Computer Methods and Programs in Biomedicine Update

Interpretable Machine Learning Text Classification for Clinical Computed Tomography reports – A Case Study of Temporal Bone Fracture

--Manuscript Draft--

Manuscript Number:	CMPBUP-D-22-00053R1
Full Title:	Interpretable Machine Learning Text Classification for Clinical Computed Tomography reports – A Case Study of Temporal Bone Fracture
Article Type:	Original Research
Section/Category:	Health Informatics
Keywords:	Machine Learning; Computed Tomography; Text Classification
Corresponding Author:	Jake Luo University of Wisconsin-Milwaukee UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Wisconsin-Milwaukee
Corresponding Author's Secondary Institution:	
First Author:	Ling Tong, MS
First Author Secondary Information:	
Order of Authors:	Ling Tong, MS
	Jake Luo, PhD
	Jazzmyne Adams, MPH
	Kristen Osinski, MS
	Xiaoyu Liu, MBA
	David Friedland, MD, PhD
Order of Authors Secondary Information:	
Abstract:	Background Machine learning (ML) has demonstrated success in classifying patients' diagnostic outcomes in free-text clinical notes. However, due to the machine learning model's complexity, interpreting mechanisms of classification results remains difficult. Methods We investigated interpretable representations of machine learning classification models. We created machine learning models to classify temporal bone fractures based on 164 temporal bone Computed Tomography (CT) text reports. We adopted the XGBoost, Support Vector Machine, Logistic Regression, and Random Forest algorithms. To interpret models, we used two major methodologies: (1) We calculated the average word frequency score (WFS) for keywords. The word frequency score shows the frequency gap between positive and negative classified cases. (2) We used Local Interpretable Model-Agnostic Explanations (LIME) to show the word-level contribution to bone fracture classification. Results In temporal bone fracture classification, the random forest model achieved an average F1-score of 0.93. WFS reveals a difference in keyword usage between fracture and non-fracture cases. Additionally, LIME visualized the keywords' contributions to classification results. The evaluation of LIME-based interpretation achieved the highest interpreting accuracy of 0.97. Conclusion The interpretable text explainer can improve physicians' understanding of machine learning predictions. By providing simple visualization, our model can increase the trust

	of computerized models. Our model supports more transparent computerized decision-making in clinical settings.
Suggested Reviewers:	lingyun luo, PhD Associate Professor, University of South China luoly@usc.edu.cn
	Guido Zuccon, PhD Associate Professor, The University of Queensland g.zuccon@uq.edu.au
	Vikas Chaurasia Research Scholar, VBS Purvanchal University: Veer Bahadur Singh Purvanchal University chaurasia.vikas@gmail.com
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Free Preprint Service Do you want to share your research early as a preprint? Preprints allow for open access to and citations of your research prior to publication. Computer Methods and Programs in Biomedicine Update offers a free service to post your paper on SSRN, an open access research repository, when your paper enters peer review. Once on SSRN, your paper will benefit from early registration with a DOI and early dissemination that facilitates collaboration and early citations. It will be available free to read regardless of the publication decision made by the journal. This will have no effect on the editorial process or outcome with the journal. Please consult the SSRN Terms of Use and FAQs.	YES, I want to share my research early and openly as a preprint.

Interpretable Machine Learning Text Classification for Clinical Computed Tomography reports – A Case Study of Temporal Bone Fracture

Ling Tong[1], MS, Jake Luo[1], PhD, Jazzmyne Adams[2], MPH, Kristen Osinski[3], MS, Xiaoyu Liu[4], MBA, David Friedland[2], MD, PhD

1. Department of Health Informatics and Administration, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA

2. Department of Otolaryngology and Communication Sciences, Medical College of Wisconsin, Milwaukee, Wisconsin, USA

3. Clinical and Translational Science Institute, Medical College of Wisconsin, Milwaukee, Wisconsin, USA

4. Department of Electrical Engineering and Computer Science, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA

Dear Editor and Reviewers,

We highly appreciate your detailed, instructive, and valuable comments on the manuscript, which was submitted to Computer Methods and Programs in Biomedicine Update. The suggestions are extremely helpful. Your comments allow us to further revise the manuscript.

We revised our manuscript significantly as suggested by the reviewers. We ensure that each of the reviewer's comments has been addressed carefully and that the paper is revised accordingly. Please check our newly added table of contents to read the summary of manuscript revisions. For revision details, please see the "Response to Reviewer's Comments" section for more information. In this section, we first summarize major changes since the last submission. In the major changes, we addressed all the comments and concerns raised by the reviewers. We also made additional extensive changes to improve the manuscript's expression and clarity. Then, we attached a detailed point-by-point response to every comment. After making the suggested edits, we believe our revision improved the manuscript.

We have provided both marked and unmarked manuscript files for your convenience. The track-change indicates where we made changes. The non-track change document makes for an easier reading of the manuscript.

The revision has been developed in consultation with all coauthors. Each author has given approval to the final form of this revision. We hope that the revised manuscript has reached the quality of publication for the journal.

Best Regards,

Jake Luo, PhD

Associate Professor

Department of Health Informatics and Administration

College of Health Sciences

University of Wisconsin, Milwaukee

Email: jakeluo@uwm.edu, zluo@mcw.edu

Interpretable Machine Learning Text Classification for Clinical Computed Tomography reports – A Case Study of Temporal Bone Fracture

Tong, Ling; Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, USA, ltong@uwm.edu

Luo, Jake; Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, USA, jakeluo@uwm.edu

Jazzmyne Adams, Department of Otolaryngology and Communication Sciences, Medical College of Wisconsin, Milwaukee, Wisconsin, USA, jaadams@mcw.edu

Osinski, Kristen; Medical College of Wisconsin, Clinical and Translational Science Institute of Southeastern Wisconsin, kosinski@mcw.edu

Xiaoyu Liu, Department of Electrical Engineering and Computer Science, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA, liu267@uwm.edu

David Friedland, Department of Otolaryngology and Communication Sciences, Medical College of Wisconsin, Milwaukee, Wisconsin, USA, dfriedland@mcw.edu

Acknowledgement:

Conflict of interest: All authors declare that there is no conflict of interest.

Funding:

The project described was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, Award Number UL1TR001436. The content is solely the responsibility of the author(s) and does not necessarily represent the official views of the NIH.

This project is also funded by the Research and Education Program Fund, a component of the Advancing a Healthier Wisconsin Endowment at the Medical College of Wisconsin.

Corresponding Author:

Jake Luo, PhD

Associate Professor

Director of Health Care Informatics Graduate Program

Department of Health Informatics and Administration

College of Health Sciences

University of Wisconsin, Milwaukee

Email: jakeluo@uwm.edu

Postal Address: 2025 E Newport Ave 6469, WI, 53211, Milwaukee, USA

Interpretable Machine Learning Text Classification for Clinical Computed Tomography reports – A Case Study of Temporal Bone Fracture

Table of Contents

Response to Reviewers' Comments:

Executive Summary

1. We thank the reviewers for their careful reading of the manuscript and their constructive comments. We carefully considered all the comments and made significant revisions to improve and clarify the manuscript. Because we extensively revised the manuscript, we summarized the major changes in "Major changes of the revised manuscript" for your convenience of reading.

   We also tracked changes and **underlined** the modified texts of the manuscript. We used two types of color to indicate different types of changes. The red texts indicate grammatical, phrasal, or sentence structure changes during rewriting, editing, and proofreading. The green texts indicate the change is related to the reviewer's comments. Please note that the text colors are only used in the marked version of the revised manuscripts. In addition to the marked edition, we submitted a clean copy of the revised manuscript. The clean copy of the manuscript does not include these colors and brackets.

Major Changes of the Revised Manuscript

First, we summarized out changes of manuscript in this revision from five aspects:

Grammatical and Structural Changes

We consulted the writing center and revised the article to the best of our effort. The revised manuscript has significant improvements in narratives, expressions, and word choices.

a. We made substantial grammatical and stylistic changes and chose more precise words. We corrected all grammatical errors. We rewrote long sentences and split them into two or more shorter sentences. We also paraphrased hard-to-understand sentences to make them fluent. The goal is to make sure that our manuscript is easy to read and free of grammatical mistakes.
b. We made extensive edits to the introduction, related work, methods, results, discussion, and conclusion sections suggested by a proofreader. In this way, we believe we improved the writing quality of the manuscript.
c. In addition to grammatical changes, we also made significant structural and paragraph changes to the manuscript. As the reviewer suggested, we deleted the redundant sentences. We merged similar paragraphs in each subsection. We have rewritten the abstract to better differentiate among the objectives. We edited the connections among the introduction, related work, methods, results, discussion, and conclusion sections, making the entire story more connected. We believe our structural changes significantly improve the ease of reading, and our writing can highlight our study's significance.

Content Changes

As the reviewer suggested, we added the following new paragraphs to complete the baseline model: boost existing algorithms, explain more experimental details, conduct a more in-depth analysis on word predictors, draft a new analysis section on word interpretations, and draft new paragraphs on discussions, conclusions, and limitations.

a. Title change:

   We added a word "clinical" in the manuscript title, so the title becomes "Interpretable Machine Learning Text Classification for Clinical Computed Tomography reports – A Case Study of Temporal Bone Fracture". We would like to emphasize the clinical importance of our model.

b. Evaluate the settings of the baseline model:

   We deleted the decision tree model from the baseline experiments as suggested by the reviewer. For pre-processing part, we added additional descriptions to include the used packages and codes.

c. Boost existing algorithms:

   We added the XGboost model to provide a newer machine learning model. The XGBoost can represent how boosting methods perform on our data set and provide more details about the models' performances.

d. Explain more experimental details in Methods:

   (1) We added the justification of why we chose Bag-of-Words and TF-IDF techniques to convert clinical texts into numeric matrix formats.

   (2) We also described the erroneous part of the description on cross validation and described the stratified k-fold methods appropriately.

   (3) We added additional details about the number of clinical documents, the number of training sets, the number of test sets, and the total number of features used for the baseline model.

e. Conduct a more in-depth analysis of word predictors:

   We rewrote the section "Interpretation of Machine Learning Models" in the results and re-analyzed the result of word frequency score in the discussion.

f. Draft a new analysis on word interpretations

We emphasize the importance of machine learning model interpretability. In the discussion, we compared a few examples of black box models and transparent models and how they differ in terms of interpretability.

g. Delete a paragraph from discussion:

   We deleted a paragraph regarding the discussion of the transparent model and the black box model, and which is better. Because this is a highly contentious issue that is not related to clinical document interpretation, We decided to remove them from the manuscript so it could focus on the topic of clinical document interpretation.

h. Modify conclusion

   We refined the conclusion and removed those irrelevant texts that were part of our major objective. In conclusion, we firstly discuss our experiments, including the frequency analysis and LIME application. We then state how our experiment could support computerized decision-making in clinical settings..

i. Draft a new paragraph on Limitations and future work

   We modified the narrative of Limitations. We only included one future work, as reviewer suggested, we added a second essential future direction of use highly specialized medical keywords to interpret the machine learning classification results. This direction is more valuable and will have much potential to integrate machine learning models to the clinical scenario.

Tables and Figures Changes

As a reviewer suggested, we merged charts and figures to provide a more streamlined reading experience. We replotted most figures and provided higher resolution images. We redesigned some figures to clearly show the text and numbers. We also rewrote the figure captions to indicate the illustration changes. We deleted table 1, because we do not need discussion of the pros and cons of interpreting transparent models or using LIME to explain black box models. We believe our revised tables and figures serve as strong supporting evidence for the hypothesis and conclusions.

This table summarized which figures are kept in the main text, which figures were deleted, and which figures were moved to the supplemental files.

Changes	Which Chart(s)
Kept on the main text	• Overview of our study • Comparison of gaps between fracture and non-fracture reports • Relationships between classification model's performance, number of selected features, and evaluation performances • Merge two figures: Text Explainer's Evaluation (original Figure 7) and Top Features and its weight by Text explainer (original Figure 8) • A visualization of decision trees.
Deleted chart	• Overview of Text Explainer, how text explainer explains clinical CT documents
Moved to supplemental files	• Accuracy score and the Kullback-Leibler divergence score to Evaluate the reliability of LIME Interpretation • AUROC Figure of random forest model • Classification Performance for Support Vector Machine~~, Decision Tree~~ XGBoost, Logistic Regression and Random Forest Models

Supplemental Files

As the reviewer suggested, we moved some parts of the manuscript paragraphs into the supplemental appendix. All supplemental files will be publicly available for downloading when the manuscript is finally published. Please see the "Supplemental Files" Section to download all files at the end of this manuscript.

In supplemental files, we provided the following data, code, tables, and figures:

A. The de-identified data of 164 temporal bone fracture CT reports, including 119 negative bone fracture reports and 45 positive bone fracture reports. Please search for "Supplemental X" to access this file.
B. The source code of Jupyter Notebook to implement the baseline model, conduct word-level analysis, and provide LIME interpretations. Please search for "Supplemental X" to access this file.
C. Source file for Figure 1: Source File of Study overview.
D. Source file for Figure 2: Comparison of word frequency gaps between fracture and non-fracture reports
E. Figure 3: performance evaluation of the number of keywords: the relationship between four machine learning models, the number of features in the dataset, and evaluation performance. We provided four models based on random forest, SVM, and logistic regression algorithms, correspondingly. For each model, we evaluate the performance based on the top 2, 3, 4, 5, 10, 20, 30, 50, 100, 200, 300, and 500 words as features in the training set. For each model, the performance values included accuracy, F1-score, precision, and recall. Please check "Supplemental Files" section
F. Data source for figure 2: List of all word frequency of positive and negative keywords
G. Evaluation of the LIME model: accuracy score and Kullback-Leibler divergence
H. The AUROC Figure for performances.
I. Classification Performance for Support Vector Machine, XGBoost, Logistic Regression and Random Forest Models, including Stratified 10-fold results.
J. Figure 4: Comparison of Performance between four models
K. Source file for rule-based Classifier. To answer reviewer 1's comment

L. An intuitive visualization of the stratified 10-fold cross validation, to answer reviewer 2's comment.

M. A rule-based classifier's visualization and performance calculation, to answer reviewer 1's comment.

Also, please see below for a detailed point-by-point response to all the reviewers' comments.

Response to Reviewer 1's Comments

Summary

- Reviewer 1: I had the pleasure to review the manuscript "Interpretable Machine Learning Text Classification for Computed Tomography reports - A Case Study of Temporal Bone Fracture." In this investigation, the authors demonstrated the application of Random Forest (RF), SVM, and decision trees classifier combined with commonly used NLP methods (BOW, TF-IDF) in classifying fracture cases from non-fracture cases. Word Frequency Score (WFS) and LIME were used for interpretability. A limitation of this paper, as previous reviewers have noted, is that the sample size was only 164 reports, which hinders the experimental conclusions. In addition, these 164 reports came from a single center with patients in a very small age range (60-65), which severely limits the clinical value of the proposed models. Despite its limitations, I think this manuscript is still a solid paper that illustrates a use case for how tools in explainable AI could be used to improve the transparency of ML models made for clinical purposes. The authors have revised and answered previous reviewers' comments appropriately. My main feedback includes the need for a baseline model, try boosting algorithms, and more explanation of the top predictors as well as rephrase some sentences in the discussion and conclusions.

Response: Thank you for your kind comment! We believe our revised manuscript has improved descriptions of the baseline model, boosting algorithms, and more explanation of predictors. We also make our best effort to rephrase some language parts.

Major Comments

1. In the 'Development of ML models' section, please explain why you set the minimum frequency limit threshold to 4 and the maximum frequency to 70%. Shouldn't these parameters be optimized via a parameter search?

Response: These parameters could have been optimized via a parameter search. However, we use these numbers based on our observation of word frequency analysis and a full list of top frequency words. Please check the list in the supplemental files. From the list of words and the list of frequencies, we find that a minimum frequency of five and a maximum frequency of 70% is a threshold to produce a reliable performance. Searching for the best optimum parameter is possible, but the optimized parameters are not likely to change the interpretation results. Because our study focused more on interpreting results, we relied on reasonably good parameters to create a dataset.

Also, please see the pieces of code and explanations from the Jupyter Notebook. The notebook is available in the supplemental files.

Converting Text to numbers

Machines, unlike humans, cannot understand the raw text in this format. Machines can only see numbers. Particularly, statistical techniques such as machine learning can only deal with numbers. Therefore, we need to convert our text into numbers.

Different approaches exist to convert text into the corresponding numerical form. `The Bag of Words Model` and `the Word Embedding Model` are two of the most commonly used approaches. In this article, we will use the bag of words model to convert our text to numbers.

Bag of Words

The following script uses the bag of words model to convert text documents into corresponding numerical features:

```
In [19]: from sklearn.feature_extraction.text import CountVectorizer
         vectorizer = CountVectorizer(max_features=500, min_df=5, max_df=0.7,
                                      stop_words=stopwords.words('english'))
         X = vectorizer.fit_transform(documents).toarray()

In [20]: X.shape
         # which means 164 patient's reports, each reports contains 500 features.

Out[20]: (164, 500)
```

The script above uses `CountVectorizer` class from the `sklearn.feature_extraction.text` library. There are some important parameters that are required to be passed to the constructor of the class. The first parameter is the `max_features parameter`, which is set to 1500. This is because when you convert words

to numbers using the bag of words approach, all the unique words in all the documents are converted into features. All the documents can contain tens of thousands of unique words. But the words that have a very low frequency of occurrence are unusually not a good parameter for classifying documents. Therefore, we set the `max_features` parameter to 1500, which means that we want to use 500 most occurring words as features for training our classifier.

The next parameter is `min_df` and it has been set to 5. This corresponds to the minimum number of documents that should contain this feature. So we only include those words that occur in at least 5 documents. Similarly, for the `max_df`, feature the value is set to 0.7; in which the fraction corresponds to a percentage. Here 0.7 means that we should include only those words that occur in a maximum of 70% of all the documents. Words that occur in almost every document are usually not suitable for classification because they do not provide any unique information about the document.

Finally, we remove the `stop words` from our text since, in the case of this analysis, stop words may not contain any useful information. To remove the stop words we pass the `stopwords` object from the `nltk.corpus` library to the `stop_words` parameter.

The `fit_transform` function of the `CountVectorizer` class converts text documents into corresponding numeric features.

2. For the parameter search of the number of features, where did the feature ranking (from top most feature to the bottom feature) come from?

Response: The feature ranking comes from the frequency of the words that occurred in our entire clinical document set. In other words, we used the 500 most frequently occurring words as a feature to train the classifier. The most occurring words, of course, exclude any stop words listed in the natural language toolkit (NLTK), so these top words show semantics for positive and negative documents.

3. The explanation of LIME in the 'Interpretation of ML models' sounds vague. Please explain how the local model was constructed.

Response: We used the LIME package and included the text explainer. Our pipeline completely follows the pipeline on the documentation example. Please check the documentation for Text Explainer.

https://eli5.readthedocs.io/en/latest/tutorials/black-box-text-classifiers.html#textexplainer

To answer the question of how the local model was constructed, Text Explainer generated texts similar to the document (by removing and sampling words in the document), and then trained a

white-box classifier which predicts the output of the black-box classifier. The explanation we saw is for this white-box classifier.

This approach follows the LIME algorithm; for text data, the algorithm is pretty straightforward:

1. generate slightly modified versions of the text.
2. predict probabilities for these modified texts using the black box classifier;
3. train another classifier (one of those eli5 supports) which tries to predict output of a black-box classifier on these texts.

The algorithm works because even though it could be hard or impossible to approximate a black-box classifier globally (for every possible text), approximating it in a small neighborhood near a given text often works well, even with simple white-box classifiers.

Because the local model is always an approximation of a black-box model, we included the mean_KL_divergence and accuracy score in the white-box classifier. It usually assigns the same labels as the black-box classifier on the dataset we generated, and its predicted probabilities are close to those predicted by our machine learning pipeline. The most valuable thing is that the local model could provide weights and word-level visualizations for the contribution of fracture classification.

As for the problem of *'Interpretation of ML models' sounds vague,* we described the concept in a clearer way in the manuscript: LIME provides the word-level evaluation on how each word contributes to the classification results.

4. Regarding Fig 3: what is the significant of 'left', 'right', and 'canal' in terms of interpretability? An explanation for why these words is important and meaningful to a physician would be beneficial.

Response: Please see an example of documents, which shows the words "left", "right," "fracture," "canal".

Document #1:

1. Old comminuted fracture of the right middle cranial fossa with multiple bullet fragments lodged within it as described above. There is disruption/disolution of the right ossicular chain but the inner ear structures are intact.

2. Adjacent residual right mastoid air cells are chronically opacified.

Examination reviewed by Dr. Guleria and reported findings confirmed by Dr. Rand.

Clinical Indication: Post traumatic right otalgia.

Techniques: 0.625 mm thick contiguous axial scans of temporal bones were acquired. Coronal reformats were generated and reviewed.

Comparison: None.

Findings: Again visualized are severely comminuted old fractures of the right middle cranial fossa with multiple bullet fragments within the bones of the right skull base the right middle ear cavity and right anterior mastoid air cells. The roof of the right middle a year cavity is dehiscent and there are dislocations/resorptions of components of the right ossicular chain. Only the body of the right incus is well visualized.

Multiple bullet fragments are lodged within the clivus sphenoid bone prevertebral soft tissues and in the infratemporal fossa. The residual mastoid air cells are opacified.

The left mastoid air cells appear well-aerated. The left middle ear cavity and ossicular chain are preserved. The left mastoid air cells appear unremarkable.

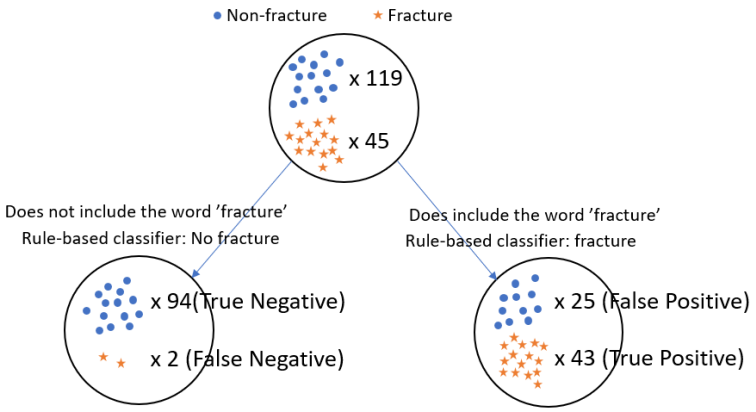
Bilateral inner ear structures appear normal in morphology and density. Internal auditory canals appear symmetrical and normal in size bilaterally. Vestibular aqueducts are not dilated.

The terms "left" and "right" "canal" appear frequently in this document as the condition of parts of anatomical words. These anatomical concepts are often a major part of descriptions in temporal CT clinical documents. A hypothesis is that once a fracture is discovered, physicians tend to provide more detailed descriptions of the anatomical parts and tissues. In clinical settings, fracture diagnosis only consists of a small percentage of documents. Physicians may take more time on abnormal CT images, and draft more detailed documents, thoroughly describing all parts of the body system. This makes the "right" and "left" "canal" show up more often in fracture texts. For non-fracture documents, we hypothesize that physicians tend to use more summarized words, such as "All [parts] appear intact, in normal morphology and density." This description is more likely to be non-fracture documents, will be shorter, and will use fewer anatomical terms.

5. Regarding Fig 3: This figure suggests that the high precision probably comes from the fact that fracture reports typically contain words like 'fracture' or 'temporal.' Also, words like 'lung' or 'calcification' imply a non-fracture case because the CT images are of the lungs or the heart. This made me think perhaps a simple rule-based model for these specific words may suffice the classification task. Would be nice to include a model as such to compare with the existing models.

Thank you very much for your suggestion. We use the word "fracture" to build a simple one-rule classifier. We wanted to keep the rule simple, because it is common to conduct keyword searches and quickly determine the classifications. This would be a good baseline to reflect the actual scenarios for bone fracture classification.

In this case, the rule-based classifier would classify documents with "fracture" as positive and documents without "fracture" as negative. This would serve as a baseline model. In supplemental files, we included the rule-based classifier and used the "if" condition to construct the classifier on our documents. We then count the true positive, false positive, true negative, and false negative cases. We measure F1, precision, and recall. Please see the following figure:



When we adopt "fracture" as the only rule of the rule-based classifier:

Precision = TP/(TP+FP) = 43/(43+25) = 0.632

Recall = TP/(TP+FN) = 43/(43+2) = 0.956

F1 = 2*Precision*recall/(precision + recall) = 0.761

Also, we provided a supplemental file named "rule-based classifier" that provided case-by-case prediction results and rule-based model details.

From the rule-based model's result, the F1-score is 0.761, which is significantly lower than our machine learning-based classifier. TP, FP, FN, and TN cases, we see more false positive cases

than false negative cases. This indicates even a simple rule-based classifier will not be likely to miss a fracture diagnosis. The precision is a bit lower than our machine learning model. To increase the precision, therefore, it makes sense to build more complicated rules or to adopt the most frequent words as machine learning features.

As we investigated in related work, the development of rule-based classifiers was mostly in the 1990s. It must be admitted that a simple rule-based classifier cannot handle complex clinical text classification tasks. If we apply multiple rules, the performance may improve, but the interpretation becomes a problem again. The rule-based classifier may not adapt to the ever-changing word usage in medical documents. The machine learning models, however, can overcome this problem. Therefore, building more complicated rules is no longer a focus of our study. We may not use a 20-year-old rule-based classifier as a baseline. Instead, starting from the machine learning model would be a better choice. Based on the rule-based classifier's performance, the ML model's development convenience, and comparisons, we decided to not include the rule-based classifier as a baseline model.

6. Regarding Fig 4: The comparison between RF, SVM, Decision Tree, and Logistic Regression provides little values. Have you tried boosting methods (XGBoost, LightGBM, CatBoost)? You could make the text sizes in the axes and legends bigger.

Thank you for the suggestion! We included an extra XGBoost model and provided performance values in two figures. Generally speaking, the XGBoost model shows comparable performance to the Random Forest model. The precision, recall, and F1 values are similar. The XGBoost model's performance has been added to the revised figures.

7. In addition, you mentioned that from these results that you chose 500 as the number of topics, but the RF plot suggests that 20 is the best number.

Response: Thank you for pointing out this issue. Please see the graph below to see the relationship between the machine learning model, the number of features in the dataset, and the Evaluation Performances. This figure is a simplified visualization of previous figures. The performance metrics did not change.

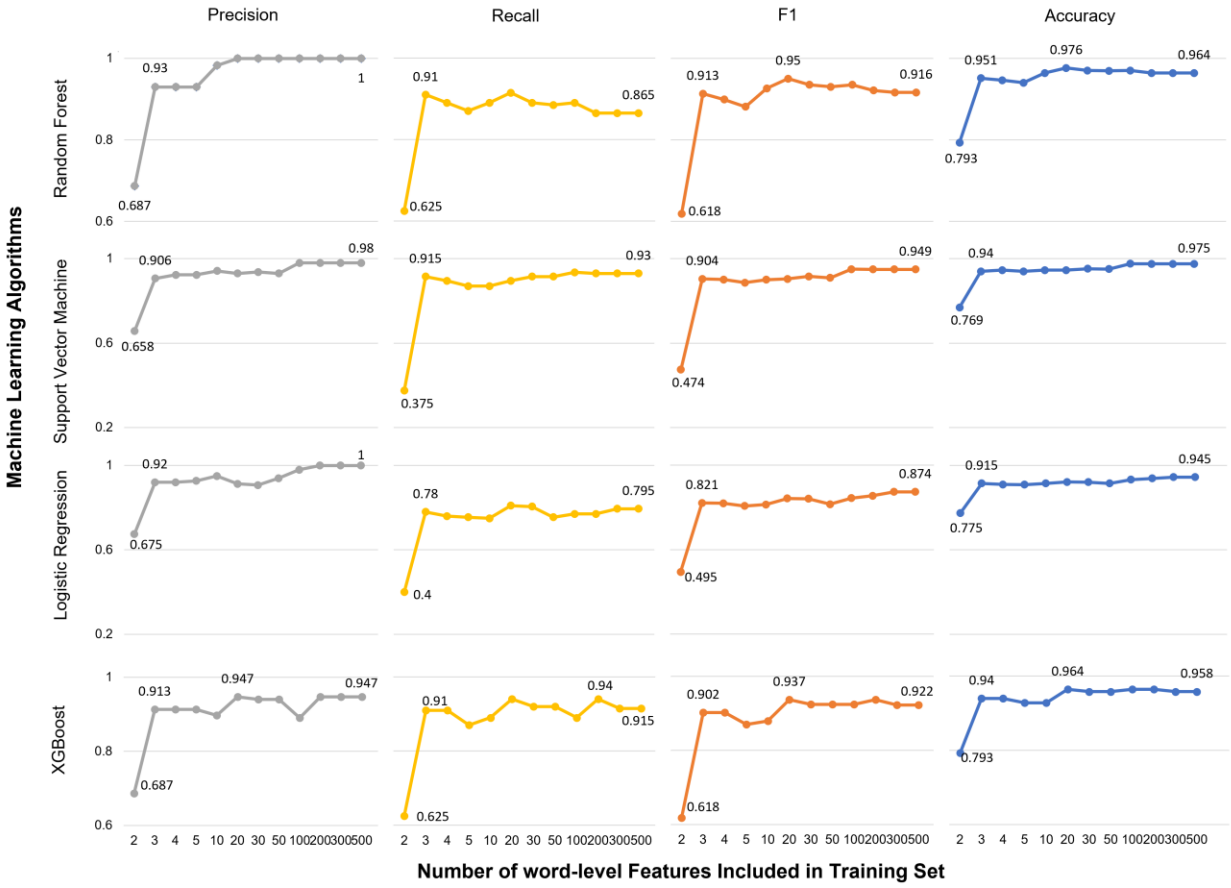


Figure 4: Relationships between classification model's performance, number of selected features, and evaluation performances for Random Forest, XGboost, Support Vector Machine, and Logistic Regression model.

In this set of figures, we can quickly find that most performance is better as the number of selected words gets larger. Therefore, we keep our main models to a 500-word limit.

8. I think either Fig 4 or Fig 5 should be in the main body, and the other one could be moved to the Supplement.

Response: We moved figure 5 into the supplemental files section. Figure 3 would show the model's performance from comprehensive perspectives.

9. Regarding Fig 6: The AUC is very high, that made me think whether the problem of classifying fracture/non-fracture cases is challenging enough to utilize ML. If it is a relatively simple problem, a simple model would suffice. This could be answered by including a simple rule-based model in the analysis to serve as a baseline.

We built a rule-based classifier and reported the performance. Please refer to the response in comment 5.

Here is the answer to the question of whether the problem of classifying fracture and non-fracture cases is challenging enough to utilize ML: Machine learning models need to be used because they work much better than a simple rule-based classifier.

10. Figure 7 and 8 could be merged into one. Similarly, some figures are redundant (see below), and 10 total figures is a lot. Please merge some figures and/or put some in the Supplements.

11. Fig 10 is nice. Not a new method but a nice and definitely interpretable visualization to summarize the logic behind a model.

Comment 10 and 11 are both related to figures, so we combined them and responded to the comments together:

After we discussed it with the coauthors, we agreed to make the following edits to the charts:

Action	Which Chart(s)
Leave these figures on the main text:	• Overview of our study • Comparison of gaps between fracture and non-fracture reports • Relationships between classification model's performance, number of selected features, and evaluation performances • Merge two figures: Text Explainer's Evaluation (original Figure 7) and Top Features and its weight by Text explainer (original Figure 8) • A visualization of decision trees.
Delete this chart:	• Overview of Text Explainer, how text explainer explains clinical CT documents
Move these to supplemental files:	• Accuracy score and the Kullback-Leibler divergence score to Evaluate the reliability of LIME Interpretation • AUROC Figure of random forest model • Classification Performance for Support Vector Machine~~, Decision Tree~~ XGBoost, Logistic Regression and Random Forest Models

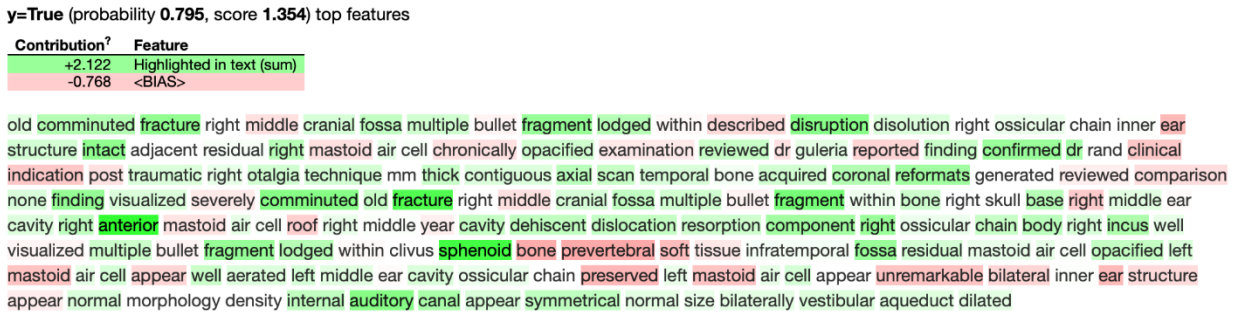
Please see our updated manuscript for detailed changes of the charts.

12. I appreciate the author's answer to a reviewer's comment regarding the reason that authors used BOW instead of Word-2-vec. Would be nice to include it in the main text.

Response: Thank you for the suggestion. We moved the justification of using BOW instead of word-2-vec into the main text.

13. From my understanding, LIME can only explain individual reports, and not globally. Since LIME trains a local linear model around the individual prediction to approximate the prediction of the 'black box' model. Thus, I am confused as whether the model could get an aggregated explanation?

Here is the original figure 7 and figure 8:

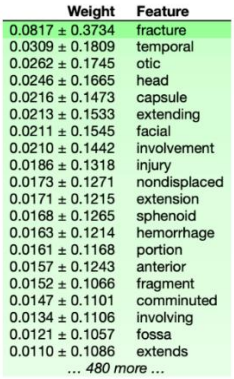


y=True (probability **0.795**, score **1.354**) top features

Contribution[?]	Feature
+2.122	Highlighted in text (sum)
-0.768	<BIAS>

old comminuted fracture right middle cranial fossa multiple bullet fragment lodged within described disruption disolution right ossicular chain inner ear structure intact adjacent residual right mastoid air cell chronically opacified examination reviewed dr guleria reported finding confirmed dr rand clinical indication post traumatic right otalgia technique mm thick contiguous axial scan temporal bone acquired coronal reformats generated reviewed comparison none finding visualized severely comminuted old fracture right middle cranial fossa multiple bullet fragment within bone right skull base right middle ear cavity right anterior mastoid air cell roof right middle year cavity dehiscent dislocation resorption component right ossicular chain body right incus well visualized multiple bullet fragment lodged within clivus sphenoid bone prevertebral soft tissue infratemporal fossa residual mastoid air cell opacified left mastoid air cell appear well aerated left middle ear cavity ossicular chain preserved left mastoid air cell appear unremarkable bilateral inner ear structure appear normal morphology density internal auditory canal appear symmetrical normal size bilaterally vestibular aqueduct dilated

Figure 7: Text Explainer's Evaluation on word's Contribution using Random Forest Algorithm



Weight	Feature
0.0817 ± 0.3734	fracture
0.0309 ± 0.1809	temporal
0.0262 ± 0.1745	otic
0.0246 ± 0.1665	head
0.0216 ± 0.1473	capsule
0.0213 ± 0.1533	extending
0.0211 ± 0.1545	facial
0.0210 ± 0.1442	involvement
0.0186 ± 0.1318	injury
0.0173 ± 0.1271	nondisplaced
0.0171 ± 0.1215	extension
0.0168 ± 0.1265	sphenoid
0.0163 ± 0.1214	hemorrhage
0.0161 ± 0.1168	portion
0.0157 ± 0.1243	anterior
0.0152 ± 0.1066	fragment
0.0147 ± 0.1101	comminuted
0.0134 ± 0.1106	involving
0.0121 ± 0.1057	fossa
0.0110 ± 0.1086	extends
… 480 more …	

Figure 8: Top Features and its weight by Text explainer using Random Forest Algorithm

First, we would like to clarify that figures 7 and 8 are not relevant. We use the word "contribution" in figure 7, and "weight" in figure 8. We recognize that putting Figure 7 and Figure 8 together may cause confusion very easily. Here are explanations for figures 7 and 8:

Figure 7 is a visualization of a text explainer (https://eli5.readthedocs.io/en/latest/tutorials/black-box-text-classifiers.html#textexplainer). The text explainer's only job is to evaluate each word's contribution to each document. The text explainer's evaluation is based on an approximated white-box model, as we mentioned. In the explainer, each word's contribution could be positive (shown in green) or negative (shown in red). The red word means a contribution to negative classifications. The green word means a contribution to positive classification. For example, in Figure 7 text example, we can say the words "comminuted" and "fracture" contribute to a positive result. "Unremarkable" contributes to negative results. In other words, in the text explainer, the word has two directions. Therefore, figure 7's contribution is only related to each document. Figure 7's word-level contributions are not aggregated evaluations.

In Figure 8, LIME evaluated the weight of the completed model based on a global view. The weight is based on the entire document set and does not relate to any form of classification results. We can say that "weight" is an alternative expression of "feature importance". Of course, they are based on different algorithms, but both key concepts are to evaluate which feature serves as a deciding factor in classification.

The "fracture" has a weight of 0.0817. In Figure 8, the information only shows "fracture" is often used as the primary deciding factor in decision trees from the random forest. "Fracture" serves as a key deciding criteria in many decision trees. Therefore, the weight is calculated based on the completed machine learning model, which is trained based on the entire training set. Therefore, the figure 8 weight list is based on aggregated evaluations.

We thank you for your comments. With your comment, we find that our paragraph may cause confusion in the discussion section "Interpretation of Machine Learning Models." To avoid confusion, we provide the following additional clarification in the main text:

Figure 7 shows a visualization of a text explainer. In this text explainer, each word has been assigned a contribution score, showing the words lead to positive or negative classification. The text explainer's evaluation is based on the individual text level.

The feature list in Fig. 8 displays the random forest model's most important word list. The word list is aggregated from multiple decision trees. The selection of each word is calculated by whether the word serves as a deciding factor in a decision tree. Higher weight words are often used as a key factor in the classification results. The feature list corresponds to the text explainer's assessment of figure 7.

14. Regarding the second paragraph of the 'Summary of Text feature analysis' in the Discussion comment: I don't quite agree with the remark on using highly specialized language patterns are highly conserved. Sometimes it is the case that clinical texts are unavoidably full of domain-specific keywords, and use of a specialized corpus instead of a general one could achieve higher performance.

Response: We agree with your argument. Sometimes, clinical texts are unavoidably full of domain-specific keywords.

Use of a specialized corpus instead of a general one could achieve higher performance. We included this direction as a part of our future work. Our work's next step is to only visualize the medical specialized words in clinical documents. This can reduce the effect of other irrelevant words and make a more accurate prediction.

Regarding our statements in the last paragraph of the 'Summary of Text feature analysis' in the Discussion, we have modified the texts, and added the following statements:

Physicians often draft reports with highly specialized medical terms. These medical terms often serve as reliable predictors. According to our results, we suggest investigating if non-experts can easily interpret the medical terms.

Our LIME results show that specialized medical words contribute significantly to classification. As a result, we believe that interpretable AI has the potential to help explain more complex diagnostic conditions. This pipeline can also be used to interpret other clinical texts, classify diagnoses, and give an explanation that is easy to understand.

15. I don't quite agree with the claim in the Discussion that LIME is significantly easier to interpret than decision tree, and many would say the same. This claim is too strong in my opinion. Each has its own merits. Perhaps LIME could be slightly more suitable for non-numeric data.

Response: Thank you for your suggestion. However, we believe our claim is reasonable, especially for text data. We did not change our claim but have modified the claim slightly. Our scope of discussion focused on text data.

We believe LIME would be significantly more explainable than a decision tree. Our decision tree example only includes five keywords. However, a real decision tree has significantly more depth and more nodes. A short decision tree like this example is easy to understand, but in most clinical cases, the tree must be very deep to achieve satisfactory performance. The actual decision tree often reaches 100 or more nodes. In this case, the physician must iterate through ten levels of depth and choose from ten different rules. This is not practical. If a similar LIME network was constructed, the visualization would only show a list of weights like in figure 8. This list makes it easier for physicians to determine the importance of each word to the classification results.

16. Regarding Table 1: not quite sure what it means for inherently transparent models to require deep AI knowledge. Please elaborate or word differently. In addition, there are several papers that argue otherwise (pro for fully transparent models and against post-hoc methods). This is a highly controversial topic that is also not quite related to the sole purpose of interpretable classification of fracture reports from texts.

We agree that it is a highly controversial topic about choosing which forms of models to use. After carefully considering your comments, we believe deleting this part of the discussion would help focus on our topic of "Interpretable models on clinical document classification". As you mentioned, the topic of selecting which types of models is outside the scope of this study. Therefore, we decided to delete that part of the discussion. The deletion allows the paper to focus on the LIME and the WFS analysis parts of the work.

17. Regarding the conclusions: conclusion #2 about RF achieved the highest accuracy, in my opinion, is not novel nor significant.

Response: Thank you for the comments. We deleted the RF part and focused more on the significance of our model's interpretation in the conclusion.

Minor Comments

18. Regarding Text pre-processing: please provide in the Methods more details on which parts in the preprocessing step were automated and which packages were used for them. It will only take one or two sentences.

Response: We added a brief description of the text processing on the section about the following text pre-processing steps.

We removed all non-word elements in clinical reports, including numbers, punctuation, and special characters. We converted all words to lowercase letters. We perform these changes using regular expressions. Then, we followed a Natural Language Toolkit stop-word list to remove stop words, lemmatized words, and incorrect spellings and acronyms. All words were free of noun declination and verb conjugations. The supplemental code book shows how we pre-process the documents.

```
In [42]: X, y = df['text'], df['fracture']

In [43]: documents = X
         import nltk
         nltk.download('wordnet')
         from nltk.stem import WordNetLemmatizer

         stemmer = WordNetLemmatizer()

         for report in range(0, len(X)):
             # Remove all the special characters
             document = re.sub(r'\W', ' ', str(X[report]))

             # remove all single characters
             document = re.sub(r'\s+[a-zA-Z]\s+', ' ', document)

             # remove all numbers
             document = re.sub(r'[0-9]+', ' ', document)

             # Remove single characters from the start
             document = re.sub(r'\^[a-zA-Z]\s+', ' ', document)

             # Substituting multiple spaces with single space
             document = re.sub(r'\s+', ' ', document, flags=re.I)

             # Removing prefixed 'b'
             document = re.sub(r'^b\s+', '', document)

             # Converting to Lowercase
             document = document.lower()

             # Lemmatization
             document = document.split()

             document = [stemmer.lemmatize(word) for word in document]
             document = ' '.join(document)

             documents.append(document)
```

We thank you very much for your thoughtful review and valuable comments.

Response to Reviewer 2's Comments

Summary

Reviewer 2: The paper uses four machine learning techniques, Support Vector Machine(SVM), Decision Tree(DT), Logistic regression, and Random Forest, to do binary classification on a case study of Temporal Bone Fracture and classify 164 Electronic health reports into fracture cases and non-fracture cases.

Although the dataset size seems small, I have seen great articles for different diseases, such as "Clinical text classification of Alzheimer's drugs' mechanism of action," that perform the same task on a small dataset. Having small datasets for clinical text classification is common, and the way that the paper uses Machine Learning techniques instead of Neural Network-based (NN) models or transformers makes sense to me because using NN models and fine-tuning transformers needs large datasets.

However, the final goal of the paper is interpreting the models not achieving higher accuracy. So that is excellent work.

Response: Thank you for your comments! We would like to discuss a few points:

1. We analyzed word frequency and the gaps between negative and positive documents. We supplied the analysis in supplemental files. In the analysis, there are fewer than 1000 valid words as features. For this size, we think neural network-based models or transformers would not achieve an advantage over machine learning models. Should we include a larger set of clinical documents and a more complex classification task, the neural network might achieve a significantly better outcome. In future work, we will consider neural networks and transformers and apply them to a larger set of clinical documents to process complex classification tasks.
2. We did not notice the paper "*Clinical text classification of Alzheimer's drugs' mechanism of action*" in the literature review. They are very valuable work. In this revision, we added the relevant publication and our comments in related work section.

Major Comments:

1. If the paper's work is on interpretability, which I believe it is, I strongly recommend that the authors remove the Decision Tree (DT) from the models because, in essence, it is not a black-box model. It is a white box model. Moreover, One category of Stanford's interpretability proposed methods is trying to draw a DT model and write an Interpretable decision set (IDS), so considering the DT model as a block-box model is entirely wrong.

Response: Thank you for pointing out the issue. We have removed all content related to decision trees from the models and performance evaluations. However, we still like to discuss the model interpretability in decision trees and use decision trees as an example of a white box model. because we focus on the model's interpretability. It is essential to mention that the decision tree is unique because it is an intrinsically interpretable machine learning model. We would like to compare this model with other explanation techniques, such as LIME.

2. On the other hand, Both RF and SVM models are black box, so I believe your job on these two models makes sense.

Response: Thank you for your comments.

3. You need to talk about the dataset details. It is the only comment you have not addressed correctly. I know you provided the code, but I assume that the results are unreliable if the dataset is imbalanced in small datasets. Therefore, the interpretation is not correct. Please add a table and provide information about the train, test, and evaluation size.

Response: First, we apologize for not using the most precise words in the manuscript. The 10-fold cross validation should be stratified by the 10-fold cross validation. The procedure was shown in the Jupyter notebook.

We acknowledge that we did not provide sufficient information on training, testing, and validation in the last revision. We are sorry if there is any confusion in the manuscript, but we believe our corrected language has reflected the actual case after revision. We added new descriptions of the cross validation and train/test split percentages so our study can be reproduced. For the concerns of imbalanced data and small data sets, we believe our experiment and manuscript provided the correct interpretation to the maximum extent possible. If the manuscript needs more description, please inform us what part of the model creation information is critical to the reader. If you have concerns about codes, please suggest modifications to our codes. We will perform additional experiments.

In this study, we adopted stratified k-fold cross-validation to provide a reliable model performance. It is a technique to stratify the sampling by the class label, and this technique can tackle small and imbalanced datasets.

For more details of stratified k-fold, please see the response to Comment 6.

4. Please define the acronym for the first time the sequence is mentioned in the text and avoid using the whole sequence or redefining it in the rest of the text. For example, once you write Support Vector Machine (SVM), do not redefine the acronym again; for the rest of the article, use SVM.

Response: Response: Thank you. We have corrected all acronym issues in the manuscript. We removed a few infrequently used acronyms. The removed words are electronic health record (EHR), natural language processing (NLP), and bag of words (BOW). We revised them accordingly for the rest of the words..

5. In the Limitation and future work section, you divide your future work into two aspects. You say the first one but never clearly mention the second one. I believe the paper's English needs to be checked by the writing center. Some sentences are not clearly described.

Response: We have re-stated the two aspects of our future work, as follows:

First, we used labeled data in this preliminary study. While unlabeled data cannot be used for classification, it has the potential for unsupervised learning. We believe that by building an appropriate unsupervised model, it is possible to cluster CT reports into two categories based on text reports. Second, building a medically specialized text interpreter would highlight medical words only and achieve a clearer interpretation. For example, by adopting SNOMED-CT standards [52], it is possible to create a medical text interpreter. The model could limit the number of words to only medical terms. The word-level optimization may achieve better prediction and better interpretation.

For the language issues, we have extensively looked for English editing and proofreading services, and we have revised the manuscript to the best of our effort.

Minor Comments

6. In the comments, you mentioned that the (45 patients + 119 patients) if these are the number of samples for fracture and non-fracture cases. Is the data imbalanced? How did you deal with this problem? How are you using 10-fold for a class with 45 samples?

Firstly, we acknowledge we did not choose the most precise words for the manuscript. The 10-fold cross validation should be stratified 10-fold cross validation. The procedure was shown in the Jupyter notebook.

The data is imbalanced. In this study, we used 164 clinical texts, with 45 positive cases and 119 negative cases. The small data set is common for clinical reports, and other studies use similar small data set. The study includes the one you mentioned "Clinical text classification of Alzheimer's drugs' mechanism of action". Many studies have adopted stratified k-fold to deal with the data imbalance problem. We used the same stratified 10-fold to avoid bias as much as possible.

A stratified K-fold is a cross-validator that divides the dataset into k-folds. Stratified is to ensure that each fold of a dataset has the same proportion of observations with a given label. Here is a figure that shows how stratified k-fold works:

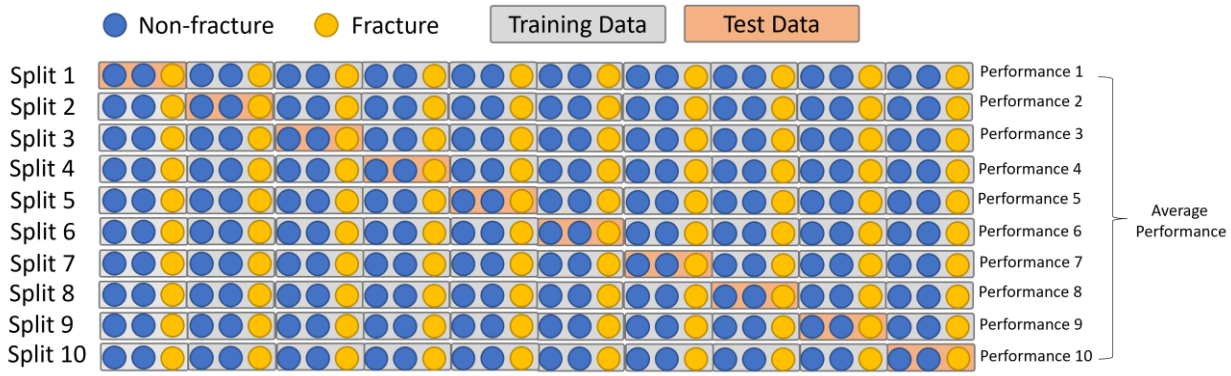


Figure 1. An example visualization of how stratified 10-fold splits the training set and test set

In this figure, fracture and non-fracture clinical texts are marked as yellow and blue dots, respectively. For easy plotting, we plotted 30 dots (20 negative, 10 positive), which is our entire data set. The stratified 10-fold cross-validation method divides the training dataset into 10 folds. The first 9 folds are used to train a model, and the 10th fold serves as the test set. This process is

repeated until each fold has a chance to be used as the holdout test set. A total of 10 models are fit and evaluated, and the model's performance is calculated as the mean of these runs.

Because stratified k-fold will evenly distribute the proportions of positive and negative cases and maintain a stable balance between test data and training data, It is widely accepted that stratified k-fold ensures that the proportion of positive to negative examples found in the original distribution is respected in all the folds. This is the best way to show an unbiased model's performance under a small data set.

As Dr Jason Brownlee writes in his blogs "How to Fix k-Fold Cross-Validation for Imbalanced Classification":

It is a challenging problem as both the training dataset used to fit the model and the test set used to evaluate it must be sufficiently large and representative of the underlying problem so that the resulting estimate of model performance is not too optimistic or pessimistic.

The two most common approaches used for model evaluation are the train/test split and the k-fold cross-validation procedure. Both approaches can be very effective in general, although train/test split can result in misleading results and potentially fail when used on classification problems with a severe class imbalance. Instead, the techniques must be modified to stratify the sampling by the class label, called stratified train-test split or **stratified k-fold cross-validation**.

As we rarely have enough data to get an unbiased estimate of performance using a train/test split evaluation of a model. Using a stratified k-fold cross validation procedure would introduce minimal bias under limited dataset. The procedure has been shown to give a less optimistic estimate of model performance on small training datasets than a single train/test split. A value of k=10 has been shown to be effective across a wide range of dataset sizes and model types.

Therefore, we hope our explanation and figures can address your concerns about the small and imbalanced dataset. In our experiment, we considered the balance and small data set and chose the appropriate methods to avoid potential problems.

Again, we thank reviewers for the careful reading of the manuscript and constructive comments. We hope our revised paper have addressed all concerns by the reviewers in sufficient details.

Interpretable Machine Learning Text Classification for Clinical Computed Tomography reports – A Case Study of Temporal Bone Fracture

Keywords:
Interpretable Machine Learning, Artificial Intelligence, Text Classification, Bone Fracture, Computed Tomography

Interpretable Machine Learning Text Classification for Computed Tomography reports – A Case Study of Temporal Bone Fracture

Ling Tong[1], Jake Luo[1], PhD, Jazzmyne Adams[2], MPH, Kristen Osinski[3], MS, Xiaoyu Liu[4], MBA, David Friedland[2], MD, PhD

1. Department of Health Informatics and Administration, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA
2. Department of Otolaryngology and Communication Sciences, Medical College of Wisconsin, Milwaukee, Wisconsin, USA
3. Clinical and Translational Science Institute, Medical College of Wisconsin, Milwaukee, Wisconsin, USA
4. Department of Electrical Engineering and Computer Science, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA

Highlights

- Random Forest classifier achieves 0.93 F1-score in classifying temporal bone fracture texts.

- Word Frequency Score reveals differences between fracture texts and non-fracture texts.

- Local Interpretable Model Explanations visualizes the word-level contribution to classification results.

- Our interpretable model provides physicians with accessible evidence for clinical diagnosis.

Interpretable Machine Learning Text Classification for Clinical Computed Tomography reports – A Case Study of Temporal Bone Fracture

Table of Contents

Acronyms

ML: Machine Learning

SVM: Support Vector Machine

WFS: Word Frequency Score

LIME: Local Interpretable Model-Agnostic Explanations

AI: Artificial Intelligence

CT: Computed Tomography

BOW: bag-of-words

TF-IDF: Term Frequency – Inverse Document Frequency

Abstract

Background

Machine learning (ML) has demonstrated success in classifying patients' diagnostic outcomes in free-text clinical notes. However, due to the machine learning model's complexity, interpreting mechanisms of classification results remains difficult.

Methods

We investigated interpretable representations of machine learning classification models. We created machine learning models to classify temporal bone fractures based on 164 temporal bone Computed Tomography (CT) text reports. We adopted the XGBoost, Support Vector Machine, Logistic Regression, and Random Forest algorithms. To interpret models, we used two major methodologies: (1) We calculated the average word frequency score (WFS) for keywords. The word frequency score shows the frequency gap between positive and negative classified cases. (2) We used Local Interpretable Model-Agnostic Explanations (LIME) to show the word-level contribution to bone fracture classification.

Results

In temporal bone fracture classification, the random forest model achieved an average F1-score of 0.93. WFS reveals a difference in keyword usage between fracture and non-fracture cases. Additionally, LIME visualized the keywords' contributions to classification results. The evaluation of LIME-based interpretation achieved the highest interpreting accuracy of 0.97.

Conclusion

The interpretable text explainer can improve physicians' understanding of machine learning predictions. By providing simple visualization, our model can increase the trust of computerized models. Our model supports more transparent computerized decision-making in clinical settings.

Introduction

Electronic Health Records have been acknowledged as a key to improving healthcare quality [1]. Computerized decision-making models are commonly used in clinical applications for disease discovery, identification, and prediction [2]. However, most current studies use structured features to build models. Unstructured data, such as free-text clinical notes, is rarely used. The limited use of free-text data is due to format issues [3]. For example, clinical texts require human-level intelligence to process complex linguistic rules, which goes beyond simple classification. To leverage clinical texts and build an accurate model, a common method is to label the clinical text. The manual process of free-text clinical notes and labels was inevitably expensive. The cost limited the wider use of free-text clinical notes.

Natural language processing [4] techniques are commonly used to build clinical classification models using free texts. Natural language processing mimics how humans learn a language by comprehending its semantics. Understanding natural language requires linguistic knowledge such as morphology, syntax, and pragmatics [5]. We have seen considerable progress in natural language processing and AI-based clinical decision-making classifiers [1]. However, understanding a model's mechanism requires extensive computer-domain knowledge [6]. Clinical practitioners need a simple method to understand the mechanisms of decision-making models.

Despite advancements in machine learning clinical classification models, only a few are used in clinical settings due to physician distrust. A common way to ensure machine learning classification's accuracy is to use a validation set [7]. However, a validation set must not replace real clinical contexts. Before using a computerized model in clinical practice, physicians must be confident that the decision-making model is applicable to patients in clinical settings. It is impossible to establish trust unless physicians understand how a model makes decisions based on medical domain knowledge. The lack of trust and transparency in decision-making models raises concerns about making incorrect decisions [3].

In this study, we aim to address this distrust by visualizing classifier interpretations. A set of untemplated narrative reports from temporal bone computerized tomography was used in our case study. These reports differentiate between those with and without fractures. Our visualization demonstrates that many aspects of clinical texts, including word frequency and word selection, will impact the final classification decision. For example, we demonstrate that some word presence,

such as ‘fracture,’ is the reason a classifier makes a positive classification. Using our visualization, physicians can combine medical domain knowledge with visualization to assess the validity of the highlighted keywords. Therefore, we believe that our visualization can boost physicians’ confidence in using classification models. This study could accelerate the adoption of ML-based decision-making systems in clinical settings.

Related Work

Clinical Text Classification

The development of automated medical text classification systems can be traced back to the 1990s [8]. Early studies focused on rule-based methods to build classifiers for medical documents [8]. For example, Aronow et al. (1999) [9] developed NegExpander, a computerized system that distinguishes between positive and negative evidence in radiological reports. The system recognizes noun and conjunctive phrases that define negation boundaries. The proposed classifier had a precision value of 93%. [10]”Thomas et al. (2005) [10] developed a text search algorithm based on association rules and implemented a computerized text classification system. The fully computerized way that radiographic reports were put into categories of “normal,” “neither normal nor fracture,” and “fracture” was accurate. A rule-based system is a simple and effective AI-based application. However, the speed and ability to handle complex tasks are limited. On the other hand, ML-based classifiers can adjust their involved parameters to adapt to the ever-changing word usage in medical documents. In recent years, machine learning studies have begun to use complex statistical models to classify clinical texts. In decision-making, Bayesian Networks [11-16], Support Vector Machines [14,17-20], and Decision Trees [12,15,21-24] have been widely used. These models outperformed the rule-based system in terms of classification accuracy. De Bruijn et al. (2006) [14] used supervised machine learning approaches to develop classifiers that automatically detect acute wrist fractures in radiological reports in 2006 [14]. They reported that the support vector machine(SVM)-based text classifier performed best overall, with 94% accuracy. Guido Zuccon et al. (2013) [19] experimented with feature engineering in SNOMED CT concepts to improve medical image classification accuracy. The classifier developed by Guido Zuccon et al. could correctly identify fractures from radiological reports. It is also stated that when using bigram or SNOMED+bigram features, the Naïve Bayes classifier had the highest F1-score. Efsun et al. [53] created a classifier for bone fracture detection using regular text classification in 2017.

Topic modeling and document similarity measurement are used to train the classifier. The lack of transparency in the classifier remains an unresolved issue. As a result, physicians struggle to understand why the classifier makes positive or negative classifications. A recent study [54] used a large dataset to implement name entity recognition and bone fracture classification. There are some attempts at machine learning models' interpretation of clinical texts. For example, a recent study [55] built five machine learning algorithms to classify Alzheimer's drugs' mechanisms of action. The author visualized a decision tree and tried to provide some text interpretations. Obviously, more attempts are needed to fully reveal how machine learning models interpret the classification results. From these studies, the model's interpretability issues have not been fully resolved. To help people understand how decision-making systems work, it is important to build an interpretable model that is clear and easy to understand.

Interpretable Machine Learning

Methods based on machine learning are effective for classifying free-text reports. An ML model, as opposed to a rule-based system, consists of an algorithm that can learn latent patterns without hard-coding fixed rules [25]. One disadvantage of ML models is the difficulty in interpreting classification results [26]. To address this weakness, recent studies have begun to interpret ML models. This field of study is known as "interpretable machine learning" [26]. An ideal solution for interpretable machine learning is to provide the evidence and reasoning for the user. Furthermore, users can discover knowledge and justify predictions based on provided evidence [27]. Therefore, interpretable machine learning models increase user trust in classifiers.

Researchers have developed two types of model interpretation techniques: model-agnostic and model-specific approaches [28]. The model-agnostic approach explains the prediction of an ML model by approximating the output of the ML model's algorithms. Shapley Values, Independent Conditional Expectation Plots, Local Interpretable Model-agnostic Explanations, Permutation Feature Importance, and Partial Dependence Plot are a few examples [28] to explain a machine learning model. Model-specific explanation methods, on the other hand, excel at explaining complex models like tree ensemble models and artificial neural networks [27]. There is also open-source software available, such as SHAP [29], Eli5 [30], and InterpretML [31]. These tools can perform a variety of tasks, including image and text classification. Interpretable machine learning has recently been used in clinical practice for a variety of medical applications, such as predicting

mortality risk [32,33], predicting abnormal ECGs [34], and finding different symptoms from radiology reports that suggest limb fracture and wrist fracture [9,10,14,19].

Despite the importance of model interpretability in machine learning models, model interpretation developments are still at preliminary stages. There is a significant gap between physicians' desire to understand the prediction and the model's lack of interpretability. In response to this need, we investigated interpretable classification models in radiological texts in our study. The research was divided into two parts: First, we created a text classifier to classify text radiological reports automatically. Then we conducted model interpretations at the text level. We explored how keywords affect model classification results. To the best of our knowledge, this is the first study that interprets classification results based on temporal bone CT reports.

Significance of Study

We provided accessible model interpretations from a physicians' perspective. Model interpretability ensures physicians' trust in the model predictions. This study benefits physicians by visualizing how the computer makes a diagnosis. Also, the physicians' feedback is valuable to adjust the model's function in clinical practice.

Method

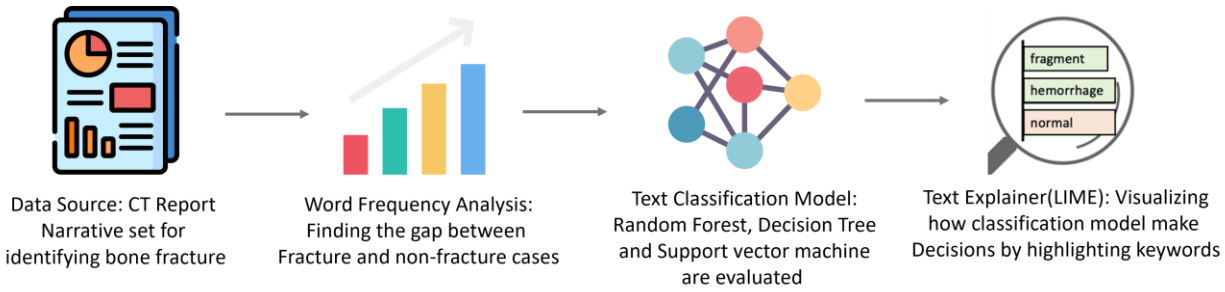


Figure 1: Overview of our study. We first used CT text reports to construct a text-based classification model. The steps are as follows: (1) text feature analysis; (2) Performance of Classification model; (3) visualizing interpretations.

Figure 1 is a graphical abstract of this study. A set of 164 clinical temporal bone CT reports were collected from the Clinical Research Data Warehouse (CRDW) of the Clinical & Translational Science Institute (CTSI) of Southeastern Wisconsin. We first created a vector representation of CT reports[35] and built text classification models. A follow-up classification performance was evaluated. To explain the machine learning model, we provided two types of model interpretation. The first type Is text feature analysis, which generates feature importance scores as well as word

frequency scores; the second type is a text explainer using LIME,[36] which provides a variety of interpretations of the classification results.

Data Source

An initial Honest Broker request was submitted to the Froedtert Health System i2b2 cohort query tool. This tool facilitates the integration of genomic and clinical data of healthcare institutions and is housed by the Clinical Research Data Warehouse (CRDW) of the Clinical & Translational Science Institute (CTSI) of Southeastern Wisconsin [37]. The diagnosis codes were selected to define a small patient cohort that would be feasible for the number of CT narratives included in the text analysis. Diagnosis codes were also selected to further define the patient cohort most likely to have a temporal bone fracture confirmed by a radiologist and therefore considered "clinically abnormal." Once this query was submitted, an identified accession list of CT exams was generated for the study team. The study team shared this accession list with the business analyst of biomedical informatics and requested a custom extraction of the imaging narratives and impressions from the data warehouse. The CT narratives and impressions were then de-identified for integration with the text analysis of this study. This query was further filtered to only include adults aged 60 – 65 to yield a final normal cohort of 119 patients, and temporal bone fracture cohort of 45 patients. Each patient's narrative was included in only one clinical text, resulting in a total of 164 documents in this study. All documents have been submitted to the supplemental files of this study.

Text Pre-processing

We removed all non-word elements in clinical reports, including numbers, punctuation, and special characters. We converted all words to lowercase letters. We perform these changes using regular expressions. Then, we followed a Natural Language Toolkit stop-word list to remove stop words, lemmatized words, and incorrect spellings and acronyms. All words were free of noun declination and verb conjugations. The supplemental code book shows how we pre-process the documents.

Text Feature Analysis – Word Frequency Score

To better understand the word distribution between positive and negative reports, we calculated the average Word Frequency Score (WFS) for each keyword**.** WFS is calculated by dividing the total number of reports by the total number of word frequencies. We generated WFS for each word

separately into positive sets and negative sets. Finally, we choose the words with the greatest frequency differences between positive and negative sets.

Machine Learning Model Development

To convert text reports to matrix formats in machine learning models, we used the bag-of-words model and the Term Frequency–Inverse Document Frequency(TF-IDF).[35,38] Bag-of-words and TF-IDF can convert each document to a fixed-length vector, allowing the ML model to process the text in vector form. In this representation, each distinct word was represented as a feature.

Word2vec is another popular topic-modeling technique to learn word associations in large texts. However, we believe that BOW and TF-IDF are better than Word2vec for text classification tasks. A bag of words is used to determine an article's topic, and the classification is determined by the type of words it contains. The Bone Fracture classification was evaluated using the TF-IDF metric, which measures word relevance. Because fracture descriptions are typically extreme, the TF-IDF can reflect this trend: some words always appear in a fracture report but almost never appear in a non-fracture report. Word2vec, on the other hand, is appropriate for discovering sub-topics. However, topic modeling is not the focus of this study. Based on these considerations, we believe the BOW and TF-IDF models are the best methods for this study.

We imposed a few additional restrictions to ensure proper conversion: First, we restricted the frequency of words that occur in all documents. The minimum frequency limit was five, and the maximum frequency was 70% of all documents. We evaluated performance when the model used a different number of top features as vectors, ranging from the two most frequent features to five hundred most frequent features, to determine the number of features used for model building. For the algorithm choice, we selected three algorithms, including XGBoost Support Vector Machine, Logistic Regression, and Random Forest. The three different algorithms represented various algorithmic and statistical interpretations. We used 20% (31) of clinical documents as our test set. which includes 9 fracture texts and 22 non-fracture texts. The remaining 80% (133) of texts were used for the training set. To train and test datset on a small and imbalanced dataset, we selected a stratified 10-fold cross-validation technique. Finally, we evaluated precision, recall, accuracy and F1-score metrics for all models. The final reported metrics are based on stratified 10-fold cross-validation using the top five hundred features as vectors in the feature sets.

Interpretation of Machine Learning Models

To interpret the classification mechanisms, we adopted the LIME[36] method on CT reports. LIME is a framework that explains the classifier predictions in an interpretable manner. First, a machine learning classifier classifies temporal bone fracture cases to bone fracture (positive) or non-bone fracture (negative) cases. The classification model learns the difference between positive and negative word distributions from clinical texts. Finally, LIME highlights the keywords in texts that help with prediction.

To view the LIME's evaluation accuracy, we performed an evaluation of the LIME's explanation results in two metrics. To achieve explanation, LIME creates an alternate white-box model based on text features. During this step, the vector representation of the black box model is compared to the vector representation of the white box model. We used two metrics30 to assess the performance of LIME's interpretation results, including accuracy score and Kullback-Leibler divergence.

(1) An accuracy score is calculated by dividing the generated sample by the cosine distance between the generated sample and the original documents (i.e., text which is closer to the example that is more important). A perfect match of text patterns between two models yields a score of one.

(2) The Kullback-Leibler divergence demonstrates how interpretable models approximate the machine learning model. A lower score indicates that two models tend to classify the same results. A value of zero indicates a perfect match across all classification results.

Results

Word Frequency Score and Clinical Text Summary

We first summarize the clinical documents. Among 164 selected text documents, forty-five were diagnosed with a bone fracture and 119 were diagnosed without a fracture. The positive CT reports have an average length of 299.8 words (Standard deviation (SD) = 124.3), significantly shorter than normal CT reports (average = 480.6, SD = 235.9). The top-five common words that appear in normal reports are 'normal' (total frequency = 487), 'right' (393), 'canal' (356), 'CT' (347), and 'left' (334). The top five common words that appear in fracture reports are 'left' (432), 'fracture' (394), 'right' (381), 'bone' (337) and 'temporal' (312). Figure 2 shows the normal-report-favored and fracture-report-favored word lists that show the largest word frequency score gaps between the two categories.

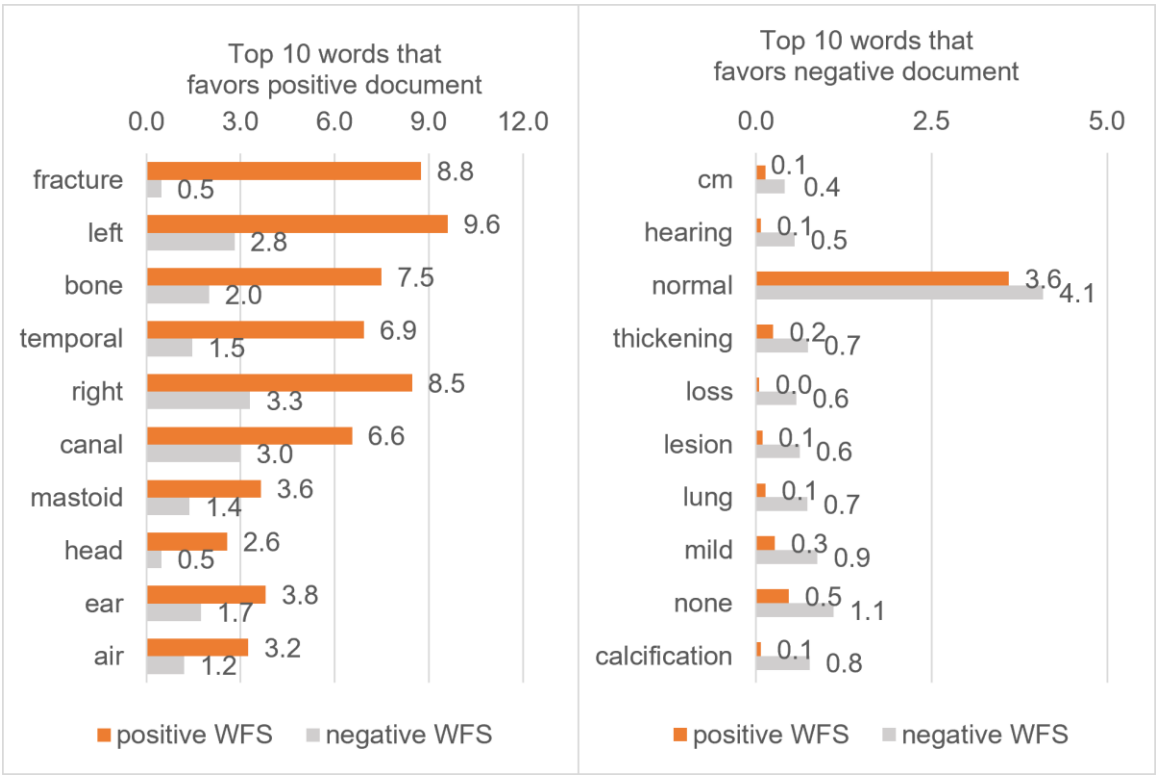


Figure 2: Comparison of gaps between fracture and non-fracture reports. The red bar stands for the frequency of fracture reports, and the blue bar stands for non-fracture reports. The left-side chart shows the top ten words that appear more often in fracture sets, whereas the right-side chart shows the top ten words that appear more often in non-fracture sets.

Classification Models' Parameters and Performance:

Figure 3 shows the relationship between classification model performance and the number of keywords used in models. Each sub-figure uses either random forest, SVM, or logistic regression algorithms. Figure 3 shows that there are positive correlations between the number of keywords and classification performance. As the number of topics increases in the Random Forest model, precision and accuracy remain high, but recall begins to decline. The SVM and Logistic Regression models did not exhibit a decreasing trend. Considering the relationship between the number of selected keywords and performance, we eventually adopted five hundred keywords into the feature set.

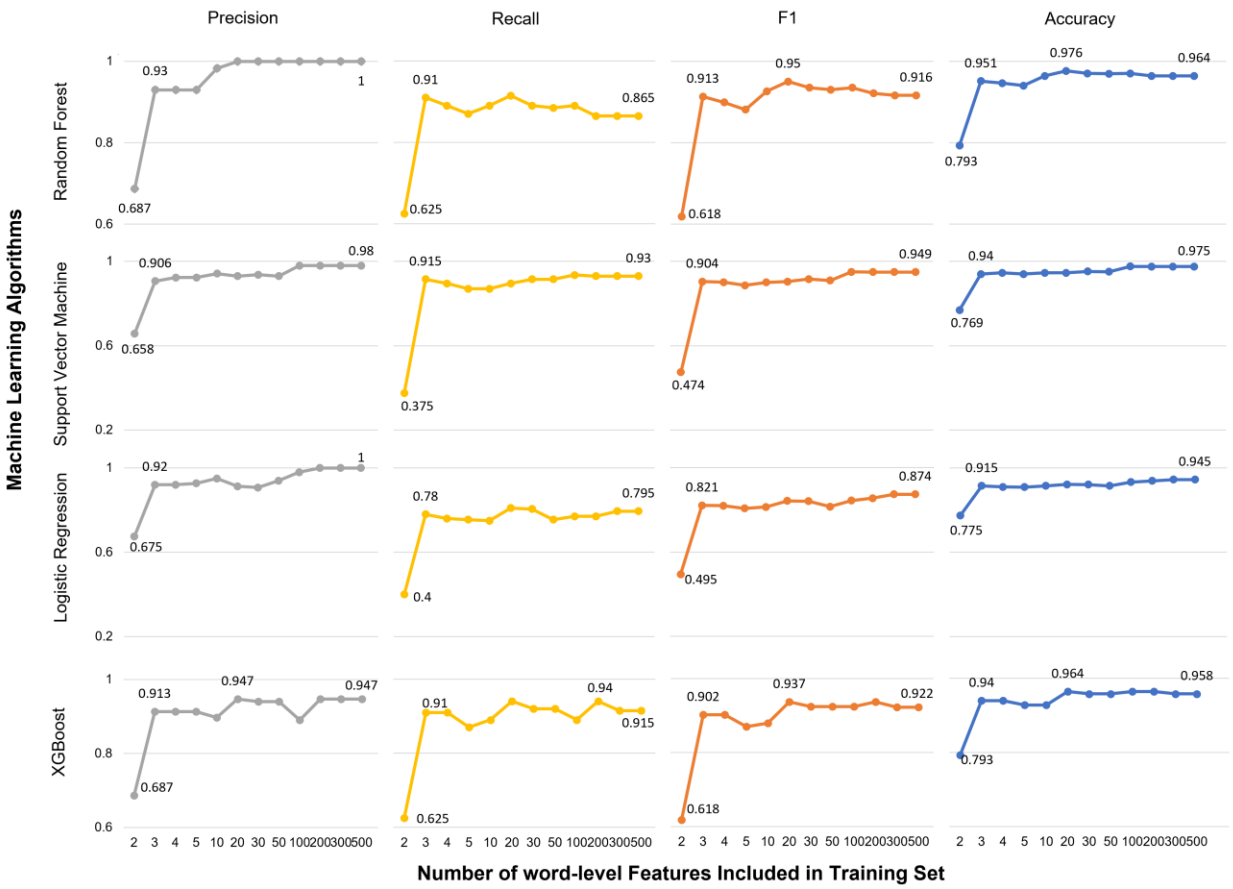


Figure 3: Relationships between classification model's performance, number of selected features, and evaluation performances for Random Forest, Support Vector Machine, and Logistic Regression model.

Interpretation of Machine Learning Models

A LIME Text Explainer visualizes text features that influence the classification positively or negatively. Figure 4 illustrates an example of a bone clinical text case. LIME used a Random Forest algorithm to provide explainable results in this document. Important keywords that contribute to the final classification result are highlighted in the visualization. A similar visualization was produced by the support vector machine algorithms. They did, however, produce a slightly different selection of keywords in the keyword feature sets. The Random Forest classifier predicts a fracture result with 79.5% certainty and a z-score of 1.354. The words in green were explained as having contributed to the model's positive classification result. The words 'comminuted,' 'fracture,' 'lodged,' 'fossa,' 'disruption,' 'ossicular,' and 'temporal' were ranked among the most predictive words for the positive classification result in the prediction result. As a follow-

up weight evaluation, figure 8 shows how the explainer considers the weight of each word that predicts the most positive outcome.

Figure 4 shows a visualization of a text explainer. In this text explainer, each word has been assigned a contribution score, showing the words lead to positive or negative classification. The text explainer's evaluation is based on the individual text level.

The feature list in Figure 4 displays the random forest model's most important word list. The word list is aggregated from multiple decision trees. The selection of each word is calculated by whether the word serves as a deciding factor in a decision tree. Higher weight words are often used as a key factor in the classification results. The feature list corresponds to the text explainer's assessment. Both show that LIME can successfully discover keywords for classification.

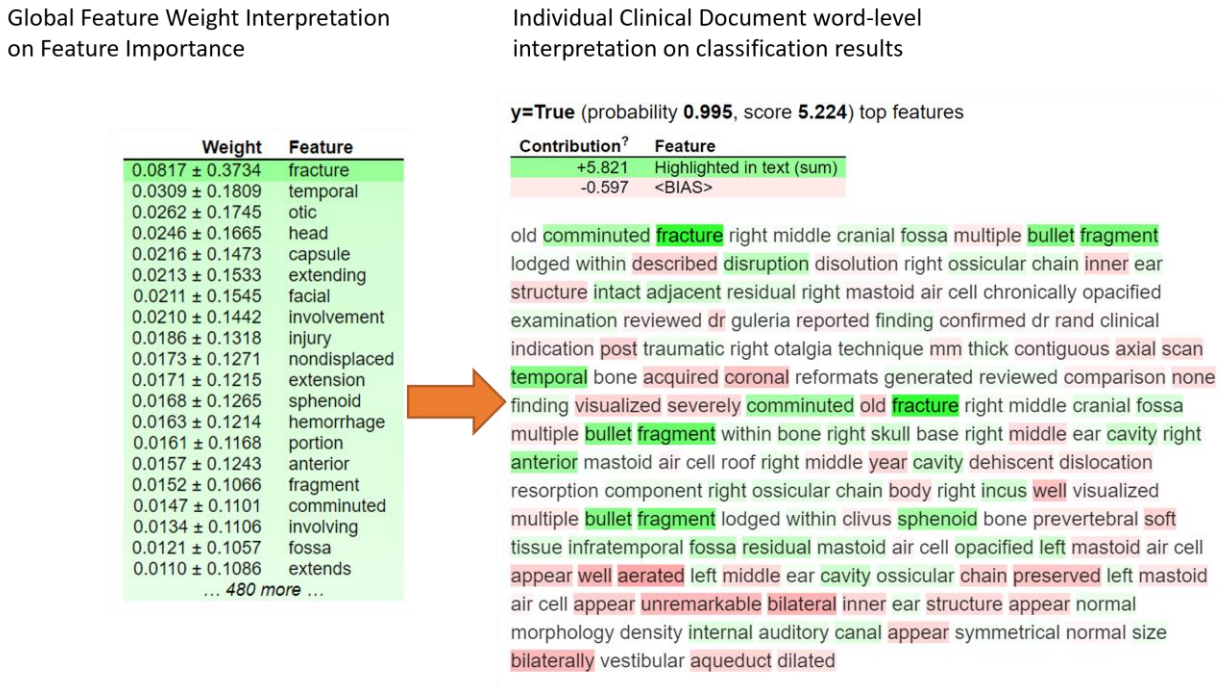


Figure 4: How LIME evaluate the importance of each word features and use the weight of features to visualize the word-level contribution for each document to calculate classification results

The reliability of LIME's interpretation was assessed by the accuracy score and the Kullback-Leibler divergence score between the LIME interpretation framework and the machine learning model. A subsequent evaluation of the accuracy score between our Random Forest model and the explainable model is 0.867. It means 86.7% of reports will generate the same prediction

result between two classifiers. A mean Kullback-Leibler divergence for all target classes showed how well probabilities are approximated between two models. The Kullback-Leibler divergence value is 0.015. This means two models will have a 1.5% probability that they will predict the same report into different categories. With these evaluations, we can state that the Text Explainer model is a highly trustworthy model that can predict the behavior of support vector machine models in CT classification tasks.

Discussion

Study Significance

Clinical text reports are one of the most important, yet underutilized, resources in electronic health records. [9] An interpretable model not only assists medical practitioners in making informed decisions, but it also increases physicians' trust in the model. This trust can expedite the adoption of computer-based diagnosis [7]. In this study, we used untemplated CT narratives from electronic health records to train ML classifiers to classify bone fracture cases. We outperformed a similar study using a crowdsourcing method, which achieved an accuracy of 0.799 [42]. We presented word-level contributions to prediction using the LIME framework. Clinicians can use the word-level list to help them validate the model's validity. The model can aid in clinical decision-making by providing understandable explanations.

A LIME-based model explains how word-level contributions in clinical texts are retrieved. The words "fragment" and "hemorrhage" are shown in green in the graphical abstract of Figure 1, indicating their contribution to bone fracture prediction. Previous experience can lead physicians to automatically associate the terms "fragment" and "hemorrhage" with bone fracture diagnosis. Physicians' domain knowledge frequently corresponds to the model's explanation. As a result, physicians can accept a model's prediction if they understand how the model's algorithm interprets documents and makes predictions. We believe that providing such interpretations will increase clinical acceptance of automated systems. The interpretation of a model that represents the algorithm increases physicians' trust, resulting in transparency. The transparency facilitates the transition from manual to automated processes.

Summary of Text Feature Analysis

We discovered that text features contribute to the prediction of results. The word usage demonstrates a significant difference between positive and negative reports. For example, the word 'fracture' was identified as the most important feature in our evaluation; it appears frequently in the positive set (frequency = 394) but infrequently in the negative data set (55). "head" (frequency = 116 in the positive set, 57 in the negative set), "temporal" (312 in the positive set, 173 in the negative set), and "hemorrhage" are other examples (87 in the positive set, 17 in the negative set). All these words demonstrate the frequency difference between positive and negative sets. We conclude that the WFS gap between fracture and non-fracture reports can measure how the words are used differently.

In Figure 2, the WFS result indicates that "fracture," "left," "bone," "temporal," and "right" are the top five words that appear more frequently in positive sets than in negative sets. The top five words that appear more frequently in negative sets than in positive sets are "calcification," "none," "mild," "lung," and "lesion." The difference in WFS between fracture and non-fracture reports is a predictor of classification. Physicians often draft reports with highly specialized medical terms. These medical terms often serve as reliable predictors. According to our results, we suggest investigating if non-experts can easily interpret the medical terms.

In other related studies, similar clinical text features were also examined. The goal is to investigate radiologists' preferences for specific words in clinical documents. A previous study [11], for example, used Naive Bayes-based predictive machine learning models. Language patterns in clinical documents are typically consistent across specialties. The study [11] discovered, for example, that otolaryngologists use distinct language patterns in vestibular notes that are highly conserved. These patterns are highly predictive of specific vestibular diagnoses. Using a medical specialized corpus makes it easy for doctors to understand how language patterns work.

We believe that similar language patterns exist in other medical departments. According to the WFS gaps in Figure 2, we classified words as fracture-favored or non-fracture-favored. The classification yields two distinct word lists. Documents with a bone fracture prefer one list of words, while documents without a fracture prefer the other. As a result, our WFS analysis identified text patterns associated with classification results. Incorporating Electronic Health Records into decision-making models has been used to treat a variety of diagnoses and conditions, including

heart failure symptoms [39], vestibular diagnoses [40], and gastrointestinal diagnoses [41]. Our results on LIME show specialized medical words make significant contributions to the classification. As a result, we believe that interpretable AI has the potential to help explain more complex diagnostic conditions. This pipeline can also be used to interpret other clinical texts, classify diagnoses, and give an explanation that is easy to understand.

Summary of Classification Performances

We demonstrated that a model using five hundred major topic words and stratified 10-fold cross validation can achieve an average performance of 0.95 accuracy on Random Forest classifiers. A comparable study of medical text labeling using crowdsourcing methods achieves 0.799 accuracy[42]. Therefore, our model's performance is competitive compared to other human-labeled studies. We observed our model's precision is high, but recall is lower. It indicates our model tends to predict more positive outcomes than negative outcomes. As a result, we developed a cautious model that may help to avoid serious errors in clinical practice. As a result, our model may help to result in fewer false negative cases and may avoid serious errors in clinical practice.

In four models, we discovered a positive correlation between the number of keywords used in the feature set and model performance (Figure 3). Increasing the number of features can improve prediction performance, especially when the number of keywords is less than ten. When the number of keywords exceeds two hundred, however, performance remains stable. We hypothesize that the performance is hampered by the limited impact of less-important keywords. Because those keywords have lower WFS, their contribution to classification is minimal. A growing number of adopted features in other algorithms, such as Support Vector Machine algorithms, will provide more information to the machine learning model. SVM algorithms perform well in high-dimensional spaces; they can use word and text features to provide accurate classification. To summarize, both three algorithms can complete classifications at high performance.

Summary of Interpretation Results

LIME has been widely used in research studies. Pan et al[6] used LIME to investigate the contribution level of features of new instances for predicting central precocious puberty in girls. Ghafouri-Fard et al[44] used the same approach for diagnosing autism spectrum disorder. Palatnik de Sousa et al used LIME to classify the metastases of lymph nodes[45]. Other interpretation of target

conditions using LIME include acute kidney injury[24], chest injury[46], electrocardiogram-aided cardiovascular diseases[34], radiology reports[9], and so on. Overall, LIME can provide visualized results for various diagnoses to help clinicians evaluate the reliability of clinical-decision model.

There are two approaches to implementing classification model interpretation: using an intrinsically transparent model such as decision tree23. Another approach to achieving interpretability is to use post-hoc methods such as LIME. This method was chosen primarily for the ease of interpretation of results at the word level. For ML-based explanation methods to be chosen in the future, it is important to talk about the pros and cons of each implementation of interpretation.

Interpretability of Machine Learning Models

Some machine models are inherently transparent and interpretable. Because of their simple mechanisms, the output of these models is directly interpretable. For example, linear regression and logistic regression are inherently more transparent. Clinicians have extensive experience in interpreting coefficients, effect sizes, and p-values. For example, our previous studies[48] explored the social determinants of tertiary rhinology care utilization using linear regression techniques, which require no AI-based knowledge.

A decision tree[12] is a slightly complicated yet transparent model in computer science.[12]. In machine learning, this concept can be used to define a preferred sequence of attributes to investigate to most rapidly narrow down to a specific state. Such a sequence is called a decision tree. This process is known as decision tree learning. Usually, an attribute with high mutual information should be preferred to other attributes. A higher information gain can be applied to the split for each node. As we stated, we can calculate the information gain for specific words. When a decision tree is used for text classification, it consists of internal tree nodes labeled by term, branches departing from them are labeled by test on the weight, and leaf nodes represent corresponding class labels. By going through the query structure from root to leaf, which is the goal of classification, decision trees can classify the document based on predetermined rules.

Figure 5 shows a simplified decision tree showing how to make a diagnosis decision based on the frequency of specific words in CT reports. The decision tree is generated from the example document. The square indicates each criterion when a document is working as an input. The number in each square represents the frequency of words occurring in the document. Each

condition criterion is followed by a percentage number. The percentage number means the percentage of all documents that fall into this condition. Each document will be classified and go into one of the categories, finally being classified into a "positive" or "negative" result. The decision rules are learned by machine learning techniques by information gain, where a bit of statistical knowledge is required to learn to understand how to build the decision tree. Fortunately, the decision rules and the sequence are directly interpretable by clinicians, especially if a tree is small. There is a significant difference between decision tree and LIME methods in the complexity of interpretation. A decision tree requires clinicians to analyze whether an entire sequence of a tree is reasonable, whereas a LIME method only requires clinicians to determine if featured words are associated with target outcomes. However, a decision tree can be too large to interpret. It is also possible a decision tree may not generate a meaningful representation. Therefore, the LIME method is significantly easier to interpret a model.

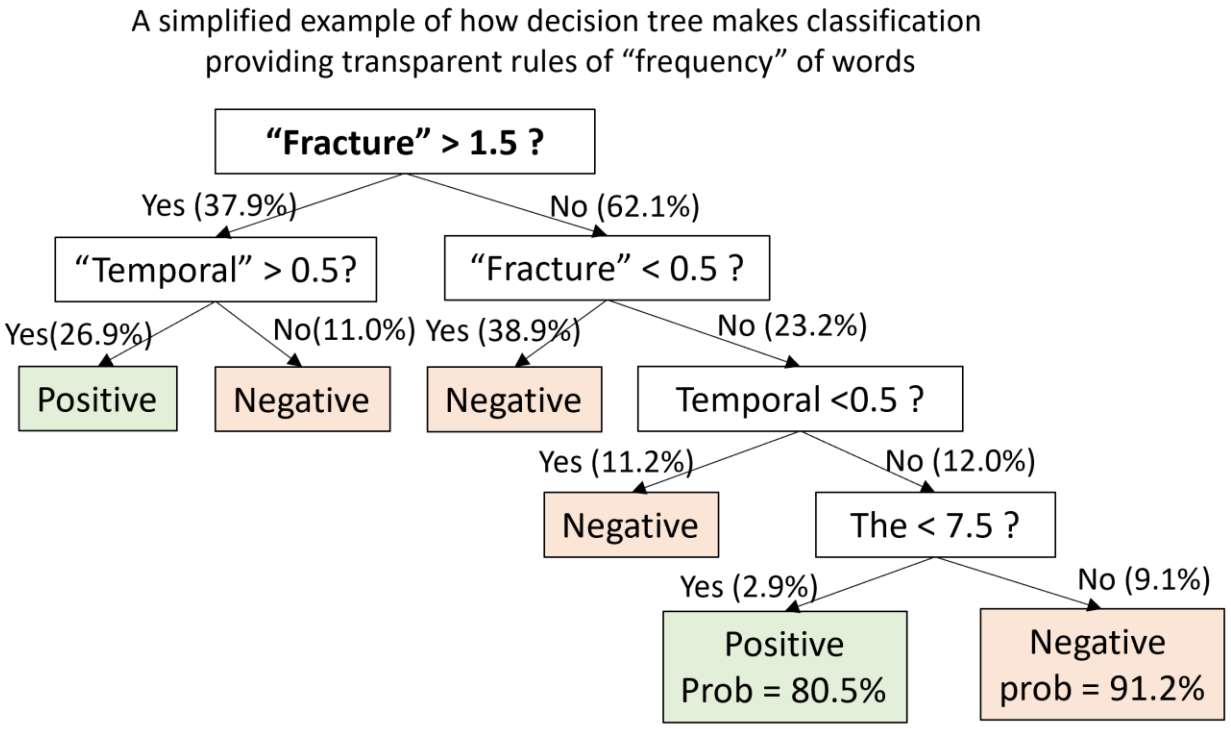


Figure 5: A visualization of how a transparent decision tree model determines the classification result from a sample CT text report. The percentage number shows the proportion of reports falling into each category. The frequency of a specific word determines the model's classification result.

Limitations and future work

An obvious limitation is the limited variety and quantity of temporal bone CT reports, with only 164 documents available. All reports were limited to a single health care system in Wisconsin, which may introduce potential bias. A larger set of clinical reports may also lead to unbiased model construction and more accurate classification performance. Our future work will consist of two aspects: First, we used labeled data in this preliminary study. While unlabeled data cannot be used for classification, it has the potential for unsupervised learning. We believe that by building an appropriate unsupervised model, it is possible to cluster CT reports into two categories based on text reports. Second, building a medically specialized text interpreter would highlight medical words only and achieve a more transparent interpretation. For example, by adopting SNOMED-CT standards [52], we can create a medical text interpreter. By using medical terms only, the model could narrow down the choices of words. The word-level optimization may achieve better prediction and better interpretation.

Conclusion

Machine learning models are often incomprehensible for clinical providers. There is a need for interpreting clinical text in simple ways. To interpret models, we used two major methodologies: (1) We calculated an average word frequency score for keywords. (2) Using Local Interpretable Model-Agnostic Explanations, we visualized the contribution weight of keywords to bone fracture. We concluded that interpretable text explainers can improve physicians' understanding of machine learning predictions. By providing simple visualization, our model can increase the trust of computerized models and support computerized decision-making in clinical settings.

Understanding the decision-making system's mechanism is critical for promoting its use in clinical settings. By providing physicians with simple visualization, our model can help them make decisions. This interpretation increases confidence in computerized models. Our proposed method could be used as a computer-assisted tool in CT report classification. It could be used as an adjunct tool to assist clinicians in making decisions in their daily practice. Overall, this study laid the groundwork for the development and validation of credible explanations. Our model has the potential to be integrated into contemporary clinical decision-making environments for clinical practitioners.

Acknowledgement

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We confirm that the manuscript has been read and approved by all named authors. No other persons who satisfied the criteria for authorship are listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

The project described was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, Award Number UL1TR001436. The content is solely the responsibility of the author(s) and does not necessarily represent the official views of the NIH.

Ethics Approval

This is an observational study. The University of Wisconsin—Milwaukee Institutional Research Ethics Committee has confirmed that no ethical approval is required.

Competing Interest

The authors declare that they have no competing interests. The sponsor was not involved in the design and conduct of the study; the collection management, analysis, and interpretation of the data; the preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication.

References

1. Shortliffe EH, Cimino JJ. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. 4th ed. Springer; 2014.
2. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet 2012 136*. 2012;13(6):395-405. doi:10.1038/nrg3208
3. Greenes R. *Clinical Decision Support: The Road to Broad Adoption*. Academic Press. Academic Press; 2014. Accessed July 15, 2021.
4. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Informatics Assoc*. 2019;26(4):364-379. doi:10.1093/JAMIA/OCY173
5. Shao Y, Taylor S, Marshall N, Morioka C, Zeng-Treitler Q. Clinical Text Classification with Word Embedding Features vs. Bag-of-Words Features. *Proc - 2018 IEEE Int Conf Big Data, Big Data 2018*. Published online January 22, 2019:2874-2878. doi:10.1109/BIGDATA.2018.8622345

6. Liyan, Liu G, Mao X, et al. Development of Prediction Models Using Machine Learning Algorithms for Girls with Suspected Central Precocious Puberty: Retrospective Study. *JMIR Med Inf 2019;7(1)e11728 https//medinform.jmir.org/2019/1/e11728*. 2019;7(1):e11728. doi:10.2196/11728

7. da Cruz HF, Pfahringer B, Martensen T, et al. Using interpretability approaches to update "black-box" clinical prediction models: an external validation study in nephrology. *Artif Intell Med*. 2021;111:101982. doi:10.1016/J.ARTMED.2020.101982

8. Mujtaba G, Shuib L, Idris N, et al. Clinical text classification research trends: Systematic literature review and open issues. *Expert Syst Appl*. 2019;116:494-520. doi:10.1016/J.ESWA.2018.09.034

9. Aronow DB, Fangfang F, Croft WB. Ad Hoc Classification of Radiology Reports. *J Am Med Informatics Assoc*. 1999;6(5):393-411. doi:10.1136/JAMIA.1999.0060393

10. Thomas BJ, Ouellette H, Halpern EF, Rosenthal DI. Automated Computer-Assisted Categorization of Radiology Reports. *Am J or Roentgenol*. 2005;184(2):687-690. doi:10.2214/AJR.184.2.01840687

11. Luo J, Erbe C, Friedland DR. Unique clinical language patterns among expert vestibular providers can predict vestibular diagnoses. *Otol Neurotol*. 2018;39(9):1163-1171. doi:10.1097/MAO.0000000000001930

12. Lewis DD. A Comparison of Two Learning Algorithms for Text Categorization 1 Introduction 2 Text Categorization : Nature and Approaches. *Proceeding Third Annu Symp Doc Anal Inf Retr*. 1994;33:1-14.

13. Raja Srinivasa Reddy B, Kadaru BB. An integrated hybrid feature selection based ensemble learning model for parkinson and alzheimer's disease prediction. *Int J Appl Eng Res*. 2017;12(22):11989-12003. Accessed April 17, 2020. http://www.ripublication.com

14. de Bruijn B, Cranney A, O'Donnell S, Martin JD, Forster AJ. Identifying Wrist Fracture Patients with High Accuracy by Automatic Categorization of X-ray Reports. *J Am Med Informatics Assoc*. 2006;13(6):696-698. doi:10.1197/JAMIA.M1995

15. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. *Assoc Adv Artif Intell*. 1998;752(1):41-48. Accessed July 15, 2021. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf

16. Schneider KM. Techniques for improving the performance of naive bayes for text classification. In: *Lecture Notes in Computer Science*. Vol 3406. ; 2005:682-693. doi:10.1007/978-3-540-30586-6_76

17. Wang Y, Sohn S, Liu S, et al. A clinical text classification paradigm using weak supervision and deep representation 08 Information and Computing Sciences 0801 Artificial Intelligence and Image Processing 17 Psychology and Cognitive Sciences 1702 Cognitive Sciences. *BMC Med Inform Decis Mak*. 2019;19(1):1-13. doi:10.1186/s12911-018-0723-6

18. Qin YP, Wang XK. Study on multi-label text classification based on SVM. *6th Int Conf Fuzzy Syst Knowl Discov FSKD 2009*. 2009;1:300-304. doi:10.1109/FSKD.2009.207

19. Zuccon G, Wagholikar AS, Nguyen AN, et al. Automatic Classification of Free-Text Radiology Reports to Identify Limb Fractures using Machine Learning and the SNOMED CT Ontology. *AMIA Summits Transl Sci Proc*. 2013;2013:300. Accessed July 15, 2021. /pmc/articles/PMC3845773/

20. Joachims T. Text categorization with Support Vector Machines: Learning with many relevant features. *Eur Conf Mach Learn*. Published online 1998:137-142. doi:10.1007/BFB0026683

21. Chaurasia V, Pal S. Data Mining Approach to Detect Heart Diseases. *Int J Adv Comput Sci Inf Technol*. 2014;2(4):56-66. Accessed April 17, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376653

22. Vateekul P, Kubat M. Fast induction of multiple decision trees in text categorization from large scale, imbalanced, and multi-label data. *ICDM Work 2009 - IEEE Int Conf Data Min*. Published online 2009:320-325. doi:10.1109/ICDMW.2009.94

23. Johnson DE, Oles FJ, Zhang T, Goetz T. A decision-tree-based symbolic rule induction system for text categorization. *IBM Syst J*. 2002;41(3):428-437. doi:10.1147/SJ.413.0428

24. Freitas Da Cruz H, Schneider F, Schapranow M-P. Prediction of Acute Kidney Injury in Cardiac Surgery Patients: Interpretation using Local Interpretable Model-agnostic Explanations. *HEALTHINF*. Published online 2019:380-387. doi:10.5220/0007399203800387

25. Dipanjan Sarkar. The Importance of Human Interpretable Machine Learning. Towards Data Science. Published May 24, 2018. Accessed July 15, 2021. https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476

26. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy 2021, Vol 23, Page 18*. 2020;23(1):18. doi:10.3390/E23010018

27. Molnar C, Casalicchio G, Bischl B. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *Commun Comput Inf Sci*. 2020;1323:417-431. doi:10.1007/978-3-030-65965-3_28

28. Molnar C. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. Leanupb; 2019. Accessed July 15, 2021. http://leanpub.com/interpretable-machine-learning

29. Scott Lundberg. SHAP documentation. Published 2018. Accessed July 15, 2021. https://shap.readthedocs.io/en/latest/index.html

30. Mikahail Korobov, Konstantin Lopuhin. ELI5 documentation. Published 2017. Accessed July 15, 2021. https://eli5.readthedocs.io/en/latest/index.html

31. InterpretML Team. InterpretML documentation. Published 2021. Accessed July 15, 2021. https://interpret.ml/docs/intro.html

32. Allyn J, Allou N, Augustin P, et al. A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: A decision curve analysis. *PLoS One*. 2017;12(1). doi:10.1371/journal.pone.0169772

33. Cerna AEU, Pattichis M, VanMaanen DP, et al. Interpretable Neural Networks for Predicting Mortality Risk using Multi-modal Electronic Health Records. *Arxiv*. Published online January 23, 2019. Accessed July 15, 2021. https://eugdpr.org/

34. Neves I, Folgado D, Santos S, et al. Interpretable heartbeat classification using local model-agnostic explanations on ECGs. *Comput Biol Med*. 2021;133:104393. doi:10.1016/J.COMPBIOMED.2021.104393

35. Singh AK, Shashi M. Vectorization of Text Documents for Identifying Unifiable News Articles. *IJACSA) Int J Adv Comput Sci Appl*. 2019;10(7). Accessed July 15, 2021. www.ijacsa.thesai.org

36. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*. Published online August 13, 2016:1135-1144. doi:10.1145/2939672

37. Clinical Research Data Warehouse (CRDW). Accessed July 15, 2021. https://ctsi.mcw.edu/investigator/ctsi-tools/i2b2/

38. Church K, Gale W. Inverse Document Frequency (IDF): A Measure of Deviations from Poisson. In: Springer, Dordrecht; 1999:283-295. doi:10.1007/978-94-017-2390-9_18

39. Vijayakrishnan R, Steinhubl SR, Ng K, et al. Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record. *J Card Fail*. 2014;20(7):459-464. doi:10.1016/J.CARDFAIL.2014.03.008

40. Friedland DR, Tarima S, Erbe C, Miles A. Development of a Statistical Model for the Prediction of Common Vestibular Diagnoses. *JAMA Otolaryngol Neck Surg*. 2016;142(4):351-356. doi:10.1001/JAMAOTO.2015.3663

41. Mehrotra A, Dellon ES, Schoen RE, et al. Applying a natural language processing tool to electronic health records to assess performance on colonoscopy quality measures. *Gastrointest Endosc*. 2012;75(6):1233-1239.e14. doi:10.1016/J.GIE.2012.01.045

42. Costa J, Silva C, Antunes M, Ribeiro B. On using crowdsourcing and active learning to improve classification performance. *Int Conf Intell Syst Des Appl ISDA*. Published online 2011:469-474. doi:10.1109/ISDA.2011.6121700

43. Craven MW, Shavlik JW. Extracting Thee-Structured Representations of Trained Networks. *Adv Neural Inf Process Syst*. 1995;8:24-30.

44. Ghafouri-Fard S, Taheri M, Omrani MD, Daaee A, Mohammad-Rahimi H, Kazazi H. Application of Single-Nucleotide Polymorphisms in the Diagnosis of Autism Spectrum Disorders: A Preliminary Study with Artificial Neural Networks. *J Mol Neurosci 2019 684*. 2019;68(4):515-521. doi:10.1007/S12031-019-01311-1

45. Sousa IP de, Vellasco MMBR, Silva EC da. Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases. *Sensors (Basel)*. 2019;19(13). doi:10.3390/S19132969

46. Kulshrestha S, Dligach D, Joyce C, et al. Comparison and interpretability of machine learning models to predict severity of chest injury. *JAMIA Open*. 2021;4(1):1-8. doi:10.1093/JAMIAOPEN/OOAB015

47. Wang Y, Sohn S, Liu S, et al. A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Informatics Decis Mak 2019 191*. 2019;19(1):1-13. doi:10.1186/S12911-018-0723-6

48. Poetker DM, Friedland DR, Adams JA, Tong L, Osinski K, Luo J. Socioeconomic Determinants of Tertiary Rhinology Care Utilization: *OTO open*. 2021;5(2). doi:10.1177/2473974X211009830

49. Sethi T, Kalia A, Sharma A, Nagori A. Interpretable artificial intelligence: Closing the adoption gap in healthcare. *Artif Intell Precis Heal*. Published online January 1, 2020:3-29. doi:10.1016/B978-0-12-817133-2.00001-X

50. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Informatics Assoc*. 2020;27(7):1173-1185. doi:10.1093/jamia/ocaa053

51. Fan A, Jernite Y, Perez E, Grangier D, Weston J, Auli M. ELI5: Long Form Question Answering. *ACL 2019 - 57th Annu Meet Assoc Comput Linguist Proc Conf*. Published online July 22, 2019:3558-3567. Accessed July 15, 2021. https://arxiv.org/abs/1907.09190v1

52. Amgad M, Elfandy H, Hussein H, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*. Published online 2019. doi:10.1093/bioinformatics/btz083

53. Dai, Z., Li, Z., & Han, L. (2021). BoneBert: A BERT-based Automated Information Extraction System of Radiology Reports for Bone Fracture Detection and Diagnosis. In IDA (pp. 263-274).

54. Kayi, E. S., Yadav, K., Chamberlain, J. M., & Choi, H. A. (2017). Topic Modeling for Classification of Clinical Reports. arXiv preprint arXiv:1706.06177.

Figure 1

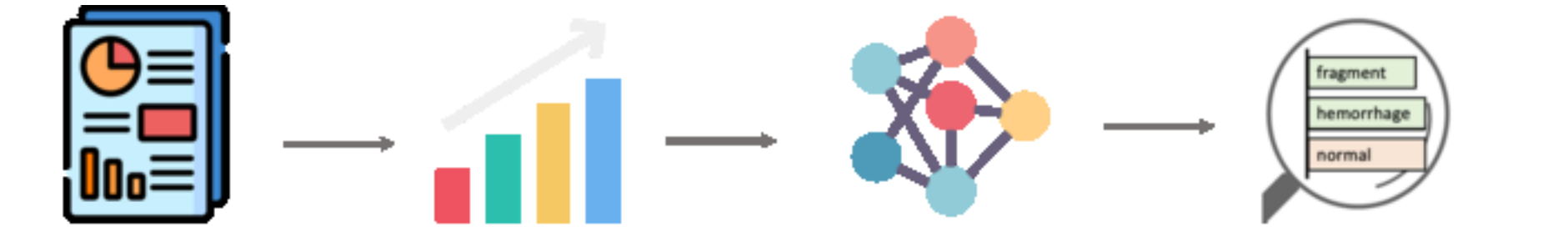


Figure 2

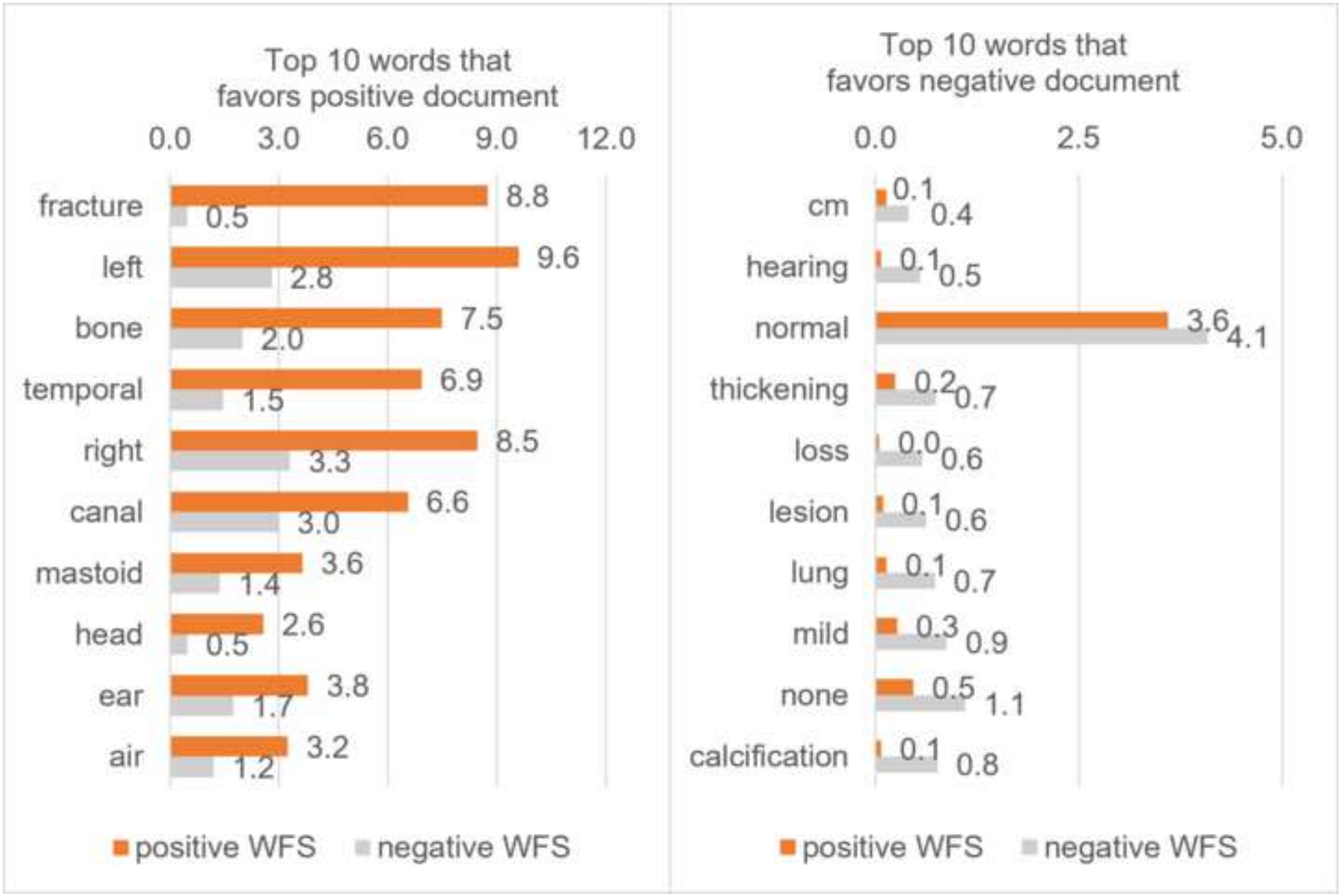


Figure 3

Click here to access/download;Figure;Figure-3-main-text-performance-and-number-of-keywords.tif

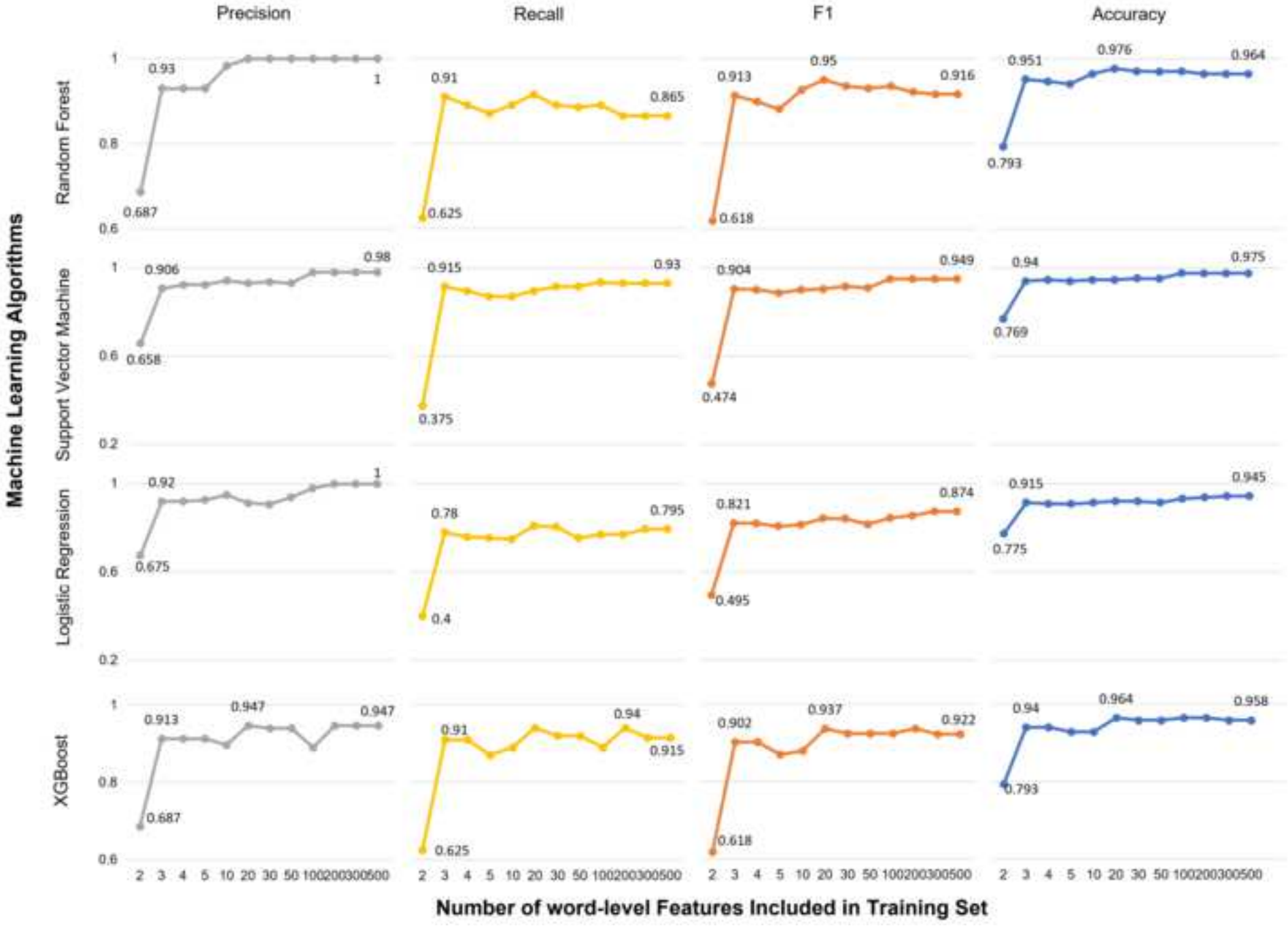


Figure 4

Global Feature Weight Interpretation on Feature Importance

Weight	Feature
0.0817 ± 0.3734	fracture
0.0309 ± 0.1809	temporal
0.0262 ± 0.1745	otic
0.0246 ± 0.1665	head
0.0216 ± 0.1473	capsule
0.0213 ± 0.1533	extending
0.0211 ± 0.1545	facial
0.0210 ± 0.1442	involvement
0.0186 ± 0.1318	injury
0.0173 ± 0.1271	nondisplaced
0.0171 ± 0.1215	extension
0.0168 ± 0.1265	sphenoid
0.0163 ± 0.1214	hemorrhage
0.0161 ± 0.1168	portion
0.0157 ± 0.1243	anterior
0.0152 ± 0.1066	fragment
0.0147 ± 0.1101	comminuted
0.0134 ± 0.1106	involving
0.0121 ± 0.1057	fossa
0.0110 ± 0.1086	extends
… *480 more* …	

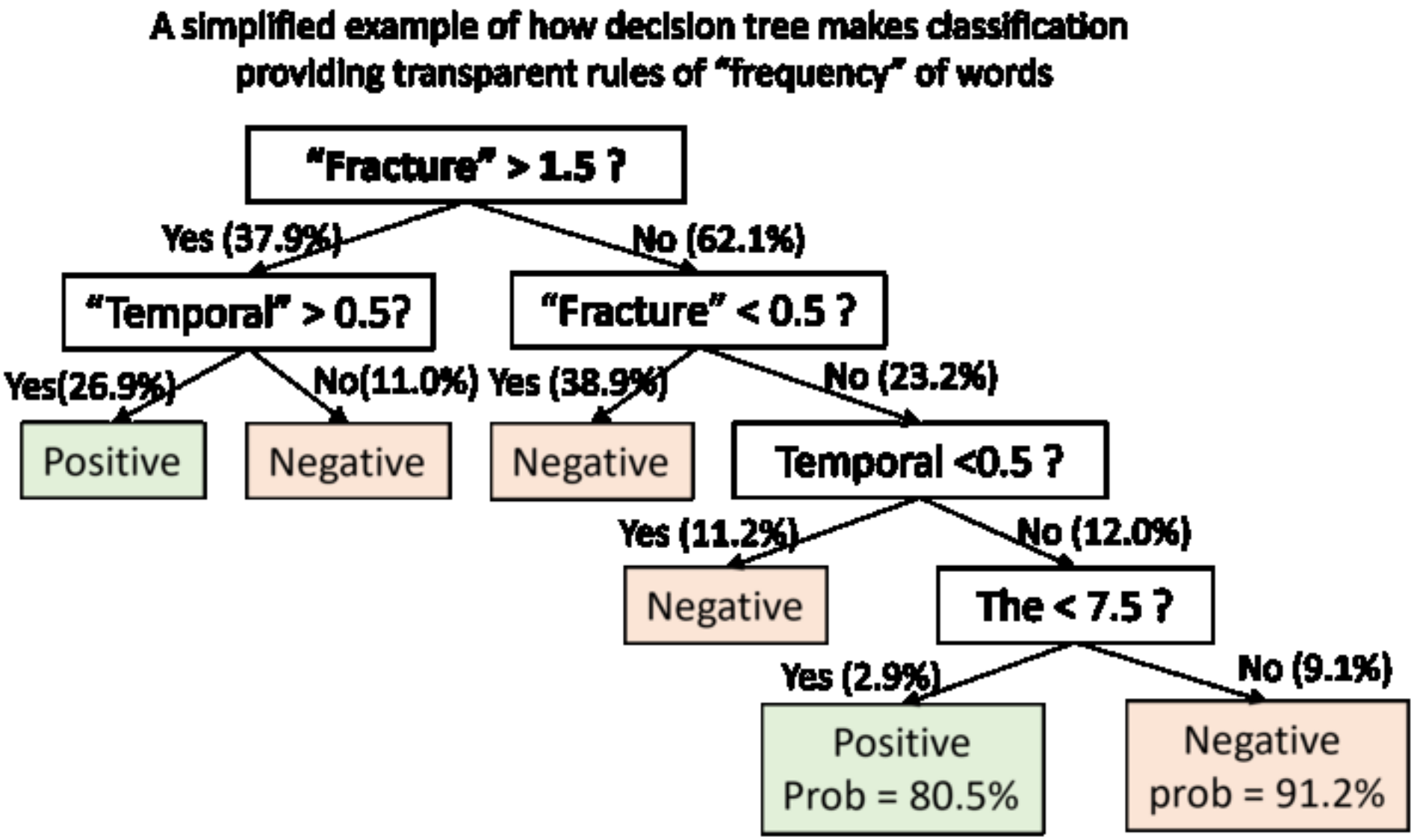
Individual Clinical Document word-level interpretation on classification results

y=True (probability **0.995**, score **5.224**) top features

Contribution[?]	Feature
+5.821	Highlighted in text (sum)
-0.597	<BIAS>

old comminuted fracture right middle cranial fossa multiple bullet fragment lodged within described disruption disolution right ossicular chain inner ear structure intact adjacent residual right mastoid air cell chronically opacified examination reviewed dr guleria reported finding confirmed dr rand clinical indication post traumatic right otalgia technique mm thick contiguous axial scan temporal bone acquired coronal reformats generated reviewed comparison none finding visualized severely comminuted old fracture right middle cranial fossa multiple bullet fragment within bone right skull base right middle ear cavity right anterior mastoid air cell roof right middle year cavity dehiscent dislocation resorption component right ossicular chain body right incus well visualized multiple bullet fragment lodged within clivus sphenoid bone prevertebral soft tissue infratemporal fossa residual mastoid air cell opacified left mastoid air cell appear well aerated left middle ear cavity ossicular chain preserved left mastoid air cell appear unremarkable bilateral inner ear structure appear normal morphology density internal auditory canal appear symmetrical normal size bilaterally vestibular aqueduct dilated

Figure 5



Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Authors:
Tong, Ling; Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, USA, ltong@uwm.edu
Luo, Jake; Department of Health Informatics and Administration, University of Wisconsin-Milwaukee, Milwaukee, USA, jakeluo@uwm.edu
Jazzmyne Adams, Department of Otolaryngology and Communication Sciences, Medical College of Wisconsin, Milwaukee, Wisconsin, USA, jaadams@mcw.edu
Osinski, Kristen; Medical College of Wisconsin, Clinical and Translational Science Institute of Southeastern Wisconsin, kosinski@mcw.edu
Xiaoyu Liu, Department of Electrical Engineering and Computer Science, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA, liu267@uwm.edu
David Friedland, Department of Otolaryngology and Communication Sciences, Medical College of Wisconsin, Milwaukee, Wisconsin, USA, dfriedland@mcw.edu

Click here to access/download

Supplementary Material

Supp-A-deidentified-data-source.xlsx