Role of Hybrid of Rule Based and Machine Learning Natural Language Processing in classifying Free Text Radiology Reports, with special emphasis on identifying Pleural Effusion

Ashutosh Tripathi

Institution: DPS International Pushp Vihar, New Delhi, India, 110017

E-Mail Address: tripathiashutosh14@gmail.com

Role of Hybrid of Rule-Based and Machine Learning Natural Language Processing in classifying Free Text Radiology Reports, with special emphasis on identifying Pleural Effusion

Ashutosh Tripathi*

* High School Student in DPS International Pushp Vihar, New Delhi, India

Abstract- — Radiological reports, particularly the free-text, are a good source of clinical data which can be used to assist with surveillance of disease. Pleural Effusion and other radiological findings on chest X-ray or chest computed tomography (CT) scans are one type of relevant result to, both, health services and the medical community at large. In this study, we examined the ability of a Hybrid system to identify Pleural Effusion from free-text radiological reports. We used a hybrid of a machine learning and rule-based NLP system. The system encoded the reports, and then a defined set of rules were created aimed at the identification of the pleural effusion. The rules were executed against the encodings of the radiological reports, followed by further classification. Four different methods for classification were used to compare and conclude the best approach. The accuracy of the reports was compared with a Clinician review of the Radiological Reports. We find that NLP based computable rules are accurate enough for the automated bio-surveillance of Pleural Effusion from radiological reports. However, this requires further validation with multiple large databases and more diverse database

Index Terms- Classify, Free-Text, Machine Learning, Natural Language Processing, Radiology, Pleural Effusion, Text Mining

I. INTRODUCTION

This project seeks to determine the accuracy of a Natural Language Processing (NLP) based system for the identification of patients with Pleural Effusion from a corpus of radiological reports.

A pleural effusion is an excessive accumulation of fluid in the pleural space. It indicates an imbalance between pleural fluid formation and its removal. Pleural effusions accompany a wide variety of disorders of the lung, pleura, and systemic disorders. Therefore, a patient with pleural effusion may present not only to a pulmonologist but to a general internist, rheumatologist, gastroenterologist, nephrologist, or surgeon. Due to the wide variety of fields where this disease could affect the patients, it is rather fitting to study identification and classification of pleural effusion to determine the accuracy of an NLP and Rule-Based system in identification of patients with Pleural Effusion.

The problem this study tries to address is the applicability of a hybrid classification system in identification of pleural effusion in free-text radiology reports. This problem is further described in the introduction section below.

The NLP Technique used is one based on a hybrid of pre-defined rules and machine learning. There have been multiple studies to test the application of different types of NLP systems in the past. A hybrid NLP system was selected due to several reasons. Firstly, identification of Pleural Effusion, alone, requires a very vast clinical knowledge of specific terms used in the radiology reports. In a rule-based NLP system, Clinical knowledge can be manually incorporated. For instance, if we were to expand this algorithm's use, we could use the Unified Medical Language System (UMLS).

On the other hand, a Machine Learning Algorithm alone would require annotation of these terms. Since there can be many such terms, annotation not only becomes tedious but error-prone as well. A combination of both reduces the work and dramatically increases the efficiency, in theory.

In a rule-based NLP system, rules can be readily added and modified to accommodate a new target. For example, if the goal was shifted from diagnosing Pleural Effusion to one that was diagnosing Ascites, for instance.

Furthermore, multiple previous radiology report parsing studies[30][31] done have indicated that Machine learning-based NLP systems are inferior to one that is hybrid of rule-based and machine learning.

Rule-based NLP also foregoes of the unnecessary hassle in machine learning approach, because unstructured text cannot be directly interpreted by a machine, due to text's ambiguity and subtlety of natural language. These problems, combined with variations among different radiologist and healthcare organizations, leads to an inevitable bottleneck is not only a machine learning-based algorithm but this as well.

Although in recent years, there has been a slight shift in trends in radiological reports, with a more standardized and structured reporting method being utilized, the majority of the stories remain unstructured and in free form language. This particular study, therefore, focuses on one specific healthcare organizations. This leads to a degree of uniformity in the structure and vocabulary used in free-text radiological reports.

The classifier developed in this case, furthermore, focuses solely on two instances - a negative for Pleural Effusion, or a positive (or suggestive positive) for the same. Although an over-simplification of the process of interpretation of the data, this allows for the portrayal of the fact that NLP systems can be used efficiently for interpreting radiology reports, and determine, with a certain degree of confidence, presence of Pleural Effusion.

II. METHOD

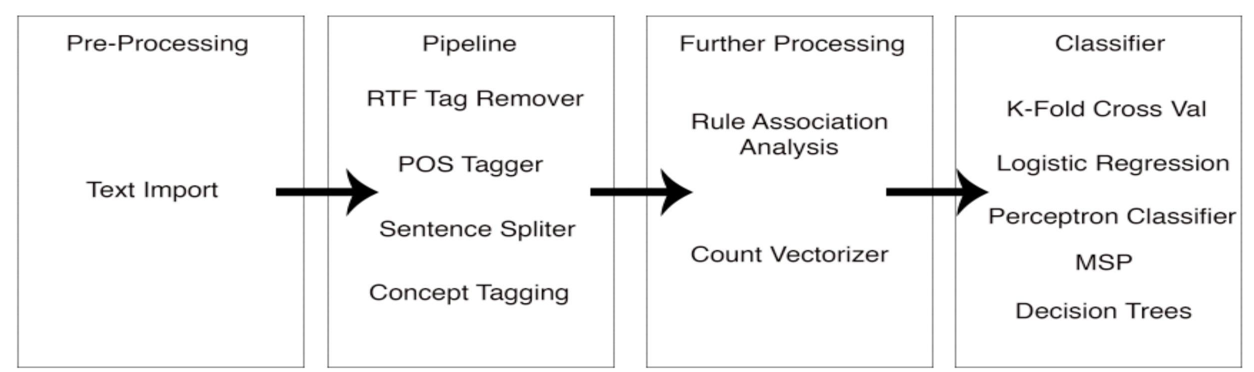

Fig. 1: Brief Overview of the Method used

Briefly, the method used by our system is a straightforward one. As described in figure 1 (Fig 1), the method can be divided into four main stages – **Pre-Processing, Pipelining, Further Processing and Classification.**

Pre-processing is the stage during which the text is imported and pre-processed. Pre-processing involves stages from the Pipelining stage, and includes processes such as removing RTF tags (as was common in the radiology reports studied at the hospital in question). Alongside this, POS taggers, sentence splitters and concept tagging is used to make the process of processing the data easier.

Afterwards, the data goes through a count-vectorizer(initialized/based on the initial dataset). The count-vectorizer rule-association analysis was based on the words found to be relevant to identification of pleural effusion symptoms during the study, as has been described below.

Lastly, the data is classified. As described later in this study, we used four different methods of classification–**Logistic Regression, MSP, Decision Trees , Perceptron Classifier.** The reason for the same has been described later in this document.

A. Dataset

The dataset used contains ~2300 Patient Records, due to which, the data is not being made public. These were a combination of X-Rays, Abdomen USG, CT Scans, MRI Scans and Ultrasound scans. Some of these scans were present in the corpus even though they had no relation to chest or even abdomen. These were there to ‘throw off’ the machine–basically to make the entire system more robust by providing it training for false reports from the start.

Although efforts were taken to make the training set less skewed, there was clearly more data not relevant to pleural effusion than there were those relevant to pleural effusion–the distribution was still, however, a respectable 3:1 split. Furthermore, there were way more reports that had pleural effusion present compared to those that had no pleural effusion a 2:3 split in this case.

However, the algorithm can be used, in theory, with any radiology report corpus. Along with patient records, it also contained radiology free-text reports, and the test type. For study purposes, the test-type field was removed from the final-data that was used. The corpora contained different types of radiology reports of patients, stored as rtf files. Furthermore, out of these data, ~2000 records were initially labelled by a Clinician for the train-test splits. Each label indicated either presence or absence of Pleural Effusion. Labels used were as shown in Table 1.

TABLE 1

Labels	**Description**
0	No significant evidence of presence of Pleural Effusion
1	There is significant evidence of presence of Pleural Effusion

B. Pre-Processing

For Firstly, we made an RTF Tag Remover. Since each of these documents were of the Rich Text Format, it was quintessential to remove these RTF specific tags to get some sort of meaningful result.

This was followed by normal pre-processing steps such as changing all the text to lower case. After this, report-specific list of ‘stop words’-words that really did not contribute much to the meaning of the sentence-were made. These were the same as those normally used in libraries, but certain words such as ‘no’ and ‘not’ were important, as defined later, and thus excluded from this list.

Using a partially Rule-Based Approach meant we needed to make a negation detector from the very scratch. Although the method used was grossly inadequate, we proceeded with it (the possible suggestions and continuations are mentioned later in this report) Using a Rule-Based approach furthermore also meant that we had to define a dictionary of words that were commonly occurring in reports and that were related to the objective – to detect pleural effusion.

Thus, after manually analyzing the pre-annotated reports, we decided to proceed by making a dictionary of the most frequently occurring unigrams, bigrams, trigrams, and quadrigrams from the list of words that were important features of Pleural Effusion- such as “left upper lobe”. We found occurrence of such words and concatenated them into one. So, ‘left upper lobe’ became ‘left_upper_lobe’.

We had a similar approach to negation detection. For instance, if there was a phrase ‘no evidence’, it indicated a negation, and thus we concatenated the words into one (‘no_evidence') and proceeded as detailed below. This particular method was opted for due to the fact that the structure of the sentences wasn’t very complex.

Evidently, there were certain words that we must have missed, or those that were redundant and didn’t contribute much to our study. To handle this, we used Rule-Association Mining.

We used Python’s Apyori Library’s Apiori’s algorithm to conduct Rule-Association mining with the pre-defined features to get a sense of the applicability of the features we were using. Rule-Association mining led to a better determination of the dictionary to finally use for the study. Every single time, the dictionary was composed of words that were related to pleural effusion and their symptoms. These were made alongside the clinicians to get a better sense of symptoms commonly used in radiology reports.

This alongside a manual analysis of the database led to a more robust dictionary at the end.

C. Mapping and Tagging of Words

After the pre-processing stage, all the important n-grams were now one single words, thus we could proceed with the mapping and tagging phase.

We used the manually analyzed list of important features (such as 'left_pleural_effusion') and indicated them as 'FTR' (meaning feature). Similarly, for the negations indicating no evidence, we now replaced them with the words 'SAFE'. This we did for words that indicated 'RISK' and adjectives as well (which were annotated as 'ADJ').

This process of mapping and tagging of words was essential. We had now defined our rules – the existence of these defined dictionary of words- and had implemented it, essentially, in our pre-processing step to make the learning easier.

D. Classification Algorithm

We tried multiple different approaches before sticking with 4 different algorithms -logistic regression, perceptron learning algorithm for binary classifier, multi-layer perceptron learning algorithm, and decision tree learning.

But before going directly to logistic regression, we needed to vectorize our corpus. We use a Count Vectorizer to vectorize our edited corpus of data into a count-vectorizer. Then, we fed this into our above stated algorithms and trained it using this Count Vectorizer.

For initial Testing, we did an 80-20 train-test split. We used the vectorizers created from the training data and did a vector transform on the testing data. We then trained each of our algorithms using a train data, and then initially tested it using the test data.

We further tested each algorithm on unseen (by the machine) dataset of radiology reports and compared it to manual annotations of these reports by multiple clinicians.

III. EVALUATION

For a fairer (and less optimistic) judgement of our algorithms, we used k-cross validation, with k=15. We then tested our algorithms further to get a better view of each algorithm.
Each algorithm was tested in 6 ways – we saw their performance on training data, the testing data, and the unseen data, followed by performance on the same datasets, with k-cross validation.

A. Performance Measures

We use the standard performance measures (as used in both medical and computing literature) to assess the performance of classification tasks. The formulations for our measures are below, stated in terms of true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

- Sensitivity (or Recall): TP/(TP+FN)
- Specificity: TN/(TN+FP)
- Precision: TP/(TP+FP)
- Accuracy: TP+TN/(TP+FP+TN+FN)
- F1 score: (2*TP)/(2*TP+FP+FN)
- Area Under the Receiver Operating Characteristic Curve
- Mean Squared Errors (MSE)

The mean squared error (MSE) of an estimator measures the average squared difference between the estimated values and what is estimated. MSE is a function, corresponding to the expected value of the squared error loss. MSE is used to make better conclusions on whether a model is an underfit, an overfit, or a decent/good fit.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

In the clinical setting, the main reason to use an automated classification technique is to reduce the amount of data that clinicians must review to make decisions. Thus, recall is critical for reducing liability and precision is critical for minimizing time needed for secondary review.

B. Logistic Regression

Logistic regression was used due to the fact that it was probably the best regression model for this classification task for this type of data, where the annotations were binary in nature, or, in technical terms, the dependent variable was dichotomous.

a). Training Results

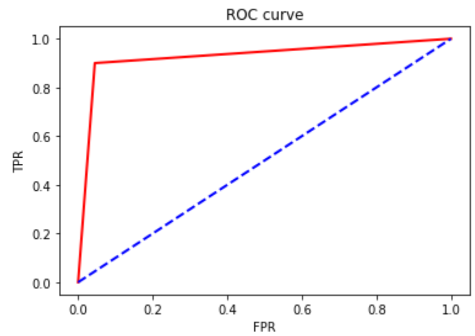

Fig. 2: ROC Curve for Logistic Regression on Training Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 2

	Regular
Sensitivity	0.9000
Specificity	0.9552
Precision	0.80769
Accuracy	0.945679
F1 score	0.85135
AUROC	0.92761
MSE	0.09972

b). Test-Data Results

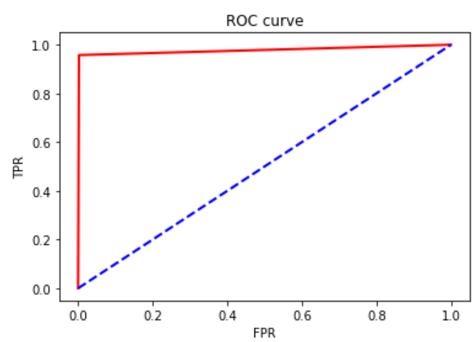

Fig. 3: ROC Curve for Logistic Regression on Testing Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 3

	Regular
Sensitivity	0.957516
Specificity	0.99771
Precision	0.9898
Accuracy	0.99009
F1 score	0.973421
AUROC	0.97761
MSE	0.08407

c). Results on Unseen Data

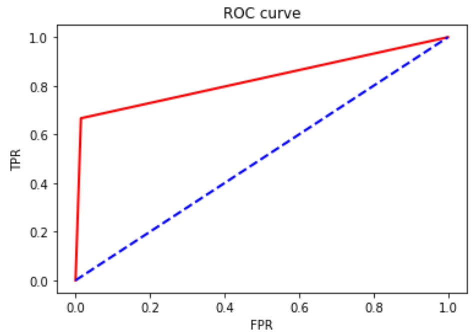

Fig. 4: ROC Curve for Logistic Regression on Unseen Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 4

	Regular
Sensitivity	0.66666
Specificity	0.98551
Precision	0.952380
Accuracy	0.8889
F1 score	0.78431
AUROC	0.82609
MSE	0.15397

d). Training Results (K-Cross Validation)

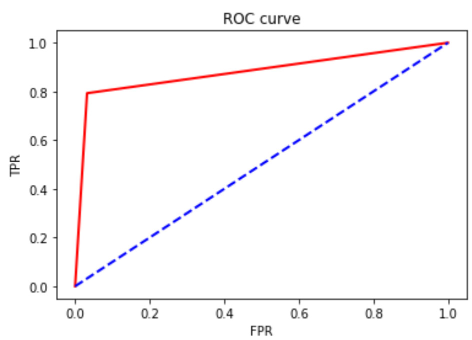

Fig. 5: ROC Curve for Logistic Regression on Training Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 5

	Regular
Sensitivity	0.79276
Specificity	0.9679878
Precision	0.85159
Accuracy	0.93502
F1 score	0.82112
AUROC	0.880375

e). Test-Data Results (K-Cross Validation)

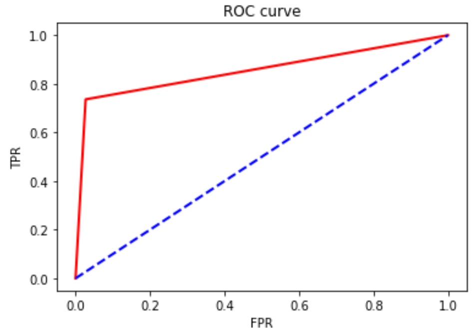

Fig. 6: ROC Curve for Logistic Regression on Test Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 6

	Regular
Sensitivity	0.7361
Specificity	0.97297
Precision	0.8548
Accuracy	0.93086
F1 score	0.7910
AUROC	0.8545

f). Results on Unseen Data (K-Cross Validation)

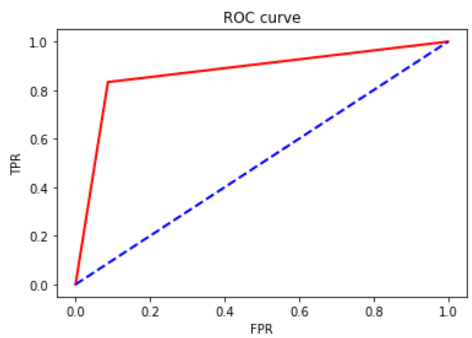

Fig. 7: ROC Curve for Logistic Regression on Unseen Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 7

	Regular
Sensitivity	0.8333
Specificity	0.9130
Precision	0.80645
Accuracy	0.88889
F1 score	0.81967
AUROC	0.873188

C. Perceptron Classifier

A supervised learning algorithm for binary classification, Perceptron is a single layer neural network. Since this is a single layer neural network approach, it was imperative to see if it was useful in classification of the radiology reports.

a). Training Results

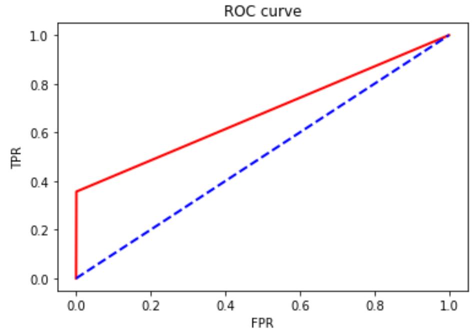

Fig. 8: ROC Curve for Perceptron Classifier on Training Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 8

	Regular
Sensitivity	0.3566
Specificity	0.99924
Precision	0.9907
Accuracy	0.87995
F1 score	0.5245
AUROC	0.6779
MSE	0.091079

b). Test-Data Results

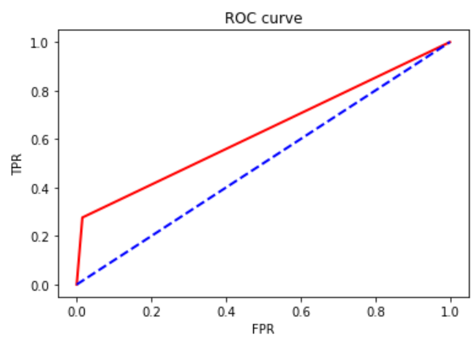

Fig. 9: ROC Curve for Perceptron Classifier on Testing Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 9

	Regular
Sensitivity	0.2763158
Specificity	0.9848
Precision	0.80769
Accuracy	0.85185
F1 score	0.41176
AUROC	0.63055
MSE	0.101594

c). Results on Unseen Data

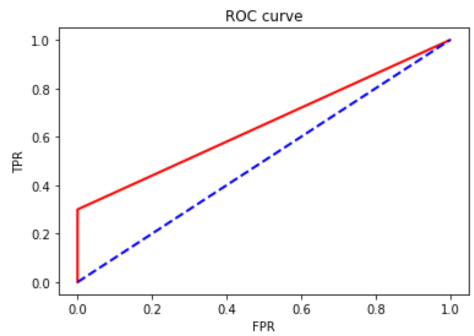

Fig. 10: ROC Curve for Perceptron Classifier on Unseen Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 10

	Regular
Sensitivity	0.3
Specificity	1.0
Precision	1.0
Accuracy	0.78788
F1 score	0.4615
AUROC	0.65
MSE	0.153968

d). Training Results (K-Cross Validation)

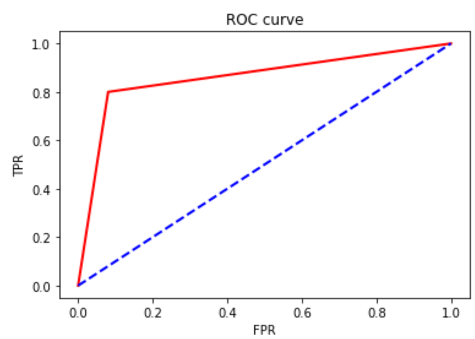

Fig. 11 : ROC Curve for Perceptron Classifier on Training Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 11

	Regular
Sensitivity	0.8
Specificity	0.91945
Precision	0.69364
Accuracy	0.897277
F1 score	0.7430
AUROC	0.85972

e). Test-Data Results (K-Cross Validation)

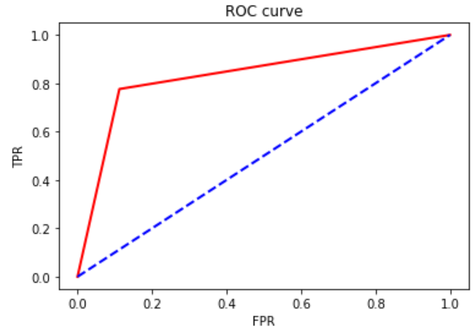

Fig. 12: ROC Curve for Perceptron Classifier on Test Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 12

	Regular
Sensitivity	0.77632
Specificity	0.88754
Precision	0.61458
Accuracy	0.86667
F1 score	0.68605
AUROC	0.83193

f). Results on Unseen Data (K-Cross Validation)

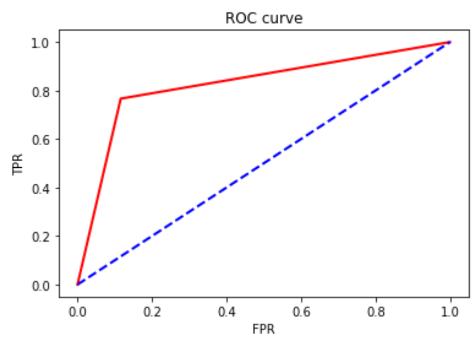

Fig. 13: ROC Curve for Perceptron Classifier on Unseen Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 13

	Regular
Sensitivity	0.766667
Specificity	0.88406
Precision	0.7419
Accuracy	0.84848
F1 score	0.754098
AUROC	0.82536

D. Multi-Layer Perceptron Classifier (MLP)

A Multi-Layer Perceptron Classifier is more commonly known as a Neural Network. To be able to solve nonlinearly separable problems, MLP is composed of more than one perceptron. They are composed of an input layer to receive the signal, an output layer that makes a decision or prediction about the input, and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP. MLPs with one hidden layer are capable of approximating any continuous function.

There are pretty high chances that the classifier needed for this project might need to be non-linear in nature, and thus we chose MLP

a). Training Results

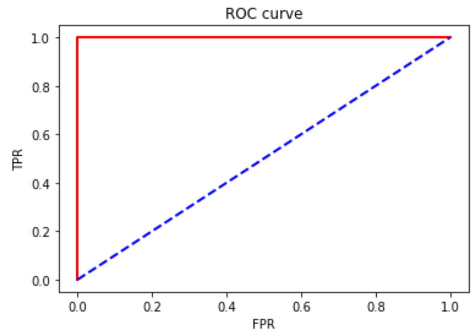

Fig. 14: ROC Curve for MLP on Training Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 14

	Regular
Sensitivity	1.0
Specificity	1.0
Precision	1.0
Accuracy	1.0
F1 score	1.0
AUROC	1.0
MSE	0.07058

b). Test-Data Results

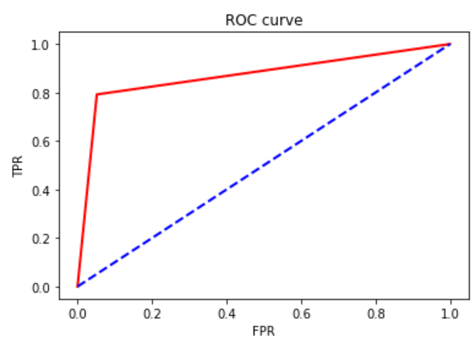

Fig. 15: ROC Curve for MLP on Testing Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 15

	Regular
Sensitivity	0.7922
Specificity	0.94817
Precision	0.78205
Accuracy	0.9185
F1 score	0.7871
AUROC	0.8702
MSE	0.1063

c). Results on Unseen Data

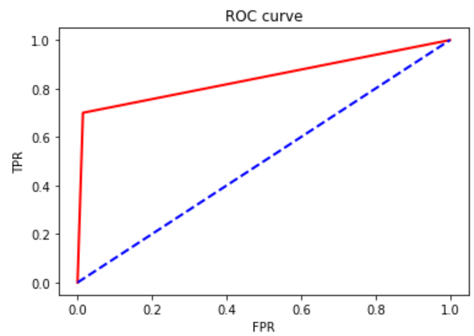

Fig. 16: ROC Curve for MLP on Unseen Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 16

	Regular
Sensitivity	0.7
Specificity	0.9855
Precision	0.9545
Accuracy	0.89899
F1 score	0.80769
AUROC	0.84275
MSE	0.1222

d). Training Results (K-Cross Validation)

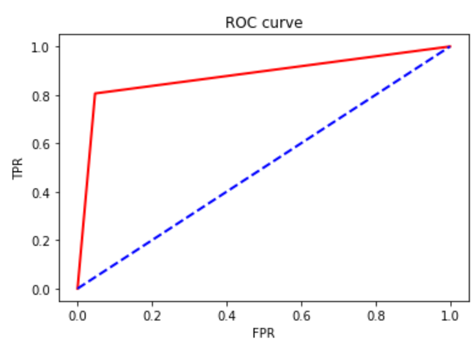

Fig. 17: ROC Curve for MLP on Training Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 17

	Regular
Sensitivity	0.8060
Specificity	0.9529
Precision	0.7954
Accuracy	0.92574
F1 score	0.80066
AUROC	0.87947

e). Test-Data Results (K-Cross Validation)

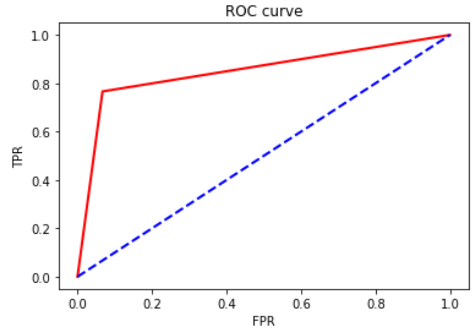

Fig. 18: ROC Curve for MLP on Test Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 18

	Regular
Sensitivity	0.76623
Specificity	0.9329
Precision	0.7284
Accuracy	0.9012
F1 score	0.7468
AUROC	0.84958

f). Results on Unseen Data (K-Cross Validation)

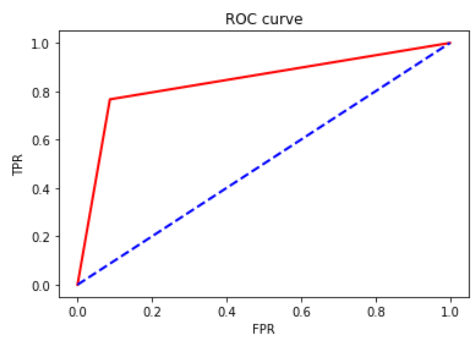

Fig. 19: ROC Curve for MLP on Unseen Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 19

	Regular
Sensitivity	0.7667
Specificity	0.9130
Precision	0.79310
Accuracy	0.86869
F1 score	0.77966
AUROC	0.83986

E. Decision Trees

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question, edges represent the answers the to these questions, and the leaves represent the actual output. Decision Trees are a supervised learning method used for non-linear classification and regression tasks. Since this classification task is most-probably non-linear in nature, we use Decision Trees as a classifier algorithm. We use ID3 algorithm for Decision Tree learning.
However, using ID3 increases probability of overfit models.

a). Training Results

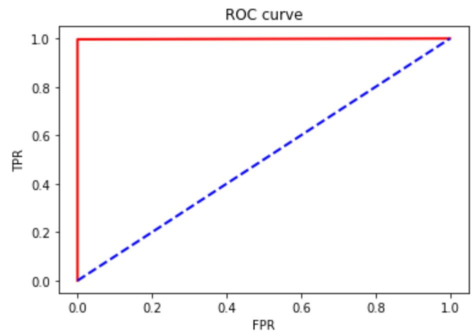

Fig. 20: ROC Curve for Decision Trees on Training Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 20

	Regular
Sensitivity	0.996587
Specificity	1.0
Precision	1.0
Accuracy	0.9994
F1 score	0.9983
AUROC	0.99829
MSE	0.08109

b). Test-Data Results

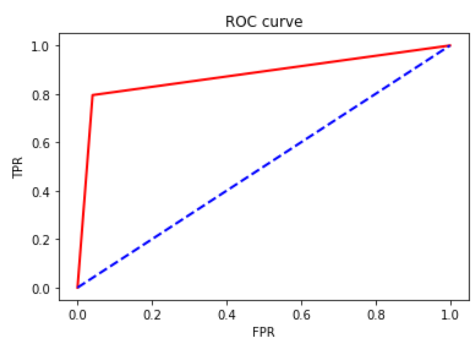

Fig. 21: ROC Curve for Decision Trees on Testing Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 21

	Regular
Sensitivity	0.79518
Specificity	0.959627
Precision	0.8354
Accuracy	0.9259
F1 score	0.8148
AUROC	0.8774
MSE	0.1193

c). Results on Unseen Data

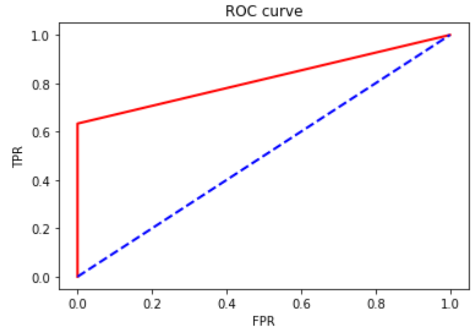

Fig. 22: ROC Curve for Decision Trees on Unseen Data (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 22

	Regular
Sensitivity	0.63333
Specificity	1.0
Precision	1.0
Accuracy	0.8889
F1 score	0.7755
AUROC	0.81667
MSE	0.179365

d). Training Results (K-Cross Validation)

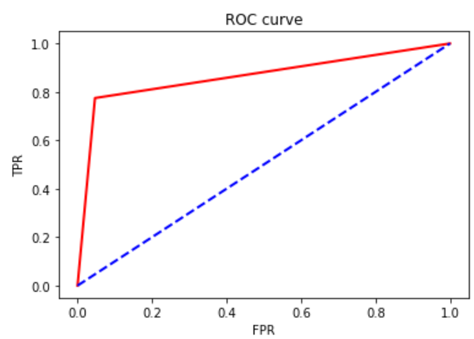

Fig. 23: ROC Curve for Decision Trees on Training Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 23

	Regular
Sensitivity	0.77474
Specificity	0.95314
Precision	0.78547
Accuracy	0.92079
F1 score	0.7800687
AUROC	0.86394

e). Test-Data Results (K-Cross Validation)

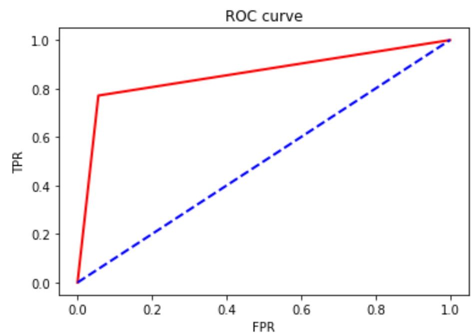

Fig. 24: ROC Curve for Decision Trees on Test Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 24

	Regular
Sensitivity	0.7711
Specificity	0.9441
Precision	0.7805
Accuracy	0.9086
F1 score	0.77576
AUROC	0.85759

f). Results on Unseen Data (K-Cross Validation)

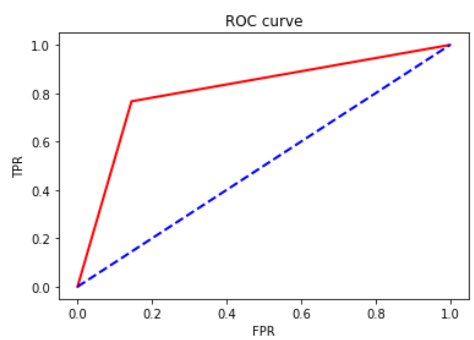

Fig. 25: ROC Curve for Decision Trees on Unseen Data with k-cross validation (in red) and ROC Curve for a perfectly random classifier (in dotted blue)

TABLE 25

	Regular
Sensitivity	0.76667
Specificity	0.85507
Precision	0.69697
Accuracy	0.82828
F1 score	0.7301587
AUROC	0.81087

IV. RESULTS AND DISCUSSION

Our Rule-Based Machine Learning Hybrid NLP approach is certainly non-linear, as established from the poor performance of the single perceptron classifier (without k-cross validation).

What we can further conclude that Decision Trees isn't the best classification algorithm to use, due to the AUROC for training data being ~18% greater than that for the testing data (without k-cross validation).

From the above considered algorithms, the best performers were the Logistic Regression, followed by MSPs, then by Decision Trees, with the worst performing classifier being Single Perceptron Classifier.

(*Note from the author: The algorithm for the study can be found at https://github.com/Ashutoshtripathi14/rdiotxtpe.*)

The results obtained are encouraging, for we have shown that this approach can be used to classify Radiology Reports based on the presence of Pleural Effusion to an acceptable degree of efficiency.

V. CONCLUSION

Since the current algorithm is based on a dictionary, a better approach would be to make a parser to parse the sentence, divide it into categories, and then accordingly annotate, instead of annotating just few occurrences of a particular word. This will, further, deal with the problem of negations and context. Not to mention, a better dictionary through the use of possible market-basket analysis is also possible.

Although a larger training data might be useful, it isn't indicative of real-world industry projects, where this is the most probable data size available.

The applicability and performance of this method is now solely dependent on one factor and one factor alone – the robustness of the dictionary we used. Since we were only analyzing presence of Pleural Effusion meant that we could do manual checking and make a dictionary manually. However, if the case was that we were making it a more general program, another approach or a pre-defined database of medical words would have become necessary.

Our approach to detection of findings of Pleural effusion was done through using a limited set of rules based on knowledge and using patterns from multiple sources of data. We have shown that we are able to extract useful information for set of rules and combine it with machine learning to successfully identify items or traces of items – in this case Pleural Effusion - to be found in an unstructured document – in this case the free-text radiology report.

REFERENCES

[1] W. Christopher Baughman, Eamon Johnson and Gultekin Ozsoyoglu, *Mixing Domain Rules with Machine Learning for Radiology Text Classification,* 2014.

[2] C. A. Bejan, F. Xia, L. Vanderwende, M. M. Wurfel, and M. Yetisgen-Yildiz. *Pneumonia identification using statistical feature selection*. Journal of the American Medical Informatics Association: JAMIA, 19(5):817–23, 2011.

[3] B. Chapman and S. Lee. *Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm*. Journal of biomedical . . ., 44(5):728–737, 2011.

[4] P. a. Dang, M. K. Kalra, M. a. Blake, T. J. Schultz, E. F. Halpern, and K. J. Dreyer. *Extraction of recommendation features in radiology with natural language processing: exploratory study.* AJR. American journal of roentgenology, 191(2):313–20, Aug. 2008

[5] N. Haven. V. Garla, V. L. Re, A. Justice and C. Brandt, *Automating the classification of radiology reports indicating hepatic decompensation.* page 2010, 2008.

[6] V. Liu, M. P. Clark, M. Mendoza, R. Saket, M. N. Gardner, B. J. Turk, and G. J. Escobar. *Automated identification of pneumonia in chest radiograph reports in critically ill patients.* BMC medical informatics and decision making, 13(1):90, Jan. 2013.

[7] S. Dublin, E. Baldwin, R. L. Walker, L. M. Christensen, P. J. Haug, M. L. Jackson, J. C. Nelson, J. Ferraro, D. Carrell, and W. W. Chapman. *Natural Language Processing to identify pneumonia from radiology reports*. Pharmacoepidemiology and drug safety, 22:834–41, 2013.

[8] A, Maghoosdi, M. Sevenster, J. Scholtes, and G. Nalbantov. *Sentence-based classification of free-text breast cancer radiology reports.* 2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS), pages 1-4, June 2012

[9] H.H. Abujudeh, R. Kaewlai, K. Farsad, E. Orr, M. Gilman, J.A. Shepard *Computed tomography pulmonary angiography: an assessment of the radiology report* Acad Radiol, 16 (2009), pp. 1309-1315

[10] Aronis JM, Cooper GF, Kayaalp M, Buchanan BG. *Identifying patient subgroups with simple Bayes'. Proc* AMIA Symp; 1999: 658–62.

[11] Aronson AR. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*. Proc AMIA Symp 2001:17–21.

[12] L. Berlin *Pitfalls of the vague radiology report* AJR Am J Roentgenol, 174 (2000), pp. 1511-1518

[13] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan *A simple algorithm for identifying negated findings and diseases in discharge summaries* J Biomed Inform, 34 (2001), pp. 301-310

[14] W.W. Chapman, L.M. Christensen, M.M. Wagner, P.J. Haug, O. Ivanov, J.N. Dowling, *et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing* Artif Intell Med, 33 (2005), pp. 31-40

[15] W.W. Chapman, G.F. Cooper, P. Hanbury, B.E. Chapman, L.H. Harrison, M.M. Wagner *Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders* J Am Med Inform Assoc, 10 (2003), pp. 494-503

[16] W.W. Chapman, M. Fiszman, P.R. Frederick, B.E. Chapman, P.J. Haug *Quantifying the characteristics of unambiguous chest radiography reports in the context of pneumonia* Acad Radiol, 8 (2001), pp. 57-66

[17] Christensen LM, Harkema H, Haug PJ, Irwin JY, Chapman WW. *Onyx: a system for the semantic analysis of clinical text.* In: BioNLP '09: proceedings of the workshop on BioNLP. Association for Computational Linguistics, Morristown, NJ, USA; 2009. p. 19–27.

[18] P.L. Elkin, S.H. Brown, B.A. Bauer, C.S. Husser, W. Carruth, L.R. Bergstrom, *et al. A controlled trial of automated classification of negation from clinical notes* BMC Med Inform Decis Mak, 5 (2005), p. 13

[19] Friedman C. *A broad-coverage natural language processing system*. Proc AMIA Symp 2000:270–4.

[20] M. Fiszman, W.W. Chapman, D. Aronsky, R.S. Evans, P.J. Haug *Automatic detection of acute bacterial pneumonia from chest X-ray reports* J Am Med Inform Assoc, 7 (2000), pp. 593-604

[21] Huang Y, Lowe H. *A grammar-based classification of negations in clinical radiology reports.* Proc AMIA Annu Fall Symp 2005:988.

[22] K.S. Jones *A statistical interpretation of term specificity and its application in retrieval* J Doc, 28 (1972), pp. 11-21

[23] Mowery DL, Harkema H, Dowling J, Lustgarten J, Chapman WW. *Distinguishing historical from current problems in clinical reports – which textual features help?* In: BioNLP workshop of the 47th annual meeting of the Association of Computational Linguistics, Boulder, CO; 2009.

[24] P.G. Mutalik, A. Deshpande, P.M. Nadkarni *Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS* J Am Med Inform Assoc, 8 (2001), pp. 598-609

[25] S. Pakhomov, S. Bjornsen, P. Hanson, S. Smith *Quality performance measurement using the text of electronic medical records* Med Decis Making, 28 (2008), pp. 462-470

[26] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, *et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications* J Am Med Inform Assoc, 17 (5) (2010), pp. 507-513

[27] Ö. Uzuner, X. Zhang, T. Sibanda *Machine learning and rule-based approaches to assertion classification* J Am Med Inform Assoc, 16 (1) (2009), pp. 109-115

[28] B.R. South, S. Phansalkar, A.D. Swaminathan, J. Anthony, S. Delisle, T. Perl, *et al. Text-processing of VA clinical notes to improve case detection models for influenza-like illness* Adv Dis Surveill, 2 (2007), p. 28

[29] Wilson RA. *Automated ancillary cancer history classification for mesothelioma patients from free-text clinical reports.* Master's thesis, University of Pittsburgh; 2010.

[30] Bethany Percha, Houssam Nassif, Jafi Lipson, Elizabeth Burnside & Daniel Rubin (2012). *Automatic classification of mammography reports by BI-RADS breast tissue composition class*

[31] John Zech, Margaret Pain, Joseph Titano, Marcus Badgeley, Javin Schefflein, Andres Su, Anthony Costa, Joshua Bederson, Joseph Lehar & Eric Karl Oermann (2018*). Natural Language-based Machine Learning Models for the Annotation of Clinical Radiology Reports*