GANs)—A systematic review," *Academic Radiol.*, vol. 27, no. 8, pp. 1175–1185, Aug. 2020.

[5] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás, "3D deep learning on medical images: A review," *Sensors*, vol. 20, no. 18, p. 5097, Sep. 2020.

[6] J. Zhao, X. Hou, M. Pan, and H. Zhang, "Attention-based generative adversarial network in medical imaging: A narrative review," *Comput. Biol. Med.*, vol. 149, Oct. 2022, Art. no. 105948.

[7] P. Kora, C. P. Ooi, O. Faust, U. Raghavendra, A. Gudigar, W. Y. Chan, K. Meenakshi, K. Swaraja, P. Plawiak, and U. R. Acharya, "Transfer learning techniques for medical image analysis: A review," *Biocybernetics Biomed. Eng.*, vol. 42, no. 1, pp. 79–107, Jan. 2022.

[8] H. Trivedi, J. Mesterhazy, B. Laguna, T. Vu, and J. H. Sohn, "Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's natural language processing algorithm," *J. Digit. Imag.*, vol. 31, no. 2, pp. 245–251, Apr. 2018.

[9] Y. H. Lee, "Efficiency improvement in a busy radiology practice: Determination of musculoskeletal magnetic resonance imaging protocol using deep-learning convolutional neural networks," *J. Digit. Imag.*, vol. 31, no. 5, pp. 604–610, Oct. 2018.

[10] P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. Ureña-López, and M. T. Martín-Valdivia, "Automatic medical protocol classification using machine learning approaches," *Comput. Methods Programs Biomed.*, vol. 200, Mar. 2021, Art. no. 105939.

[11] A. D. Brown and T. R. Marotta, "A natural language processing-based model to automate MRI brain protocol selection and prioritization," *Academic Radiol.*, vol. 24, no. 2, pp. 160–166, Feb. 2017.

[12] C. M. Sandino, E. K. Cole, C. Alkan, A. S. Chaudhari, A. M. Loening, D. Hyun, J. Dahl, A. S. Wang, and S. S. Vasanawala, "Upstream machine learning in radiology," *Radiol. Clinics*, vol. 59, no. 6, pp. 967–985, Nov. 2021.

[13] G. W. Boland, R. Duszak Jr., and M. Kalra, "Protocol design and optimization," *J. Amer. College Radiol.*, vol. 11, no. 5, pp. 440–441, 2014.

[14] A. Kalra, A. Chakraborty, B. Fine, and J. Reicher, "Machine learning for automation of radiology protocols for quality and efficiency improvement," *J. Amer. College Radiol.*, vol. 17, no. 9, pp. 1149–1158, Sep. 2020.

[15] A. D. Brown and T. R. Marotta, "Using machine learning for sequence-level automated MRI protocol selection in neuroradiology," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 5, pp. 568–571, May 2018.

[16] N. A. Muhammad, A. Sabarudin, N. Ismail, and M. K. A. Karim, "A systematic review and meta-analysis of radiation dose exposure from computed tomography examination of thorax-abdomen-pelvic regions among paediatric population," *Radiat. Phys. Chem.*, vol. 179, Feb. 2021, Art. no. 109148.

[17] R. Khorasani, "How IT tools can help improve current protocolling performance gaps," *J. Amer. College Radiol.*, vol. 8, no. 10, pp. 675–676, Oct. 2011.

[18] A. C. Yu, B. Mohajer, and J. Eng, "External validation of deep learning algorithms for radiologic diagnosis: A systematic review," *Radiol., Artif. Intell.*, vol. 4, no. 3, May 2022, Art. no. e210064.

[19] A. S. Chaudhari, C. M. Sandino, E. K. Cole, D. B. Larson, G. E. Gold, S. S. Vasanawala, M. P. Lungren, B. A. Hargreaves, and C. P. Langlotz, "Prospective deployment of deep learning in MRI: A framework for important considerations, challenges, and recommendations for best practices," *J. Magn. Reson. Imag.*, vol. 54, no. 2, pp. 357–371, Aug. 2021.

[20] S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, G. Fortino, and M. Buckland, "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis," *IEEE Access*, vol. 4, pp. 9145–9154, 2016.

[21] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[22] A. Gèron, *Hands-On Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2019, pp. 189–234.

[23] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. New York, NY, USA: Springer, 2012, pp. 35–49.

[24] M. Klug, Y. Barash, S. Bechler, Y. S. Resheff, T. Tron, A. Ironi, S. Soffer, E. Zimlichman, and E. Klang, "A gradient boosting machine learning model for predicting early mortality in the emergency department triage: Devising a nine-point triage score," *J. Gen. Internal Med.*, vol. 35, no. 1, pp. 220–227, Jan. 2020.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[26] H. Zafari, S. Langlois, F. Zulkernine, L. Kosowan, and A. Singer, "Predicting chronic obstructive pulmonary disease from EMR data," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–8.

[27] D. Tognetto, R. Giglio, A. L. Vinciguerra, S. Milan, R. Rejdak, M. Rejdak, K. Zaluska-Ogryzek, S. Zweifel, and M. D. Toro, "Artificial intelligence applications and cataract management: A systematic review," *Surv. Ophthalmol.*, vol. 67, no. 3, pp. 817–829, 2021.

[28] L. Liu, Y. Yu, Z. Fei, M. Li, F.-X. Wu, H.-D. Li, Y. Pan, and J. Wang, "An interpretable boosting model to predict side effects of analgesics for osteoarthritis," *BMC Syst. Biol.*, vol. 12, no. S6, pp. 29–38, Nov. 2018.

[29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.

[30] N. Al Turkestani, J. Bianchi, R. Deleat-Besson, C. Le, L. Tengfei, J. C. Prieto, M. Gurgel, A. C. Ruellas, C. Massaro, A. A. D. Castillo, and K. Evangelista, "Clinical decision support systems in orthodontics: A narrative review of data science approaches," *Orthodontics Craniofacial Res.*, vol. 24, pp. 26–36, Dec. 2021.

[31] A. Karthikeyan, A. Garg, P. K. Vinod, and U. D. Priyakumar, "Machine learning based clinical decision support system for early COVID-19 mortality prediction," *Frontiers Public Health*, vol. 9, May 2021, Art. no. 626697.

[32] M. D. Fathima and S. J. Samuel, "Improved Adaboost algorithm with regression imputation for prediction of chronic type 2 diabetes mellitus," in *Communication and Intelligent Systems*. Singapore: Springer, 2021, pp. 691–708.

[33] H. K. Sharvani, "Lung cancer detection using local energy-based shape histogram (LESH) feature extraction using AdaBoost machine learning techniques," *Int. J. Innov. Technol. Exploring Eng.*, vol. 9, no. 3, pp. 167–171, Jan. 2020.

[34] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern., C*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[35] X. Y. Liu and Z. H. Zhou, "Ensemble methods for class imbalance learning," *Imbalanced Learn., Found., Algorithms, Appl.*, vol. 1, pp. 61–82, Jun. 2013.

[36] S. Patidar, "Diagnosis of sepsis using ratio based features," in *Proc. Comput. Cardiol.*, Singapore, Sep. 2019, pp. 1–4.

[37] S.-F. Sung, L.-C. Hung, and Y.-H. Hu, "Developing a stroke alert trigger for clinical decision support at emergency triage using machine learning," *Int. J. Med. Informat.*, vol. 152, Aug. 2021, Art. no. 104505.

[38] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *Social Netw. Appl. Sci.*, vol. 3, no. 2, pp. 1–12, Feb. 2021.

[39] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[40] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *Int. J. Comput. Sci. Issues*, vol. 9, p. 272, Sep. 2012.

[41] G. Medic, M. K. Kließ, L. Atallah, J. Weichert, S. Panda, M. Postma, and A. El-Kerdi, "Evidence-based clinical decision support systems for the prediction and detection of three disease states in critical care: A systematic literature review," *FResearch*, vol. 8, p. 1728, Nov. 2019.

[42] D. Yao, J. Yang, and X. Zhan, "A novel method for disease prediction: Hybrid of random forest and multivariate adaptive regression splines," *J. Comput.*, vol. 8, no. 1, pp. 170–177, Jan. 2013.

[43] E. Alickovic and A. Subasi, "Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier," *J. Med. Syst.*, vol. 40, no. 4, pp. 1–12, Apr. 2016.

[44] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

[45] (Aug. 7, 2015). *Routine Adult Abdomen/Pelvis CT Protocols*. Accessed: Nov. 30, 2022. [Online]. Available: https://www.aapm.org/pubs/CTProtocols/documents/AdultAbdomenPelvisCT.pdf

[46] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[47] A. S. Acharya, A. Prakash, P. Saxena, and A. Nigam, "Sampling: Why and how of it," *India J. Med. Specialties*, vol. 4, no. 2, pp. 330–333, 2013.

[48] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012, p. 18.

[49] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[50] Y. Yang and Z. Xu, "Rethinking the value of labels for improving class-imbalanced learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19290–19301.

[51] H. Yuen. (2022). *MissForest 1.13*. Accessed: Feb. 2, 2023. [Online]. Available: https://pypi.org/project/MissForest/

[52] Z. C. Lipton, C. Elkan, and B. Narayanaswamy, "Thresholding classifiers to maximize F1 score," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2014, pp. 863–867.

[53] A. Ng. (2017). *Machine Learning Yearning*. Accessed: Feb. 24, 2023. [Online]. Available: http://www.mlyearning.org/
[54] M. Jaén-Vargas, K. M. Reyes Leiva, F. Fernandes, S. Barroso Gonçalves, M. Tavares Silva, D. S. Lopes, and J. J. Serrano Olmedo, "Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models," *PeerJ Comput. Sci.*, vol. 8, p. e1052, Aug. 2022.
[55] J. Wehrend, M. Silosky, F. Xing, and B. B. Chin, "Automated liver lesion detection in 68Ga DOTATATE PET/CT using a deep fully convolutional neural network," *EJNMMI Res.*, vol. 11, no. 1, pp. 1–11, Dec. 2021.
[56] M. R. Ahmed, M. S. Ahammed, S. Niu, and Y. Zhang, "Deep learning approached features for ASD classification using SVM," in *Proc. IEEE Int. Conf. Artif. Intell. Inf. Syst. (ICAIIS)*, Mar. 2020, pp. 287–290.
[57] M. H. Yap, B. Cassidy, J. M. Pappachan, C. O'Shea, D. Gillespie, and N. D. Reeves, "Analysis towards classification of infection and ischaemia of diabetic foot ulcers," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Jul. 2021, pp. 1–4.
[58] G. D. Rubin, "Costing in radiology and health care: Rationale, relativity, rudiments, and realities," *Radiology*, vol. 282, no. 2, pp. 333–347, Feb. 2017.
[59] M. Kohli, T. Alkasab, K. Wang, M. E. Heilbrun, A. E. Flanders, K. Dreyer, and C. E. Kahn, "Bending the artificial intelligence curve for radiology: Informatics tools from ACR and RSNA," *J. Amer. College Radiol.*, vol. 16, no. 10, pp. 1464–1470, Oct. 2019.
[60] S. N. Duda, N. Kennedy, D. Conway, A. C. Cheng, V. Nguyen, T. Zayas-Cabán, and P. A. Harris, "HL7 FHIR-based tools and initiatives to support clinical research: A scoping review," *J. Amer. Med. Inform. Assoc.*, vol. 29, no. 9, pp. 1642–1653, Aug. 2022.
[61] R. W. Filice and C. E. Kahn, "Biomedical ontologies to guide AI development in radiology," *J. Digit. Imag.*, vol. 34, no. 6, pp. 1331–1341, Dec. 2021.
[62] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, "The myth of generalisability in clinical research and machine learning in health care," *Lancet Digit. Health*, vol. 2, no. 9, pp. 489–492, 2020.

PEYMAN SHOKROLLAHI received the B.S. degree in electrical engineering from Shiraz University, Shiraz, Iran, in 2001, the M.A.Sc. degree in electrical and computer engineering from Ryerson University, Toronto, ON, Canada, in 2009, and the Ph.D. degree in biomedical engineering from the University of Toronto, Toronto, in 2017.

From 2012 to 2017, his research was on a magnetic resonance imaging compatible surgical robot for pediatric bone biopsy in collaboration with Hospital for Sick Children (SickKids). From 2017 to 2019, he collaborated with SickKids to develop a microrobot for sampling microbiomes. Since 2020, he has been a Research Scientist with Stanford University, developing an artificial intelligence system for medical imaging protocols.

Dr. Shokrollahi was a recipient of the Best Paper Award at the IEEE/ASME Transactions on Mechatronics on developing a magnetically actuated capsule for sampling microbiomes in the GI tract and the ISMRM Magna Cum Laude Merit Award at the International Society for Magnetic Resonance in Medicine on developing a clinical decision support system for MRI radiology titles using machine learning techniques and electronic medical records, in 2022.

JUAN M. ZAMBRANO CHAVES received the B.S. degree in biomedical engineering and the medical degree from Universidad de los Andes, Bogota, Colombia, in 2013 and 2017, respectively, and the M.S. degree in biomedical informatics from Stanford University, Palo Alto, California, USA, in 2020, where he is currently pursuing the Ph.D. degree in biomedical informatics.

He has previously published on multimodal opportunistic imaging, physical activity epidemiology and mathematical modeling, and evaluation of gene therapy. His research interests include the integration of multiple modalities for predictive models and clinical decision support in medicine.

Dr. Chaves honors and awards include the Stanford Knight-Hennessy Scholars, in 2018, and graduating Summa Cum Laude from both his undergraduate programs in biomedical engineering and in medicine.

JONATHAN P. H. LAM was born in Lake Charles, Louisiana, in 1990. He received the B.S. degree in biology from the University of Texas at Arlington, and the M.D. degree from The University of Texas Health Science Center at San Antonio. He completed the radiology residency at Mayo Clinic, Jacksonville, Florida, in 2020. He completed the Fellowship in body MRI at Stanford University, Palo Alto, California, in 2021.

From 2021 to 2022, he was a Radiologist with Baylor Scott and White, Dallas, TX, USA. He is currently with Advent Health, Orlando, FL, USA. He has presented more than 20 educational and scientific exhibits and several publications. His research interests include female pelvic floor disorders and advanced hepatobiliary MRI.

Dr. Lam's honors and awards include the Joe and Teresa Long Medical Scholar, the Paul Brand Scholarship, and the Mayo International Health Scholar.

AVISHKAR SHARMA received the M.D. degree from the Medical College of Wisconsin, in 2016. He completed the diagnostic radiology residency at the Rush University Medical Center, in 2021. He completed the Fellowship in body MRI at Stanford University, and the Imaging Informatics at the University of Maryland, in 2022. He received the board certification from the American Board of Radiology, in 2022, and the American Board of Imaging Informatics as a Certified Imaging Informatics Professional (CIIP), in 2021.

Since 2022, he has been the Director of the AI and Body Radiologist, Einstein Medical Center-Jefferson Health, and an Adjunct Clinical Instructor with the Division of Body MRI, Stanford University. His research interests include the translational impact on how AI tools help improve the quality and efficiency of the radiologist workflow.

Mr. Sharma holds positions on the Membership Committee for the Society of Imaging Informatics in Medicine (SIIM), the AI Subcommittee for Jefferson Health, and the Structure Committee for the American Board of Artificial Intelligence in Radiology (ABAIR).

DEBASHISH PAL received the B.S. degree in electrical engineering from the Indian Institute of Technology, Kharagpur, India, in 2003, and the M.S./Ph.D. degree in biomedical engineering from the Washington University in St. Louis, St. Louis, MO, USA, in 2008.

From 2008 to 2016, he worked on CT and PET image reconstruction while being employed with Philips and GE HealthCare. From 2016 to 2020, he moved to building a PET guided radiotherapy system with RefleXion. From 2020 to 2022, he was a Senior Imaging Scientist with GE HealthCare on the applications of machine learning in radiology. He is currently with Amazon Devices.

NAEIM BAHRAMI received the B.S. and M.S. degrees in electrical and biomedical engineering from the University of Tehran, Tehran, Iran, in 2006 and 2009, respectively, the M.S. degree in biomedical imaging from the University of California San Francisco, CA, USA, in 2013, and the Ph.D. degree in biomedical engineering from Virginia Tech/Wake Forest University, Winston-Salem, NC, USA, in 2016.

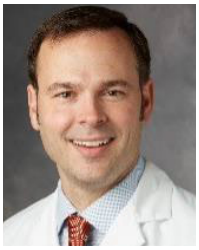
He pursued his career as a Postdoctoral Fellow with the University of California San Diego, working on neuro and cardiac MR imaging. Since 2019, he has been leading an artificial intelligence (AI) projects within GE HealthCare with regard to medical imaging.

Dr. Bahrami was a recipient of numerous awards from RSNA, ASNR, NASCI, ASFNR, and ISMRM over the course of his career for outstanding scientific contribution to the field of AI and machine learning in medical sciences.

AKSHAY S. CHAUDHARI received the B.S. degree in bioengineering from the University of California San Diego, San Diego, in 2012, and the Ph.D. degree in bioengineering and biomedical engineering from Stanford University, Stanford, CA, USA, in 2017.

From 2018 to 2019, he was a Research Scientist with Stanford University. From 2019 to 2020, he was an Instructor with the Radiology Department, Stanford University, where he has been an Assistant Professor with the Radiology Department of Biomedical Data Science, since 2020. He is the author of 40 articles and a holder of three patents. His research interests include developing new techniques for accelerated MRI acquisition and downstream image analysis, extracting prognostic insights from already-acquired CT imaging, and developing new multi-modal deep learning algorithms for healthcare that leverage computer vision, natural language, and medical records.

Dr. Chaudhary was a recipient of the W.S. Moore Young Investigator Award at the International Society for Magnetic Resonance in Medicine, in 2019, the Best Young Investigator Award at the 12th International Workshop on Osteoarthritis, in 2019, and the Best Emerging Investigator Award at the Imaging Elevated Symposium, in 2019.



ANDREAS M. LOENING received the B.S. degree in electrical science and engineering and the M.Eng. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1998 and 1999, respectively, and the Ph.D. degree in bioengineering and the M.D. degree from Stanford University, Stanford, CA, USA, in 2006 and 2008, respectively. He completed the residency training in radiology. He completed the Fellowship in body MRI at the Stanford School of Medicine, Stanford, in 2013 and 2014, respectively.

He was a Clinical Instructor of radiology with Stanford University, from 2014 to 2015, where he has been an Assistant Professor of radiology, since 2015. His research interests include molecular imaging in the abdomen, lymphatic imaging with MRI, clinical translation of new pulse sequences for use in body MRI, and artificial intelligence techniques for improving the radiologist workflow.