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

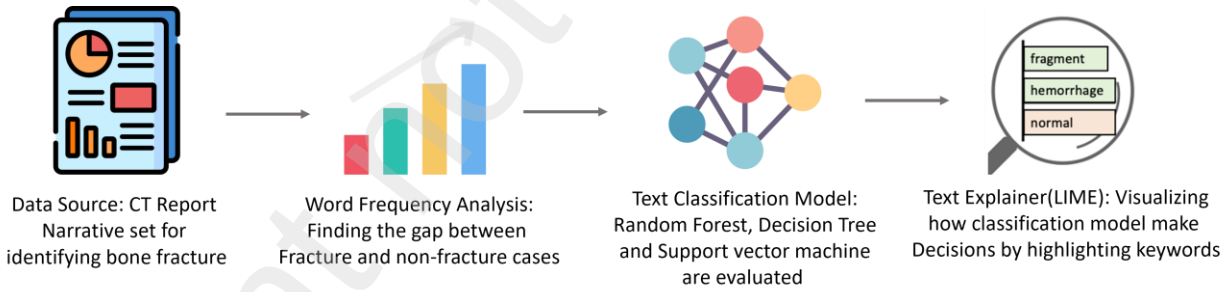

Figure 1: Overview of our study. We first used CT text reports to construct a text-based classification model. The steps are as follows: (1) text feature analysis; (2) Performance of Classification model; (3) visualizing interpretations.

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.

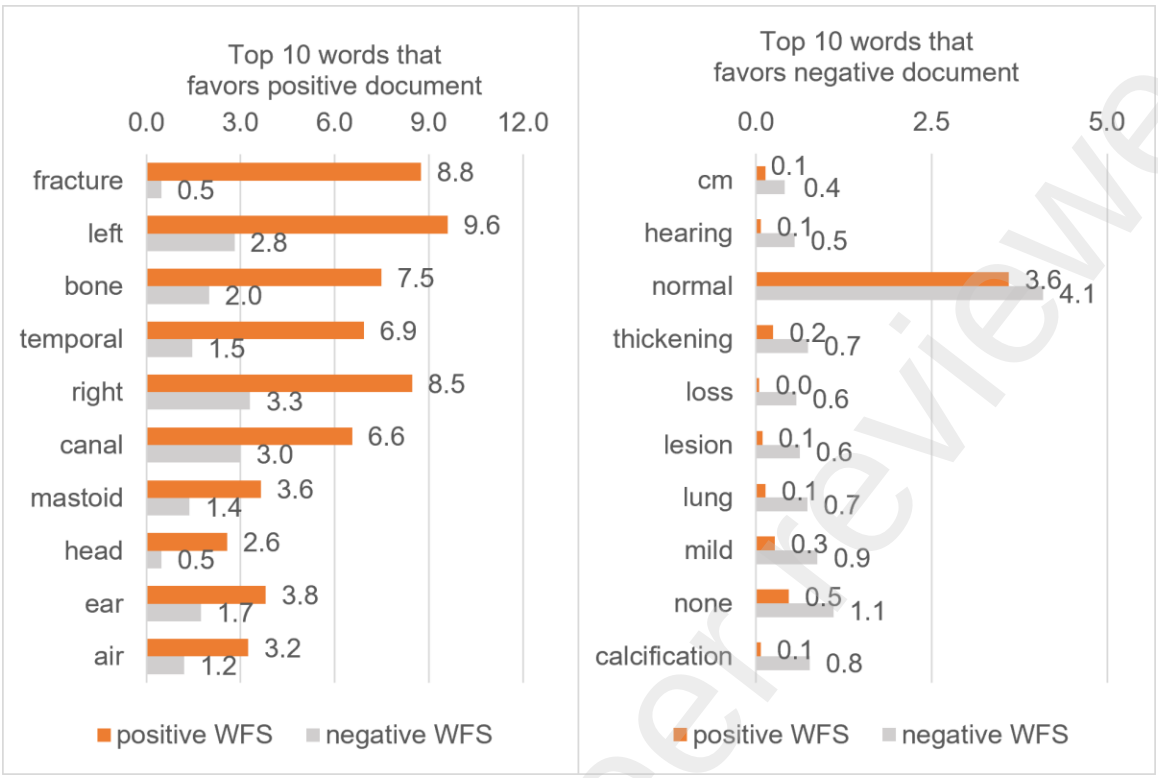

Figure 2: Comparison of gaps between fracture and non-fracture reports. The red bar stands for the frequency of fracture reports, and the blue bar stands for non-fracture reports. The left-side chart shows the top ten words that appear more often in fracture sets, whereas the right-side chart shows the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

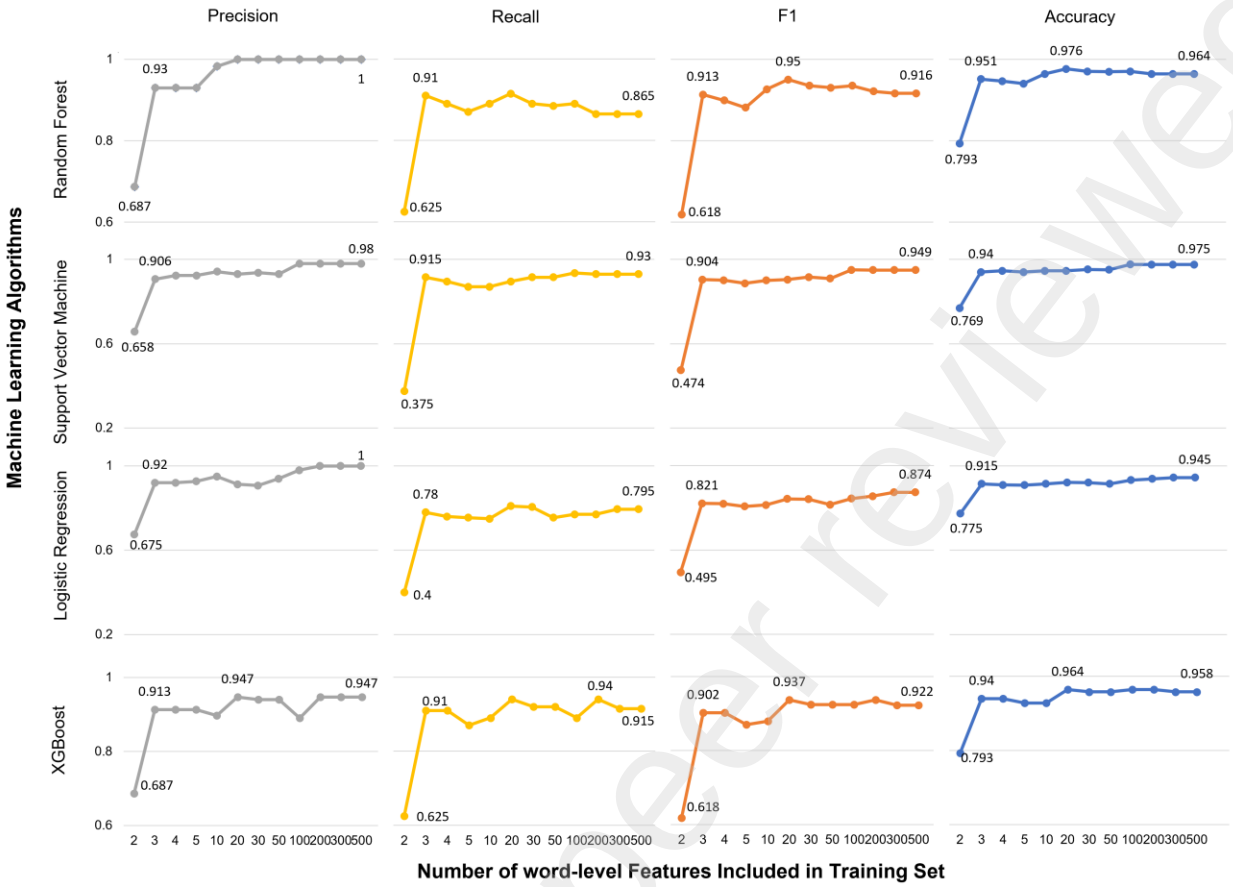

Figure 3: Relationships between classification model's performance, number of selected features, and evaluation performances for Random Forest, Support Vector Machine, and Logistic Regression model.

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.

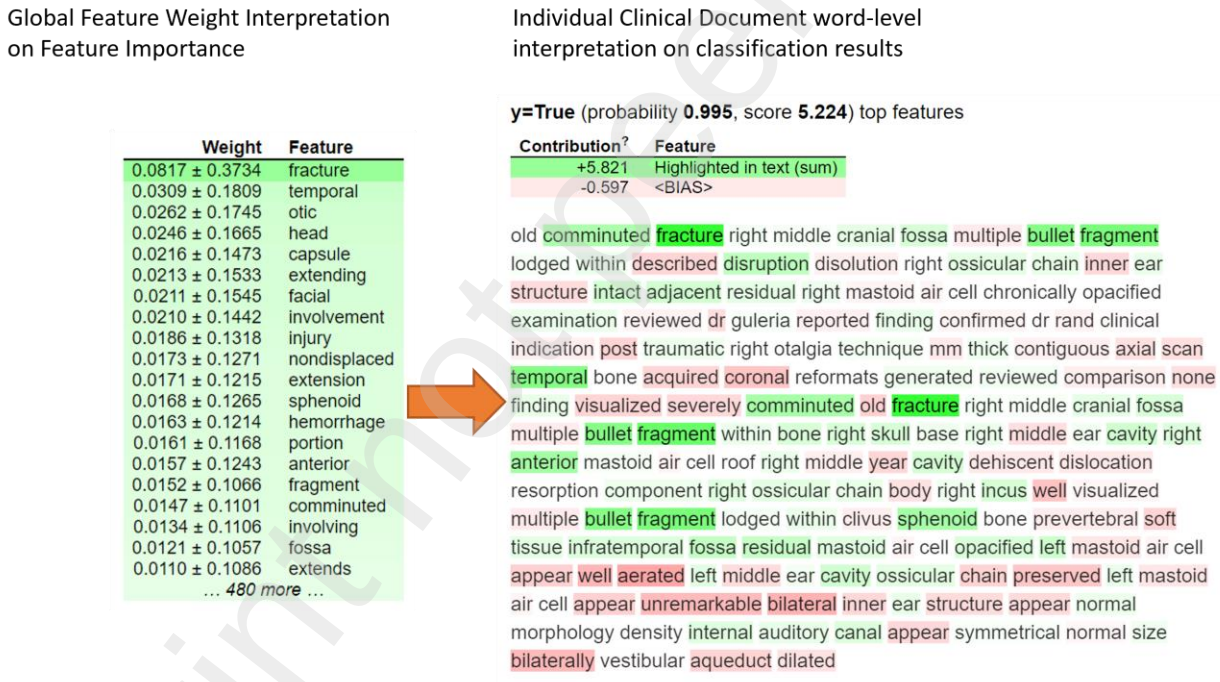

Figure 4: How LIME evaluate the importance of each word features and use the weight of features to visualize the word-level contribution for each document to calculate classification results

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.

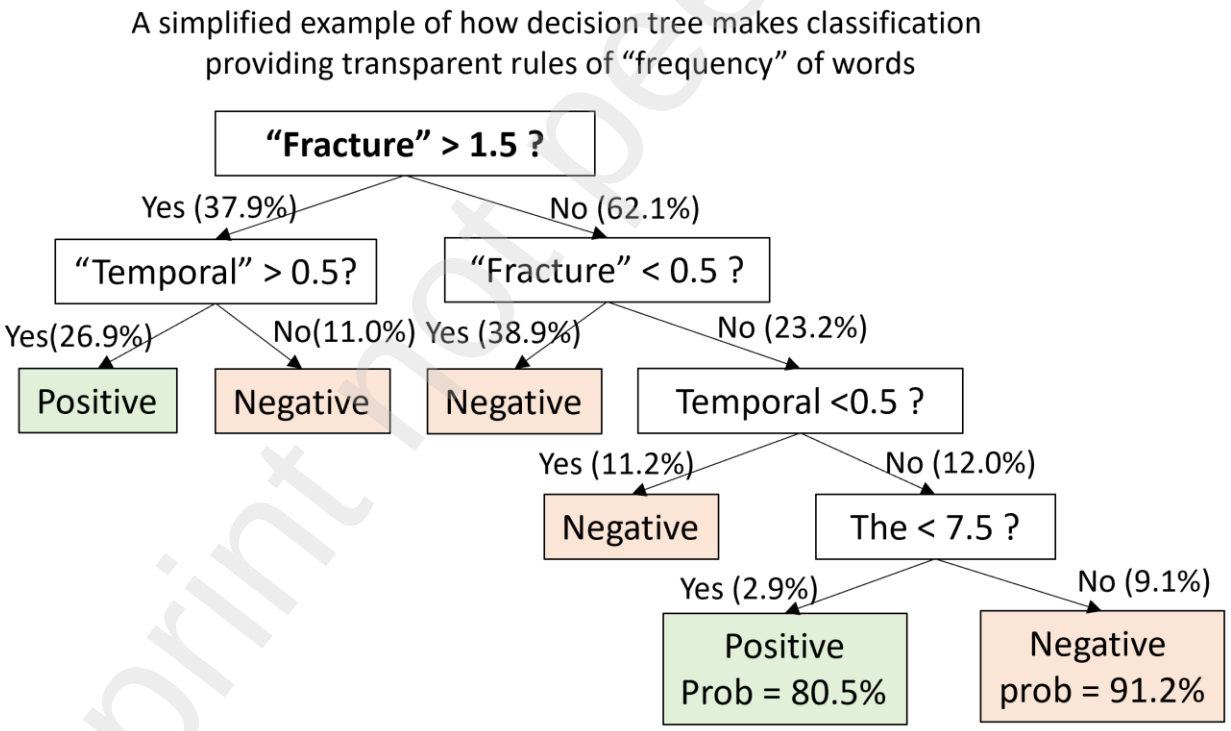

Figure 5: A visualization of how a transparent decision tree model determines the classification result from a sample CT text report. The percentage number shows the proportion of reports falling into each category. The frequency of a specific word determines the model’s classification result.

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.