HHS Public Access

Author manuscript

J Am Coll Radiol. Author manuscript; available in PMC 2020 October 05.

Published in final edited form as:

J Am Coll Radiol. 2019 March ; 16(3): 336–343. doi:10.1016/j.jacr.2018.10.020.

Use of Machine Learning to Identify Follow-Up Recommendations in Radiology Reports

E Carrodeguas[a,b], **R Lacson**[a,b], **W Swanson**[b], **R Khorasani**[a,b]

[a]Harvard Medical School, Boston, Massachusetts;

[b]Center for Evidence-Based Imaging, Department of Radiology, Brigham and Women's Hospital, Brookline, Massachusetts

Abstract

Purpose: Assess follow-up recommendations in radiology reports, develop and assess traditional machine learning (TML) and deep learning (DL) models in identifying follow-up, and benchmark them against a natural language processing (NLP) system.

Methods: This HIPAA-compliant, IRB approved study, was performed at an academic medical center generating >500,000 radiology reports annually. 1,000 randomly-selected ultrasound, x-ray, computed tomography and magnetic resonance imaging reports generated in 2016 were manually reviewed and annotated for follow-up recommendations. Traditional machine learning (Support Vector Machines, Random Forest, Logistic Regression) and deep learning (Recurrent Neural Nets) algorithms were constructed and trained on 850 reports (training data), with subsequent optimization of model architectures and parameters. Precision, recall and F1-score were calculated on the remaining 150 reports (test data). A previously-developed and validated NLP system (iSCOUT) was also applied to the test data, with equivalent metrics calculated.

Results: 12.7% of reports had follow-up recommendations. The TML algorithms achieved F1 scores of 0.75 (Random Forest), 0.83 (Logistic Regression), and 0.85 (Support Vector Machine) on the test data. DL Recurrent Neural Nets had an F1 score of 0.71; iSCOUT also had an F1 score of 0.71. Performance of both TML and DL methods by F1-scores appeared to plateau after 500–700 samples while training.

Conclusion: TML and DL are feasible methods to identify follow-up recommendations. These methods have great potential for near real-time monitoring of follow up recommendations in radiology reports.

INTRODUCTION

While imaging's importance in medicine is undeniable, studies suggest that a portion of imaging tests might be redundant, inappropriate, or otherwise unnecessary [1]. Although slowing recently, imaging utilization continues to grow and up to 12% of radiology reports include a follow-up recommendation for additional imaging [2]–[4]. Unnecessary follow-up recommendations may result in wasteful imaging, though the scope is not well studied. Furthermore, improving techniques and image resolution in diagnostic radiology has led to the detection of even subtler new or incidental findings, raising questions of management. Although some follow-up recommendations are based on well-established guidelines (e.g.,

Fleischner Society guidelines for lung nodules [5,6], American College of Radiology whitepaper on managing incidental findings [7]), many remain based on clinical preferences and vary widely [8]. Further, guidelines may contain individual recommendations with heterogeneous strength of supporting evidence [9] or practitioners may inappropriately apply them [8].

Studying follow-up recommendations is key to understanding and monitoring their prevalence, identifying unwarranted variations in their use, tracking their impact, and developing initiatives to improve their use. Nonetheless, it remains challenging to extract recommendations from radiology reports, given their free text nature and the varied language radiologists use to make follow-up recommendations. Manual annotation can accurately identify these, but the necessary human-hours make it unsustainable to study sufficiently large samples over time. Furthermore, automated identification of these recommendations is critical to build quality control systems to ensure that these recommendations and further imaging are not missed.

While several groups have applied natural language and machine learning methods to analyzing radiology reports for follow-up recommendations, most have concentrated on traditional machine learning (TML) approaches; assessing reports on a sentence-by-sentence basis or via semantic and syntax analysis [10–12]. This work has been extremely promising, but to our knowledge no group has leveraged deep learning (DL) methods or pieced together algorithms capable of considering large sections of the report. DL relies on stacked nodes and layers to automatically extract features in a way believed to be analogous to human thinking. Such work would thus leverage ever-increasing computing power, and benefit from context cues indicating follow-up recommendation much like human readers can.

OBJECTIVE

We sought to: 1) assess follow-up recommendations in radiology reports, 2) develop and assess TML and DL models to identify follow-up recommendation in free text radiology reports, and 3) benchmark them against a previously-developed and validated natural language processing (NLP) system.

MATERIALS AND METHODS

Corpus Selection

This study was performed at a large tertiary healthcare institution and approved by its Institutional Review Board. Reports were extracted from the institution's Radiology Information System from among magnetic resonance imaging (MRI), computed tomography (CT), ultrasound (US) and x-ray (XR) studies performed 1/1/2016–12/31/2106. These totaled 547,495 unique reports from inpatient, outpatient and emergency department encounters. Buderer's formula for diagnostic test sample size calculation was applied (95% confidence interval, precision: 0.05, estimated specificity: 0.95), yielding a size of 150 testing samples [13]. A total 1,000 radiology text reports were extracted randomly to have an 850:150 training-to-testing split. In addition to the radiology text report content, study modality, study description, patient date of birth and patient gender were also collected.

Training Corpus Annotation

The text reports were manually labeled for the presence of any follow-up recommendations by an annotator (Emmanuel Carrodeguas). A second annotator (Whitney Swanson) independently labeled 400 reports to assess inter-rater agreement via kappa statistic. Reports with 1+ recommendations were labeled as containing follow-up, regardless of the number of recommendations. Recommendations for follow-up were defined as any phrase that might reasonably and explicitly suggest further imaging or procedural intervention (e.g., biopsy, colonoscopy, cystoscopy, etc.). Phrases suggesting clinical correlation or establishing uncertainty but not suggesting further steps were not considered follow-up. For example, “recommend short interval follow-up” and “MRI could be performed” constituted follow-up phrases; “clinical correlation recommended” or “structure could not be well visualized” did not. The annotated corpus was then randomly divided for training and validation using 850 reports (training data), and testing using 150 reports (test data).

Follow-Up Detection Using Information Extraction Methods

To establish a baseline for all machine learning methods, test data report text was analyzed using the Information from Searching Content with an Ontology-Utilizing Toolkit (iScout), a previously-validated NLP system [14]. For follow-up detection, this system relies on detecting “imaging terms” found close to “action terms” in text. For example, “Follow-up chest CT is recommended in 12 months” would be identified as a follow-up due to the proximity of “follow-up” (an action term) and “chest CT” (an imaging term). Imaging and action term lists are provided as Supplemental Materials.

Data Pre-processing and Feature Extraction for Machine Learning Models

The radiology report text data was preprocessed to remove special characters and capitalization. In our experience with manual annotation, actionable recommendations are almost always in the Impression sections of structured reports (even if also mentioned in Findings) and therefore we restricted analysis to the Impression section when possible to reduce computational complexity. For TML algorithms and DL, text data was converted into a Bag of Words (BoW) representation using the Scikit-Learn CountVectorizer utility, and scaled to frequencies using the TfidfTransformer utility [15]. CountVectorizer converts text into a matrix of occurrence counts, and TfidfTransformer scales each frequency using term-frequency times the inverse of document frequency (TF-IDF). CountVectorizer can treat multiple words as single units, applying n-grams to preserve sequence information. For TML algorithms, we included increasing n-grams as a tuned hyperparameter (see considered parameters in Supplemental Materials) and we applied 1–3-gram representations for our deep neural nets due to computational complexity. For long short-term memory (LSTM) recurrent neural network (RNN) DL approaches, text was transformed into numerical tokens using TensorFlow’s Tokenizer function [16].

TML Models

We trained Support Vector Machine (SVM), Random Forest (RF) and Logistic Regression (LR) models using the Scikit Learn Python application programming interface (API). All models were first evaluated with default hyperparameters using 10-fold cross validation.

Hyperparameters for all three models were tuned using Scikit-Learn's grid search (GridSearchCV), with optimized models re-evaluated using 10-fold cross validation. Optimized models were trained on the entirety of the training data and tested for generalizability on the test data. The searched hyperparameter space is provided as Supplemental Materials.

SVM is a classification algorithm that represents samples as points in space, using hyperplanes with associated wide margins to separate data points into categories [17]. Data points close to the dividing plane are called support vectors.

The RF classification algorithm relies on an ensemble of decision trees, averaging their results to make a decision [17]. Each decision tree learns to make predictions based on a random subset of input features, splitting data points through hierarchical nodes (e.g., by minimizing entropy) until reaching a classification.

LR is a classifier that relies on the logistic (a sigmoid function) of the weighted sum of features and a bias term to determine classification [17].

DL Models

Keras, an API with a TensorFlow backend, was used to construct long short-term memory recurrent neural network (LSTM-RNN) models [18]. Varied architectures were trained and evaluated using 5-fold cross validation, with the optimized model tested on the test data. Models started with an arbitrarily low number of hidden layers and nodes, then increased iteratively to estimate the optimized architecture. An early callback was implemented to avoid overfitting, ending training when the validation loss function (measure of model performance on validation data) began increasing. The last layer in each model was an activation function that turned the output into a binary classification (follow-up present or absent).

LSTM-RNNs consist of multiple layers of nodes. Each node is also connected to adjacent nodes within the same layer (giving the network a sequence component). Furthermore, each node can remember previous information that persists through training steps, giving them a "memory" component. Input data was represented as a sequence of word tokens and passed into an embedding layer that was actively trained to create a vector for each word in a 100-dimensional space, spatially related to similar concepts. The models were regularized using a dropout of 0.2, representing the proportion of nodes that were randomly excluded in each iteration to prevent overfitting.

Validation Framework

Models were developed using 10-fold cross validation for TML models and 5-fold cross validation for DL models (850 training and validation reports). All TML and DL models were trained and tested on a mid-2015 Apple Macbook Pro (CPU: 2.5 GHz Intel Core i7, RAM: 16GB 1600 MHz DDR3, GPU: AMD Radeon R9 M370X 2048MB); the Scikit-Learn and Keras/Tensorflow APIs utilized the CPU with no GPU usage. Time was measured using Python's datetime functionality and estimated for the development (grid search, cross validation of DL models) and final training of all models. Previously validated, iScout was

not modified for this work.[14] Scikit-Learn's metric utilities were applied to calculate Precision (True Positives / (True Positives + False Positives)), Recall (True Positives / (True Positive + False Negatives)) and F1 score (harmonic mean of the two). Each model was characterized by receiver operating characteristic (ROC) plots and the associated area under the curves (AUCs). Top models for each algorithm were chosen based on F1-scores, and these models were then trained on all the data minus a 10% validation set (85 reports) for early stopping callbacks. Therefore, each optimized model was re-trained on 765 reports.

Evaluating Models on Test Data

The re-trained models were then used to determine generalizability on the test data (150 reports), with Precision, Recall, and F1 scores reported. Finally, to test the impact of the size of training data on training, each model was re-trained on increasing subsets of training data (100, 200, 300, 400, 500, 600, 700 and 765) and re-tested on the hold-out test data. iScout was tested on the test data (150 reports).

RESULTS

The total corpus of reports consisted of 96 MR, 249 CT, 223 US, and 432 XR examinations. Annotations revealed 127 (12.7%) reports with follow-up recommendations. CT and MR modalities had the highest rate of follow-up recommendations at 25.3% (63 reports) and 18.8% (18 reports), respectively. XR and US both had lower rates at 6.3% (27 reports) and 8.5% (19 reports), respectively. 400 reports were annotated by two raters with a percentage agreement of 97.5% and a Kappa agreement score of 0.78 (95% Confidence Interval [CI] 0.39–1.0). The characteristics of the annotated training and test sets are shown in Table 1.

TML Algorithms Training and Optimization

Initial training and validation of TML algorithms with 10-fold cross validation showed an AUC of 0.96 for SVM, 0.92 for RF and 0.95 for LR algorithms. F1-scores prior to optimization were 0.0 for SVM, 0.43 for RF and 0.21 for LR. Metrics after optimization are shown in Table 2. SVM was optimized with a 7-degree polynomial kernel with a 1- to 3-gram BoW representation. RF had 100 estimators and optimized with a 1- to 6-gram BoW representation with 5 minimum samples per leaf. LR was optimized with a 1- to 6-gram BoW representation. ROC curves for all models are displayed in Figure 1. Grid search times (Table 2) for TML models ranged from 1 minute (LR) to 35 minutes (SVM) and were heavily dependent on the size of the hyperparameter space searched.

DL Models Training and Optimization

Training and validation metrics for DL algorithms are shown in Table 3. LSTM-RNN F1-scores ranged from 0.10 to 0.60, with 100 Nodes × 2 LSTM layers representing the best architecture. Training and validation times for DL models (Table 3) were all significantly longer than TML models, ranging from 30 minutes to 300 minutes.

Testing Models on Hold-Out Data

Optimized models and iScout were tested on the test data to evaluate generalizability to unseen data (Table 4). The tuned SVM model had the best performance in terms of all

metrics with a precision of 0.88, a recall of 0.82 and an F1-score of 0.85. LSTM-RNN DL model had an F1 score of 0.71. IScout also had an F1-score of 0.71.

Sensitivity to Training Sample Size

Optimized models were further trained with increased samples sizes to establish sensitivity to sample sizes. Models were trained on increasing samples of 100 report intervals, with average F1-scores for traditional and DL approaches recorded. Compared curves are shown in Figure 2.

DISCUSSION

Follow-up recommendations are an important component of an actionable radiology report to inform clinical decision-making and care planning for patients. Follow-up recommendations are common (10–12% of reports), but a substantial portion may not be evidence-based [19,20], with significant unexplained variation among radiologists [8]. As such, assessing follow-up recommendations provides an excellent opportunity for improving care and enhancing the value of radiologist's input in devising an optimal patient care plan. However, radiology reports with follow-up recommendations are difficult to identify, in part due to their free text nature and their lack of standardized structure and content. Therefore, automated identification of reports containing follow-up recommendations would constitute a powerful tool for research and quality improvement and provide opportunities to ensure and track appropriate follow-up for a broad range of powerful clinical applications. To our knowledge, this is the first study assessing TML and DL methods in identifying follow-up recommendations in radiology reports. We find that all of these machine learning methods can identify reports with follow-up recommendations, with optimized models comparable to a previously validated system.

First, our experiments indicate the feasibility of training machine learning algorithms on textual data. A challenge of machine learning is obtaining adequate samples for training and validation. Although large datasets exist in electronic health records (EHRs), most radiology reports are free text and would require manual annotation prior to training. Current approaches to other machine learning problems use large databases (tens of thousands), an approach that is onerous for radiology reports. However, our study demonstrates stable results with less data (training on 500–700 reports), indicating that both TML and DL are feasible methods for assessing follow-up recommendations in radiology reports.

Secondly, our study highlights the ease of design and minimal preprocessing needed to achieve accurate results with machine learning. Many current information extraction systems (such as iScout) depend on carefully constructed algorithms and methods to achieve different tasks [14]. As such, they are not flexible and once constructed require expert input to adapt to changing targets. For example, utilizing iScout for follow-up detection required the careful curation of imaging and follow-up term lists. Likewise, previous machine learning approaches to information extraction have relied on extensive preprocessing (removing stop words, linking ideas, word stemming). Here we show that with minimal preprocessing and readily available "out-of-the-box" machine learning classifiers, we could match carefully engineered methods. This represents a useful option for practicing

radiologists and radiology groups seeking to track and monitor their own recommendation trends.

Thirdly, our study assesses both traditional and deep machine learning approaches and compares them to a previous rule-based NLP tool. Traditional models (SVM, RF, LR) performed well after optimization. Interestingly the DL method's performance appears to plateau with increasing training size, but it is possible that this is a function of the restricted architecture tested. Unlike previously-developed NLP tools, DL and TML can automatically select and combine features without pre-determined rules created manually by human domain experts. However, rule-based, TML and DL approaches can all be viably utilized for NLP, specifically extracting information from structured reports. It remains possible that with larger data sets and more complex architectures, DL methods would provide better classification, as DL continues to break barriers in other fields, including pathology, cancer prognosis, and drug discovery [21–24]. Similar results have been seen in radiology, with DL methods being applied broadly from lesion segmentation to diagnostic predictions from brain imaging [25–27]. As such, ML's role is likely to continue increasing in medicine and radiology.

Finally, our study opens the possibility to use automatically extracted follow-up recommendations in clinical practice. Automated systems integrated in an EHR facilitate coordination between various providers regarding collaborative care plans[28]. For reports containing further recommendations, a system can monitor and inform providers so that necessary diagnostic follow-up is not delayed. In addition, identifying follow-up recommendations within one's own practice provides an opportunity to quantify between-providers variations in follow-up recommendations for specific findings, as well as variations resulting from clinical uncertainty[29].

From a population health perspective, an integrated system built on our work could, for example, validate radiologist recommendations against accepted standards, providing them real-time feedback. Over time, such a system could track adherence to these recommendations and even quantify the impact on patient outcomes. In addition, a system that could automatically monitor follow-up recommendations could be useful for reporting in quality improvement activities and federal regulatory requirements. For example, follow-up imaging recommendations for abdominal findings (e.g., pancreatic cysts) can be included in the Center for Medicare and Medicaid Services Merit-based Incentive Payment System Measures and the system could facilitate reporting [30].

Our study has several limitations. First, we restricted our algorithms to the Impressions sections of text reports. Although most follow-up recommendations are indicated there, other contextual information (setting of study, patient characteristics, indication of study) is often useful for human raters and could have improved our models. Furthermore, while our limited preprocessing and sample size highlight the feasibility of these models, they likely also acted to limit performance. In ongoing applications of these algorithms, fine tuning these parameters could improve accuracy. Third, due to computational constraints we limited the architecture and fine tuning of our models, potentially disproportionately impacting the DL models. Fourth, our model only identified the presence of a follow-up recommendation,

extracting no information about the incidental finding or what was recommended. Finally, the study single institution has a limited number of radiologists, impacting generalizability of our models to other institutions.

In conclusion, machine learning models appear to be useful tools, capable of detecting follow-up recommendations with minimal training data in a way that may be equally applied to retrospective analyses or potentially to near real-time monitoring of radiology reports. Future work can: leverage optimized models, expanding on current architectures with more computing power; focus on training and modifying models to extract further information regarding the nature of incidental findings and subsequent recommendations; and test real-world generalizability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This study was supported in part by grant R01HS024722 from the Agency for Healthcare Research and Quality (AHRQ). We thank Laura E. Peterson for assistance in the preparation of this manuscript.

References

[1]. Hillman BJ and Goldsmith JC, “The uncritical use of high-tech medical imaging.,” N. Engl. J. Med, vol. 363, no. 1, pp. 4–6, 7 2010. [PubMed: 20573920]

[2]. Smith-Bindman R, Miglioretti DL, and Larson EB, “Rising Use Of Diagnostic Medical Imaging In A Large Integrated Health System: The use of imaging has skyrocketed in the past decade, but no one patient population or medical condition is responsible,” Health Aff. (Millwood), vol. 27, no. 6, pp. 1491–1502, 2008. [PubMed: 18997204]

[3]. Lang K, Huang H, Lee DW, Federico V, and Menzin J, “National trends in advanced outpatient diagnostic imaging utilization: an analysis of the medical expenditure panel survey, 2000–2009,” BMC Med. Imaging, vol. 13, no. 1, p. 40, 11 2013. [PubMed: 24279724]

[4]. Sistrom C, Dreyer K, Dang P, Weilburg J, Boland G, Rosenthal D, and Thrall J, “Recommendations for Additional Imaging in Radiology Reports: Multifactorial Analysis of 5.9 Million Examinations,” Radiology, vol. 253, no. 2, pp. 453–461, 2009. [PubMed: 19710005]

[5]. McDonald JS, Koo CW, White D, Hartman TE, Bender CE, and Sykes AMG, “Addition of the Fleischner Society Guidelines to Chest CT Examination Interpretive Reports Improves Adherence to Recommended Follow-up Care for Incidental Pulmonary Nodules,” Acad. Radiol, vol. 24, no. 3, pp. 337–344, 2017. [PubMed: 27793580]

[6]. Naidich DP, Bankier A. a., MacMahon H, Schaefer-Prokop CM, Pistolesi M, Goo JM, Macchiarini P, Crapo JD, Herold CJ, Austin JH, and Travis WD, “Recommendations for the Management of Subsolid Pulmonary Nodules Detected at CT: A Statement from the Fleischner Society,” Radiology, vol. 266, no. 1, pp. 304–317, 2013. [PubMed: 23070270]

[7]. Berland LL, Silverman SG, Gore RM, Mayo-smith WW, Megibow AJ, Yee J, Brink JA, Baker ME, Federle MP, Foley WD, Francis IR, Herts BR, Israel GM, Krinsky G, Platt JF, Shuman WP, and Taylor AJ, “Managing Incidental Findings on Abdominal CT : White Paper of the ACR Incidental Findings Committee,” JACR, vol. 7, no. 10, pp. 754–773, 2010. [PubMed: 20889105]

[8]. Ip IK, Mortele KJ, Prevedello LM, and Khorasani R, “Focal cystic pancreatic lesions: assessing variation in radiologists’ management recommendations.,” Radiology, vol. 259, no. 1, pp. 136–141, 4 2011. [PubMed: 21292867]

[9]. Ransohoff DF, Pignone M, and Sox HC, “How to Decide Whether a Clinical Practice Guideline Is Trustworthy,” Jama, vol. 309, no. 2, p. 139, 2013. [PubMed: 23299601]

[10]. Xu Y, Tsujii J, and Chang EI, “Named entity recognition of follow-up and time information in 20 000 radiology reports,” J. Am. Med. Inform. Assoc, vol. 19, no. 5, pp. 792–799, 2012. [PubMed: 22771530]

[11]. Dutta S, Long WJ, Brown DFM, and Reisner AT, “Automated Detection Using Natural Language Processing of Radiologists Recommendations for Additional Imaging of Incidental Findings,” YMEM, vol. 62, no. 2, pp. 162–169, 2013.

[12]. Pons E, Braun LMM, Hunink MGM, and Kors JA, “Natural Language Processing in Radiology: A Systematic Review,” Radiology, vol. 279, no. 2, pp. 329–343, 2016. [PubMed: 27089187]

[13]. Malhotra RK and Indrayan A, “A simple nomogram for sample size for estimating sensitivity and specificity of medical tests,” Indian J. Ophthalmol, vol. 58, no. 6, pp. 519–522, 7 2010. [PubMed: 20952837]

[14]. Lacson R, Andriole KP, Prevedello LM, and Khorasani R, “Information from Searching Content with an Ontology-Utilizing Toolkit (iSCOUT),” J. Digit. Imaging, vol. 25, no. 4, pp. 512–519, 8 2012. [PubMed: 22349993]

[15]. Pedregosa F, Weiss R, and Brucher M, “Scikit-learn : Machine Learning in Python,” J. Mach. Learn. Res, vol. 12, pp. 2825–2830, 2011.

[16]. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X, Brain G, Osdi I, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, and Zheng X, “TensorFlow : A System for Large-Scale Machine Learning This paper is included in the Proceedings of the TensorFlow : A system for large-scale machine learning,” USENIX Symp. Oper. Syst. Des. Implement, 2016.

[17]. Géron A, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O’Reilly Media, 2017.

[18]. F. and others Chollet, “Keras.” GitHub, 2015.

[19]. Lacson R, Prevedello LM, Andriole KP, Gill R, Lenoci-Edwards J, Roy C, Gandhi TK, and Khorasani R, “Factors associated with radiologists’ adherence to fleischner society guidelines for management of pulmonary nodules,” J. Am. Coll. Radiol, vol. 9, no. 7, pp. 468–473, 2012. [PubMed: 22748786]

[20]. Bobbin MD, Ip IK, Sahni VA, Shinagare AB, and Khorasani R, “Focal Cystic Pancreatic Lesion Follow-up Recommendations After Publication of ACR White Paper on Managing Incidental Findings,” J. Am. Coll. Radiol, vol. 14, no. 6, pp. 757–764, 2017. [PubMed: 28476609]

[21]. Kourou K, Exarchos TP, Exarchos KP, V Karamouzis M, and Fotiadis DI, “Machine learning applications in cancer prognosis and prediction,” CSBJ, vol. 13, pp. 8–17, 2015. [PubMed: 25750696]

[22]. Lavecchia A, “Machine-learning approaches in drug discovery : methods and applications,” Drug Discov. Today, vol. 20, no. 3, pp. 318–331, 2015. [PubMed: 25448759]

[23]. Jordan MI and Mitchell TM, “Machine learning: Trends, perspectives, and prospects,” Science (80-.), vol. 349, no. 6245, p. 255 LP–260, 7 2015.

[24]. Deo RC, “Machine Learning in Medicine,” Basic Sci. Clin, pp. 1920–1930, 2015.

[25]. Kohli M, Prevedello LM, Filice RW, and Geis JR, “Implementing machine learning in radiology practice and research,” Am. J. Roentgenol, vol. 208, no. 4, pp. 754–760, 2017. [PubMed: 28125274]

[26]. Zhou Y, Amundson PK, Yu F, Kessler MM, Benzinger TLS, and Wippold FJ, “Automated Classification of Radiology Reports to Facilitate Retrospective Study in Radiology,” J. Digit. Imaging, vol. 27, no. 6, pp. 730–736, 2014. [PubMed: 24874407]

[27]. Bentley P, Ganesalingam J, Lalani A, Jones C, Mahady K, Epton S, Rinne P, Sharma P, Halse O, Mehta A, and Rueckert D, “NeuroImage : Clinical Prediction of stroke thrombolysis outcome using CT brain machine learning,” NeuroImage Clin, vol. 4, pp. 635–640, 2014. [PubMed: 24936414]

[28]. K. R Hammer MM, Kapoor N, Desai S, Sivashanker K, Lacson R, Demers JP, "Adoption of a closed loop communication tool to establish and execute a collaborative follow up plan for incidental pulmonary nodules," AJR Am J Roentgenol, 2018.

[29]. Lacson R, Odigie E, Wang A, Kapoor N, Shinagare A, Boland G, Khorasani R, "Multivariate Analysis of Radiologists' Usage of Phrases that Convey Diagnostic Certainty.," Acad Radiol, 2018.

[30]. CMS, "Medicare Program; CY 2018 Updates to the Quality Payment Program; and Quality Payment Program: Extreme and Uncontrollable Circumstance Policy for the Transition Year. Final rule with comment period and interim final rule with comment period.," Fed. Regist, vol. 82, no. 220, pp. 53568–54229, 11 2017. [PubMed: 29232069]

TAKE-HOME POINTS

- Follow-up recommendations are an important component of an actionable radiology report to inform clinical decision-making, present in 12.7% of radiology reports.
- Machine learning algorithms utilized the Impression section of radiology reports with minimal textual pre-processing for identifying reports with follow-up recommendations.
- Traditional machine learning and deep learning algorithms can identify radiology reports with follow-up recommendations.
- Automatic identification of follow up recommendations could have wide implications for real time monitoring of reports to create alerts for actionable findings to ensure collaborative care plans are developed with other providers.

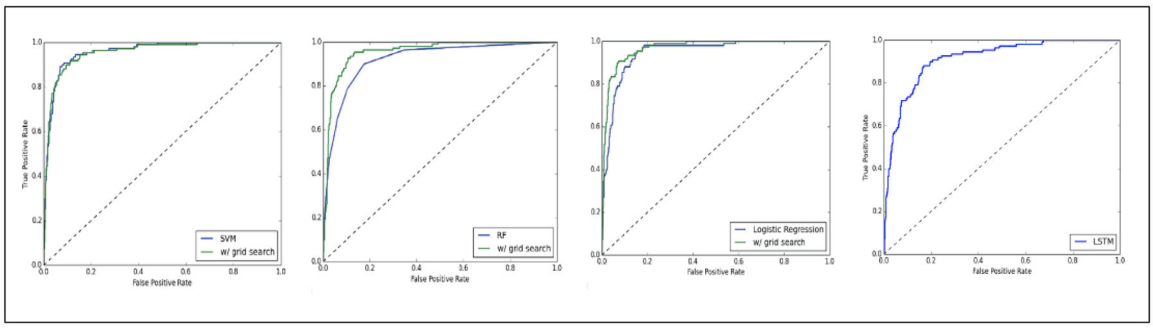

Figure 1:

(A-B) Receiver operating characteristic (ROC) curves for traditional machine learning models (blue) and optimized parameters (green). Subplots represent Support Vector Machines (A), Random Forest (B) and Logistic Regression (C). (D) Receiver operating characteristic (ROC) curve for top long short term memory deep learning architecture (100 nodes × 2 layers).

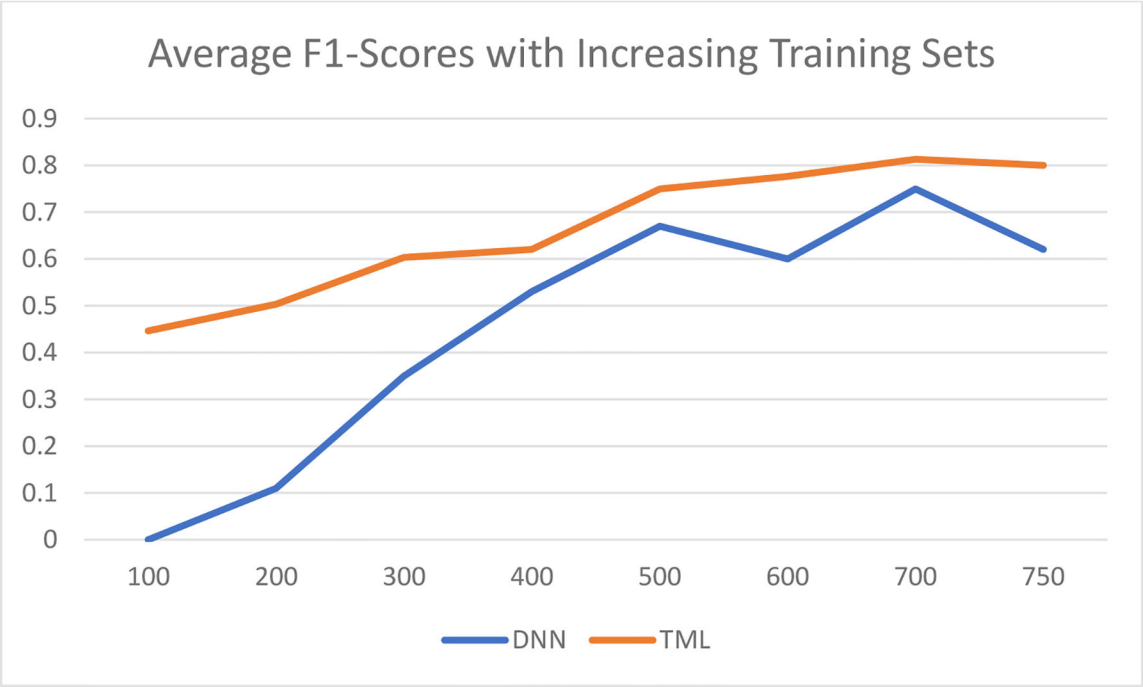

Figure 2:
Average F1 scores for deep learning (DL, blue) and traditional machine learning (TML, orange) models with increasing training data (100 to 750 samples).

Table 1:

Characteristics of training and testing corpus

	Training and Validation (N=850)	**Testing (N=150)**
Follow-Up Recommendations	110 (12.9%)	17 (11.3%)
Modality		
X-ray	365 (42.9%)	67 (44.7%)
Computed tomography	217 (25.5%)	32 (21.3%)
Magnetic resonance imaging	82 (9.6%)	14 (9.3%)
Ultrasound	186 (21.9%)	37 (24.7%)
Patient Gender		
Female	487 (57.3%)	89 (59.3%)
Male	363 (42.7%)	61 (40.7%)
Patient Age	56.3 (Standard Deviation: 19.3)	54.8 (Standard Deviation: 18.6)

Table 2:

Traditional machine learning metrics from 10-fold cross validation on training data

Model	Precision	Recall	F1-Score	AUC	Grid-Search Time
SVM	0.77	0.71	0.74	0.96	33 minutes
Random Forest	0.77	0.75	0.76	0.96	5 minutes
Logistic Regression	0.61	0.91	0.73	0.97	1.2 minutes

AUC= area under the ROC curve; SVM= support vector machine

Table 3:

5-fold cross validation deep learning algorithm architectures and metrics on training data

Model	Architecture	Precision	Recall	F1-Score	AUC	Training Time
LSTM-RNN	50 Nodes × 2 LSTM	0.71	0.44	0.54	0.89	30 minutes
	100 Nodes × 2 LSTM	**0.68**	**0.54**	**0.60**	**0.91**	**30 minutes**
	100 Nodes × 5 LSTM	0.59	0.46	0.52	0.88	90 minutes
	500 Nodes × 2 LSTM	0.68	0.54	0.60	0.85	120 minutes
	500 Nodes × 5 LSTM	0.52	0.26	0.35	0.79	300 minutes
	1000 Nodes × 2 LSTM	1.00	0.05	0.10	0.72	300 minutes

AUC= area under the curves; LSTM-RNN= long short-term memory recurrent neural network; bolded row indicates top scoring architecture

Table 4:

Optimized model testing on test data

Model	Precision	Recall	F1-Score	Training Time
SVM	**0.88**	**0.82**	**0.85**	**0.7 seconds**
Random Forest	0.80	0.71	0.75	5.5 seconds
Logistic Regression	0.79	0.88	0.83	1.2 seconds
LSTM-RNN	0.91	0.59	0.71	15 minutes
iScout	0.71	0.71	0.71	

SVM= support vector machine; LSTM-RNN= long short-term memory recurrent neural network; bolded row indicates top scoring architecture