Automation of Classification of MRI Brain Scans: An Emerging Trend

H.A.Aravinda, Prathyaj Mantha, Shilpa Das
Department of Computer Science and Engineering
School of Engineering and Technology,
Jain Deemed-to-be University, Bangalore, Karnataka, India

Abstract : The ordering of Magnetic resonance imaging (MRI) brain scans following American College of Radiology (ACR) guidelines showed a higher percentage of brain abnormalities compared to scans that do not. As the process of manually labelling patient orders obtained from a local tertiary hospital in accordance to ACR guidelines is intensive and time consuming, this study aims to develop predictive machine learning models; Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and XGBoost (XGB), to automate the classification process through text mining methods and derive insights that are useful for future clinical decision-making and resource optimization. Using 1,924 observations as the labelled training data, RF and XGB were found to be the best performing robust models with ROC values of 0.9459 and 0.9508 respectively on the validation set (481 observations). Further exploration into the interpretability of black-box algorithms using the model agnostic LIME (Local Interpretable Model-Agnostic Explanations) framework was used to generate further insights for decisions made using a separate XGB model with respect to individual patients. The LIME framework is a significant first step towards the development of a comprehensive decision support system for patient-level decisions in the ordering of MRI scans.

I. INTRODUCTION

The method of manually labelling patient orders obtained from a local tertiary hospital is challenging and time consuming, the process of automating the classification process through text mining methods and derive insights that are helpful for future clinical decision-making and resource improvement. Further exploration into the interpretability of black-box algorithms using the model agnostic LIME (Local Interpretable Model-Agnostic Explanations) framework are used to generate further insights for choices made using a separate XGB model with respect to individual patients. The LIME framework is a vital first initiative towards the development of a comprehensive decision support system for patient-level choices within the ordering of MRI scans.

Here are limitations of current work:

i. NLP machine learning models (RF, SVM, and k-nearest neighbor) on MRI brain analysis with the distinctive outcomes of choosing examination protocols and deciding the priority of MRI examinations have been done. Accuracies of these models were found to be above 80% and RF score was found to be best performing.
ii. Similar study was conducted wherein the development of a machine learning system constructed using an open source SVM framework (LIBSVM) to classify free text MRI knee reports according to normal or abnormal findings. The SVM classifier was trained using data from two different healthcare organizations and evaluated using standard metrics such as accuracy, ROC, recall, precision and F1 score on each of the same data and cross organization's data. Accuracies and F1 scores of 0.85 and 0.90 respectively were reported on train-test using the same group of data
iii. Another study carried out tells the implementation of machine learning algorithms for unstructured radiology reports would be the similar use of LIBSVM to detect tumor status from MRI reports producing an F1 score of 0.81 and ROC above 0.9 for tumor status determination.

II. COMPARITIVE STUDY

Previous studies have been found that MRI brain orders in adherence to ACR guidelines indicate a higher probability of finding an abnormality on the scan. This might translate to the right utilization of the MRI equipment and eliminating bottlenecks from waiting times because of inessential orders which do not follow ACR guidelines. NLP revolves around machine learning techniques and rule-based approach to extract specific findings.The usage of machine learning has been emerging in the radiological field because of its ability to automatically determine complex patterns in datasets and having ability to assist radiologists with intelligent clinical decision support tools on radiology data including text analysis of radiology reports using NLP.

Accuracies by the models were reported to be above 80% and RF was found to be the best performing model in terms of ROC although the scores were not reported.

III. DESIGN METHODOLOGY

A. Study Data

i. The first outcome is to ascertain the best classifying algorithm to determine if the MRI scan indications provided falls under the ACR guidelines. This has implications on the clinical workflow, overuse and appropriate prioritization of expensive and limited MRI resources to more compelling indications. The secondary outcome is to identify the key terms that distinguish between adherence to ACR or not (using LIME algorithm).
ii. The secondary outcome would be useful for the development of an MRI orders workflow decision support system to identify urgent and non-urgent scans for radiologists. The patient's healthcare data entered from hospital entries is initially stored in the enterprise data warehouses.

The extracted data contains numerous fields such as the patient's case number, patient demographics such as age and gender, textual inputs by clinicians such as the 'Indication for MRI' containing the justification for ordering MRI scans, 'Diseases' which is the existing medical condition of the patient, as well as the 'Diagnosis' containing the finalized outcome after the patient's MRI images have been examined and the report generated. An additional field containing the output labels on whether there is adherence to ACR guidelines ('1' or '0') is based on the 'Indication for MRI' text rows which are read and manually tagged by the clinicians for the purpose of this study. Together, these two tagged fields were used to train and validate the machine learning algorithms. The extracted data consists of 2,470 MRI brain order entries of patients with brain MRI scans performed from the year 2006 to 2013 at a tertiary Children and Women's Hospital, with patient age ranging from one-day-old new borns to young adults of 20 years of age.

Table I. Descriptive Statistics of Data

Parameter	Values
Age (Years):	
Max	20
Min	0.003
Mean	8.13
Median	9
3rd Quartile	13
1st Quartile	3
Gender:	
Male	1,310
Female	1,095
ACR Labels:	
ACR ('1')	1,851
Non-ACR ('0')	554

B. Data Cleaning and Preprocessing

In the pre-processing stage, the free-text fields containing the patient's initial findings prior to the MRI scan are cleansed using the below methodology. The cleaning steps involve

i. Standardizing words from British to American spelling due to American guidelines being referenced in this study, correction of spelling errors from clinical entries, as well as removal of sentences containing negated words.
ii. Spelling correction is not adjusted once the medical abbreviations used by clinicians are identified as mistakes by machine-driven spellcheckers. Words that are too poorly spelled were amended for machine readability.
iii. For medical conditions that clinicians feel are vital to be are ruled out, and have indicated this as such, in the MRI request, the removal of sentences containing negated words becomes necessary. This can be as result of text mining algorithms that are supposed to extract relevant insights from the words are unable to detect negated medical conditions effectively.
iv. Negation is also well-known major source of poor precision in medical information retrieval systems. The removal of those sentences constitutes a data loss of 65 out of 2,470 cases in this dataset, leaving 2,405 rows of data for training and validation. After splitting the remaining 2,405 observations with a training-validation ratio of 80-20, 1,924 observations from the training data and 481 from the validation data are cleaned separately to simulate real-world scenarios where new words in the validation set are absent from the train set and are dropped.

v. The text is converted into a corpus and normalized by standardizing to lowercase, removing stop words, punctuation, white spaces, numbers, and lastly stemming to reduce words to their base form. Once the corpus has been cleaned, it can be converted into a document-term-matrix (dtm) consisting of a sparse matrix with each single word or phrase existing as a variable. This is also known as the bag-of-words model in NLP.

vi. Based on the required ngram = (1,3), every variable can consist of a single word e.g. headache, two words phrase e.g. chronic headache, or three words phrase e.g. persistent chronic headache. Using the 'tm' package in R, the values of the matrix may be specified based on the frequency of occurrence of a word in a particular document which is known as term frequency (TF), or by multiplying TF with the inverse document frequency (IDF), which is the log of the total number of documents divided by the number of documents a particular term appears.

The removal of sparse terms is vital as it reduces the number of low occurrence words within the data matrix that can contribute to poorer prediction algorithm in a high dimensional matrix. While removing these terms may also enable improvement of the processing speed during modelling, this comes at the cost of information loss when the variables are dropped. The algorithms were evaluated using the remove sparse matrix term set at 99.9%, which removes rare occurring words or phrases that are absent across 99.9% of the documents. This brings the total number of variables down from 19,489 to 2,829, thereby allowing the algorithms to process the dataset much more rapidly.

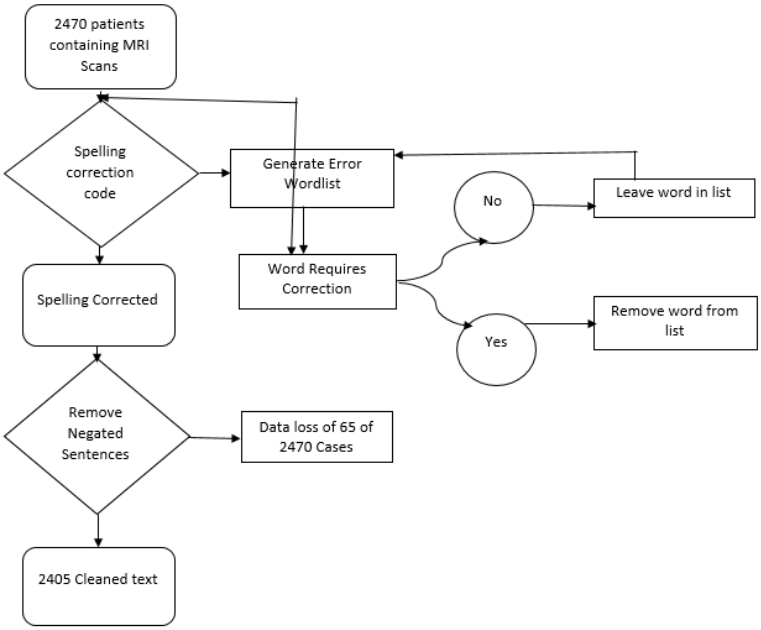

Fig 1: Flowchart showing the pre-processing algorithm

C. Algorithms Used

i. Based on commonly used techniques in NLP and the ability to provide insights and explanations from the data, the machine learning algorithms chosen for this study are Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and XGBoost (XGB). The logistic regression model is a statistical model used to measure the relationship between a binary or multinomial response against several predictors.

ii. In the case of a binary outcome (0 or 1), the data can be modelled using a binomial distribution following a probabilistic output. The logic in this model is linear in the explanatory variables. Logistic regression is a useful tool towards analyzing many sorts of data particularly in the healthcare domain.

iii. Bagging or bootstrap aggregation is a widely used technique for reducing the variance of prediction models such as trees. RF algorithm further improves bagging by modifying the variable selection in the node splitting procedure. Both bagging and RF construct each tree using different bootstrap samples from the original data.

iv. Tree boosting is a highly effective and widely used machine learning method that has been shown to give state of the-art results on many standard classification problems. An improvement to tree boosting is XGB, a scalable tree boosting algorithm that can recognize and handle sparse data, making use of a weighted quantile sketch for approximate tree learning, additional regularized objective function that helps avoid over fitting, as well as an enabled cross validation function.

IV. Proposed System

The algorithms proposed here can be evaluated and implemented for clinical audit purposes. In order to be useful for supporting the decisions of radiologists in the evaluation of clinical case notes that were recorded by the attending physicians, the machine learning models would have to be able to generate insights for each unique patient. Variable importance and partial dependency plots enable identification of significant variables that affect the models as well as the variables relationship with the output, extracting factors that drive the model prediction for a unique observation is more important for supporting the decision processes of radiologists. Despite its widespread use most machine learning models remain black boxes. Understanding the reasons behind the predictions made by such models is a vital step to earn the trust of the clinicians. The complex machine learning models that can be applied in critical applications such as healthcare where being able to understand and validate a model is essential. In order to deal with the above limitations of machine learning models such as RF and XGB, a proposed package called Local Interpretable Model-Agnostic Explanations (LIME) was also explored in this study. This is an explanation technique that explains the predictions of any classifier by learning an interpretable model locally around the prediction. The advantages of this model include its ability to be used by any type of model as well as explain unstructured data such as text or images. Of the four models shown in RF and XGB had the best overall performance with accuracy of 89.81% and 90.64%, F1-scores (weighted average of the precision and recall) 93.44% and 94.01%, ROC values of 0.9459 and 0.9508, and relatively high specificity of 74.77% respectively. This would imply that the two models have comparable performance and were able to differentiate the number of '1' and '0' cases relatively well despite an imbalanced dataset. Both LR and SVM on the other hand performed poorly in terms of model performance with ROC of 0.5946 and 0.6699 as well as specificity of 64.86% and 18.92% respectively. SVM in particular was unable to identify the non-ACR cases as evident by its low specificity of 18.92%. As TF-IDF with 99.9% sparse terms removal was the main method used to specify the data matrix, the algorithms were also subject to various conditions in order to determine how their performance would be affected.

AbbreviationsandAcronyms

MRI: Magnetic resonance imaging
ACR: American College of Radiology
LR: Logistic Regression
SVM: Support Vector Machine
RF: Random Forest
XGB: XGBoost
EHR: Electronic Health Record
NLP: Natural language processing
TF: Term frequency
TF-IDF: TF with the inverse document frequency
LIME: Local Interpretable Model-Agnostic Explanations

Table II: Performance metrics using TF-IDF with 99.9% Sparse Terms Removed

	Accuracy	ROC	Specificity	Precision	Recall	F1Score
LR	0.57	0.59	0.65	0.84	0.54	0.66
SVM	0.81	0.67	0.19	0.80	1.00	0.90
RF	0.90	0.95	0.75	0.93	0.94	0.93
XGB	0.91	0.95	0.75	0.93	0.95	0.94

As seen from the ROC plots in Figure using TF as the weighting factor improves the ROC of SVM (see Fig. 2b), while reducing the sparse terms from 99.9% to 99% improves the performance of the two poorer performing models LR and SVM (as shown in Fig.2c). In all four scenarios, both RF and XGB have been shown to be robust as their ROC values are consistently above 0.90. For ease of exposition, the variable importance and partial dependencies which are used to ascertain the significant predictors in the prediction as well as their relationship with the predicted outcome will be based on the better performing algorithms RF and XGB. Some examples of the top few stemmed words are 'headach', 'migrain', 'persist' and 'chronic'. Although XGB uses a different package than RF ('train' package on 'caret' in its R library) and evaluates the variables based on contribution to model's accuracy.

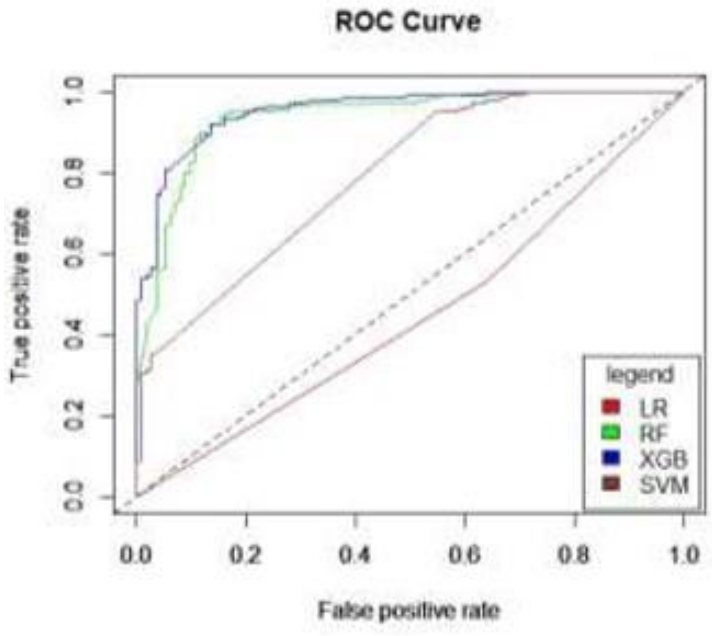

Fig 2(a): TF-IDF with 99.9% sparse terms removed

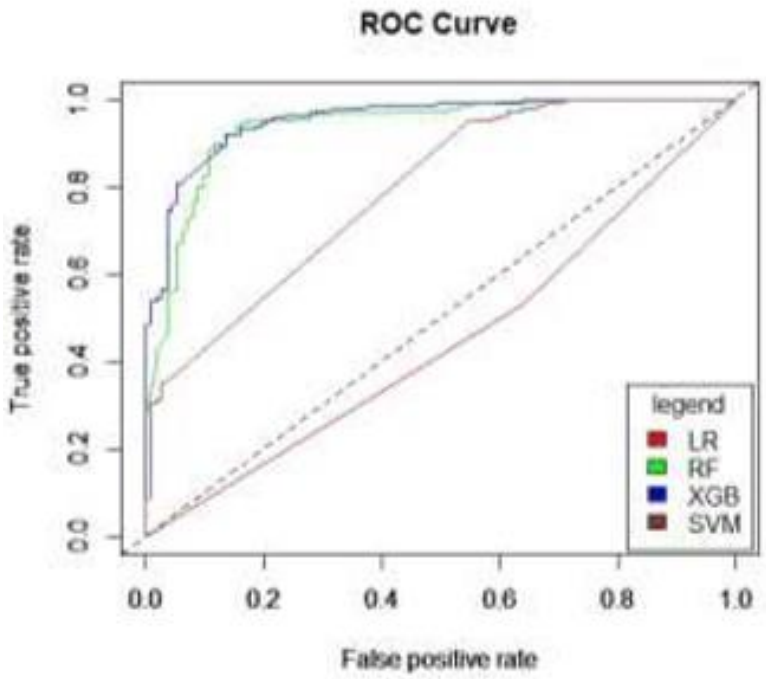

Fig 2(b): TF with 99.9% sparse terms removed

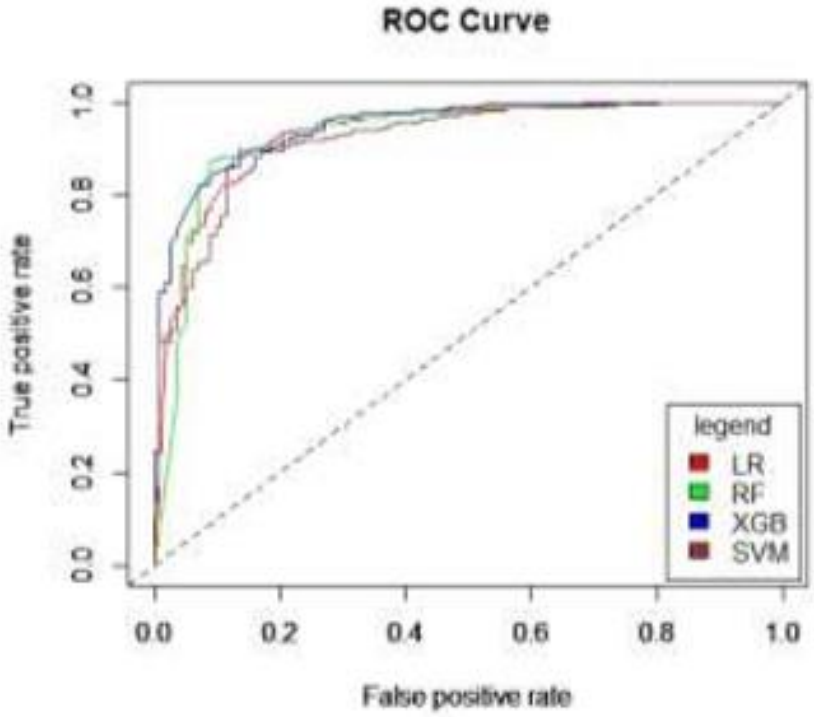

Fig 2(c): TF-IDF with 99% sparse terms removed

V. Conclusion

The ensemble model RF and XGB are the best performing robust algorithms as seen from their consistently high ROC values (>0.87) across different conditions, whereas single standalone models like LR and SVM perform worse and are less robust models, notably with higher dimensional datasets. Although still in its infancy in text analytics, the use of LIME as a model agnostic explanation for any black-box models in the field of healthcare shows promise in fine tuning the decision-making process. In the context of this study, this would mean identifying certain key words or variables which are deciding factors for a clinician's decision towards sending a patient for an MRI scan.

VI. Future Scope

Given the positive results arising from this study, expanding the scope for other types of textual MRI orders from other parts of the body such as spine or body MRI would seem viable. Using the LIME framework, the use of multimodality data becomes possible. In order to derive a practical decision support system to improve the workflows of radiologist for determining the appropriateness of MRI for individual patients, a multimodal data analysis using different types of data captured in the Electronic Health Records (EHR) as text or in the MRI images themselves, can be a topic for future research.

REFERENCES

[1] L. Rokach, R. Romano, and O. Maimon, "Negation recognition in medical narrative reports," Inf. Retr. Boston., vol. 11, no. 6, pp. 499–538, 2008. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[2] L. T. E. Cheng, J. Zheng, G. K. Savova, and B. J. Erickson, "Discerning tumor status from unstructured MRI reports—completeness of information in existing reports and utility of automated natural language processing," J. Digit. Imaging, vol. 23, no. 2, pp. 119–132, 2010.

[3] S. Wang and R. M. Summers, "Machine learning and radiology," Med. Image Anal., vol. 16, no. 5, pp. 933–951, 2012.

[4] A.C. of Radiology, "ACR-ASNR-SPR practice parameterfor the performance and interpretation of magnetic resonance imaging of the brain." 2015.

[5] E. Pons, L. M. M. Braun, M. G. M. Hunink, and J. A. Kors, "Natural language processing in radiology: a systematic review," Radiology, vol. 279, no. 2, pp. 329–343, 2016.

[6] A.D. Brown and T. R. Marotta, "A natural language processingbased model to automate MRI brain protocol selection and prioritization," Acad. Radiol., vol. 24, no. 2, pp. 160–166, 2017.

[7] S. Hassanpour, C. P. Langlotz, T. J. Amrhein, N. T. Befera, and M. P. Lungren, "Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield," Am. J. Roentgenol., vol. 208, no. 4, pp. 750–753, 2017.

[8] J.L.L. Lim, McAdory LE, Tang PH, "Investigating The Appropriateness Of MRI Brain Orders and The Relationship Between Appropriateness of Orders and Imaging Findings in A Tertiary Hospital. A Retrospective Study of MRI Brain Imaging Data.," in 73rd Korean Congress of Radiology, 2017.

[9] 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. Imaging, vol. 31, no. 2, pp. 245–251, 2018.

[10] H.T.Huhdanpaa et al., "Using Natural Language Processing of Free-Text Radiology Reports to Identify Type 1 Modic Endplate Changes," J. Digit. Imaging, vol. 31, no. 1, pp. 84–90, 2018.

[11] W.K.Tan et al., "Comparison of Natural Language Processing Rules-based and Machine-learning Systems to Identify Lumbar Spine Imaging Findings Related to Low Back Pain," Acad. Radiol., 2018.