PRAGUE UNIVERSITY
OF ECONOMICS
AND BUSINESS

Acta Informatica Pragensia
2023, Volume 12, Issue 2, pp. 260–274
https://doi.org/10.18267/j.aip.214

Article Open Access

Diagnostic Performance Evaluation of Deep Learning-Based Medical Text Modelling to Predict Pulmonary Diseases from Unstructured Radiology Free-Text Reports

Shashank Shetty [1,2], **Ananthanarayana V S** [1], **Ajit Mahale** [3]

[1] Department of Information Technology, National Institute of Technology Karnataka, Mangalore-575025, Karnataka, India
[2] Department of Computer Science and Engineering, Nitte Mahalinga Adyanthaya Memorial Institute of Technology (NMAMIT), NITTE (Deemed to be University), Udupi-574110, India
[3] Department of Radiology, Kasturba Medical College, Mangalore, Manipal Academy of Higher Education, Manipal-575001, India

Corresponding author: Shashank Shetty (shashankshetty@nitte.edu.in)

Abstract

The third most common cause of death worldwide is attributed to pulmonary diseases, making it imperative to diagnose them promptly. Radiology is a medical discipline that utilizes medical imaging to guide treatment. Radiologists prepare reports interpreting details and findings analysed from medical images. Radiology free-text reports are a rich source of textual information that can be exploited to enhance the efficacy of medical prognosis, treatment and research. Radiology reports exist in an unstructured format as are not suitable by themselves for mathematical computation or machine learning operations. Therefore, natural language processing (NLP) strategies are employed to convert unstructured natural language text into a structured format that can be fed into machine learning (ML) or deep learning (DL) models for information extraction. We propose a DL-based medical text modelling framework incorporating a knowledge base to predict pulmonary diseases from unstructured radiology free-text reports. We make detailed diagnostic performance evaluations of our proposed technique by comparing it with state-of-the-art NLP techniques on radiology free-text reports extracted from two medical institutions. The comprehensive analysis shows that the proposed model achieves superior results compared to existing state-of-the-art text modelling techniques.

Keywords

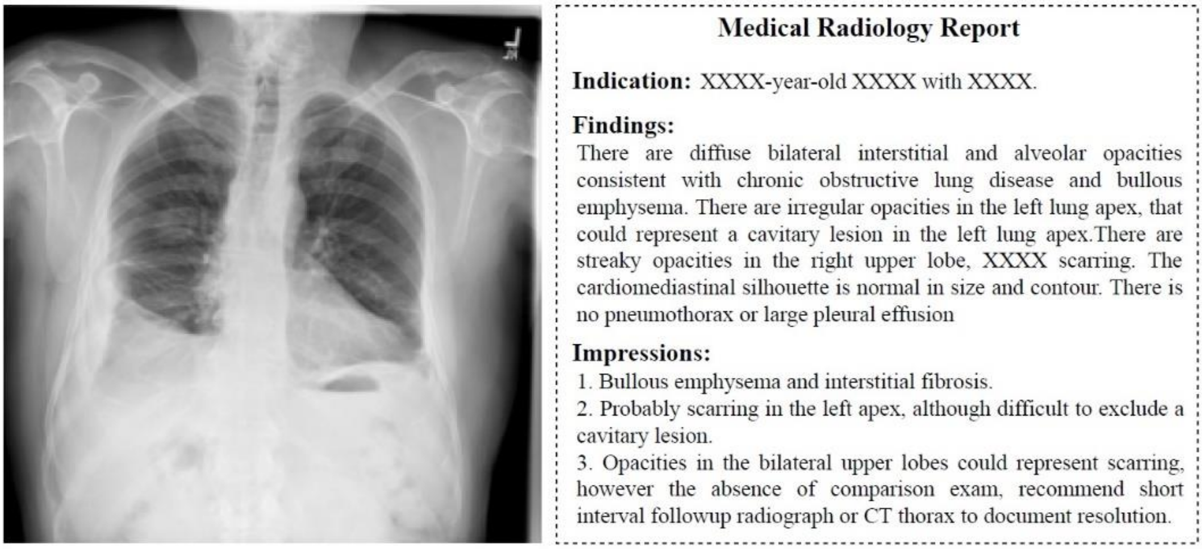
Radiology reports; Unstructured data; Natural language processing; Deep learning.

Citation: Shetty, S., Ananthanarayana, V S, & Mahale, A. (2023). Diagnostic Performance Evaluation of Deep Learning-Based Medical Text Modelling to Predict Pulmonary Diseases from Unstructured Radiology Free-Text Reports. *Acta Informatica Pragensia*, 12(2), 260–274 . https://doi.org/10.18267/j.aip.214

Academic Editor: Stanislav Vojir, Prague University of Economics and Business, Czech Republic

1 Introduction

Pulmonary diseases are a major cause of death worldwide due to tobacco smoking, air pollution, inhaling unwanted particles, radon gas and chemicals, etc. Pulmonary diseases involve various respiratory and lung disorders such as pneumonia, chronic bronchitis, pleural effusion and pulmonary fibrosis. The chances of risk involved in pulmonary diseases are high, and there is a need for timely treatment. The radiologist interprets a radiology imaging examination such as a chest X-ray to diagnose the conditions affecting the lungs. The radiologist analyses the chest X-ray and records their observations in descriptive reports to validate the prognosis (Shetty et al., 2022). Figure 1 shows an example of a radiology report prepared by a radiologist after analysing a raw chest X-ray image. These radiology reports contain rich information pertaining to patient demographics, disease findings and conclusive remarks on the abnormalities. This essential information retrieved from the clinical narratives can be leveraged to improve the efficacy of clinical assessment, treatment and research. As portrayed in the figure, the indication section depicts the examination procedure, the finding section provides a detailed outline of the clinical evaluation pertaining to the abnormalities, and the impression section indicates conclusive remarks about the disease.



Medical Radiology Report

Indication: XXXX-year-old XXXX with XXXX.

Findings:
There are diffuse bilateral interstitial and alveolar opacities consistent with chronic obstructive lung disease and bullous emphysema. There are irregular opacities in the left lung apex, that could represent a cavitary lesion in the left lung apex.There are streaky opacities in the right upper lobe, XXXX scarring. The cardiomediastinal silhouette is normal in size and contour. There is no pneumothorax or large pleural effusion

Impressions:
1. Bullous emphysema and interstitial fibrosis.
2. Probably scarring in the left apex, although difficult to exclude a cavitary lesion.
3. Opacities in the bilateral upper lobes could represent scarring, however the absence of comparison exam, recommend short interval followup radiograph or CT thorax to document resolution.

Figure 1. *Example of radiology free-text report derived from chest X-ray by radiologist.*

Usually, radiologists manually categorize clinical notes into normal (i.e., no disease) and abnormal (i.e., pulmonary diseases). Manual classification of radiology reports is labour-intensive, time-consuming and prone to human error. Rapid and accurate identification of information contained in radiology narratives will minimize the workloads, assist radiologists in decision making and prioritize patients with emergency care. Automating the task will benefit less experienced radiologists in predicting abnormality when there is a surge in cases with greater risk. There has been significant growth in the usage of ML and DL strategies for automating disease prediction tasks from EHRs. The unstructured nature of radiology free-text reports with complex vocabularies makes it difficult for ML/DL models to extract features from the raw text. NLP plays a key role in extracting structured information from clinical text (Pons et al., 2016; Shetty et al., 2023).

Automated prediction of pulmonary abnormalities from unstructured diagnostic notes can be further integrated into the existing medical diagnostic workflow to improve the health information system in the following ways:

- One potential approach would be integrating the model as a decision support tool for radiologists. In this scenario, the model could be used to automatically flag radiology reports that are likely to contain pulmonary abnormalities, which could then be prioritized for review by the radiologist. The model could also be used to suggest possible diagnoses based on irregularities detected in the report, which could help the radiologist arrive quickly at a more accurate diagnosis.

- Another approach would be to integrate the model into the electronic health record (EHR) system used by the hospital or clinic. In this scenario, the model could be used to automatically populate the patient's EHR with relevant diagnostic information based on the findings in the diagnostic notes. This could improve the accuracy and completeness of the patient's medical record, improving the quality of care delivered to the patient.
- Finally, the model could also be used as a screening tool for large-scale population health studies. By analysing radiology reports from a large cohort of patients, the model could identify patients at high risk of developing a pulmonary disease, which could inform targeted interventions and preventative care strategies.

Available unstructured text in clinical practice is scarce as most radiology reports are restricted to the private institution or are domain-specific (Shetty et al., 2022). DL models produce better results with a large cohort size. There is a need for a DL framework that accurately classifies abnormalities from radiology reports when the cohort size is small. Therefore, we propose an NLP-based DL model that incorporates a knowledge base to convert unstructured text into meaningful word embeddings.

The framework proposed in this study employs the GloVe embedding technique in combination with a knowledge base to represent the features of words. This approach is followed by the use of a deep neural network to predict the presence of pulmonary abnormalities. By utilizing GloVe embeddings, which capture semantic connections between words, the proposed framework improves the accuracy and efficiency of NLP techniques. Incorporating a knowledge base further enhances the model's predictive power by providing additional contextual information. The deep neural network component of the framework utilizes the feature representations generated by the GloVe embeddings and the knowledge base to learn patterns and relationships between words, thereby enhancing the accuracy of the predictions.

The primary contribution of this research can be summarized as follows:

- We propose a DL-based NLP task incorporating a knowledge base to predict pulmonary abnormalities from diagnostic clinical notes.
- We conduct a systematic and comprehensive performance evaluation to check the efficacy of our proposed model against standard NLP techniques.
- A quantitative analysis of the proposed model with a state-of-the-art NLP model is performed on a publicly available IU dataset and data collected from a private medical hospital.

2 Related Work

The advent of electronic health records (EHR) in clinical settings has given rise to a considerable amount of clinical data on the patient they serve. EHR comprise massive amounts of data containing valuable information in a structured (i.e., lab values and vital signs) or unstructured format (i.e., radiology narratives and clinical notes). Significant research has been carried out in the area of unstructured clinical text classification and prediction. In this section, we will highlight the related work on unstructured radiology report classification and summarize the results from the literature.

We can categorize unstructured report classification into two strategies: rule-based strategies (Sippo et al., 2013; Hassanpour et al., 2017) and standard ML-based strategies (Shetty et al., 2020; Dahl et al., 2021; Nakamura et al., 2021). The rule-based strategies completely depend on clinical ontologies that include SNOMED CT (see http://www.snomed.org), where conventional pattern matching is performed with pre-defined clinical words. The rule-based technique faces a significant challenge as its efficacy is based on the effectiveness of the pre-defined words. Machine learning-based methods use text encoding using NLP models and feature classification using ML/DL models. In their study, Shin et al. (2016) utilized term frequency-inverse document frequency (TF-IDF) as a method of text encoding and employed a convolutional neural network (CNN) that included an attention mechanism to categorize computed tomography (CT) radiology reports acquired from a private medical institution. The proposed model was compared with logistic regression (LR), random forest (RF) and support vector machine (SVM) applied to 1400 reports. The proposed attention technique achieved better performance compared to the three statistical models. Chen et al. (2018) proposed a DL framework to categorize clinical notes from CT imaging reports extracted from two private institutions. A CNN model with the GloVe word-embedding technique was utilized for classifying the reports and showcased the dominance of the presented technique compared to the traditional rule-based classifier PEfinder (Chapman et al., 2011). Dahl et al. (2021) proposed a CNN, a bidirectional recurrent neural network model (bi-LSTM) and SVM to classify and detect findings from Norwegian radiology CT reports. The bag of words (BoW) and TF-IDF word embedding techniques were utilized as an NLP strategy. The results obtained from the CNN and bi-LSTM models were marginally superior to those obtained from the conventional SVM model.

Nakamura et al. (2021) presented an automated detection and classification of actionable reports obtained from a Japanese private institution. The binary classification of CT reports was performed using four statistical methods: LR, gradient boosting decision tree (GBDT), bi-LSTM and bidirectional encoder representations from transformers (BERT). The BERT attained a significantly higher area under the precision-recall curve (AUPRC) than the other three statistical models. Bayrak et al. (2022) proposed a diagnostic note classification from magnetic resonance imaging (MRI) data acquired from a private medical institute in Turkey. The index-based word-encoding strategy for word-embedding conversion of free text and the long-short-term memory (LSTM) network, bi-LSTM and CNN for classifying the reports into epilepsy disease or not. The bi-LSTM showcased better performance compared to the other two DL strategies.

The above literature shows that the selection of the NLP task significantly affects the prediction or classification task on unstructured clinical notes. Most existing research utilizes radiology reports extracted from private medical facilities. Currently, the quantity of diagnostic notes accessible is restricted due to their confinement to private medical facilities or their focus on a specific field. There is a need for learning from unseen or rare medical words; therefore, we have proposed a DL-based NLP framework that utilizes a knowledge base to predict abnormality in radiology free-text reports.

3 Methodology

The overall proposed DL-based NLP framework for predicting pulmonary diseases from radiology free-text reports is shown in Figure 2. The findings sections of reports are extracted from the corpus and passed through the basic pre-processing procedure to cleanse the information. The refined text is fed into various NLP techniques to convert the text into meaningful word embeddings. Next, the acquired textual embeddings are utilized as an input for a DNN-based prognostic model to forecast pulmonary illnesses.

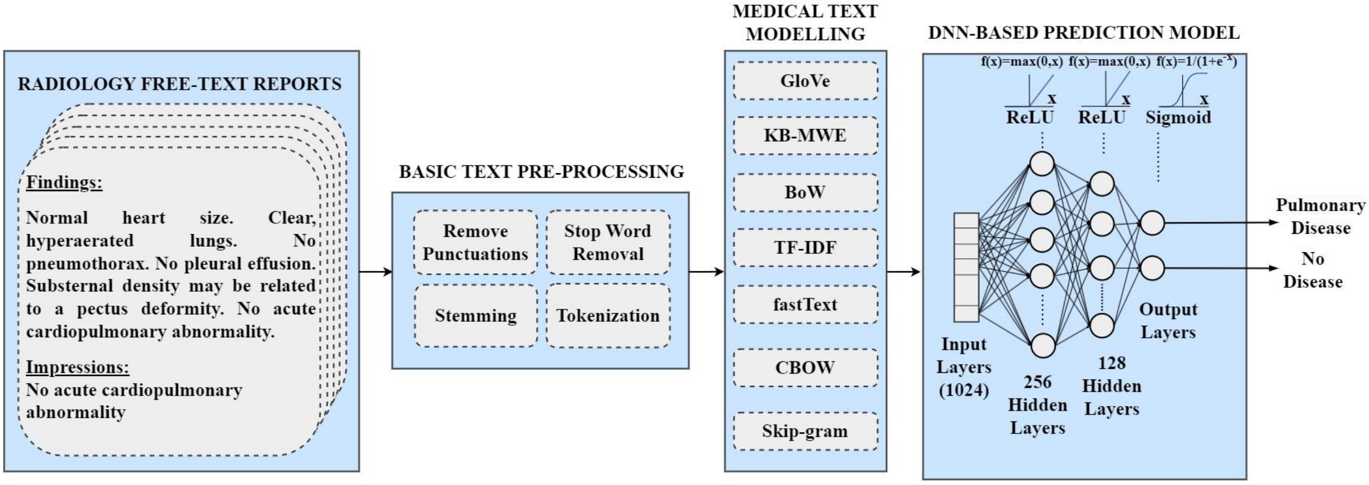

RADIOLOGY FREE-TEXT REPORTS
Findings:
Normal heart size. Clear, hyperaerated lungs. No pneumothorax. No pleural effusion. Substernal density may be related to a pectus deformity. No acute cardiopulmonary abnormality.
Impressions:
No acute cardiopulmonary abnormality
BASIC TEXT PRE-PROCESSING
Remove Punctuations
Stop Word Removal
Stemming
Tokenization
MEDICAL TEXT MODELLING
GloVe
KB-MWE
BoW
TF-IDF
fastText
CBOW
Skip-gram
DNN-BASED PREDICTION MODEL
f(x)=max(0,x)
f(x)=max(0,x)
f(x)=1/(1+e^-x)
ReLU
ReLU
Sigmoid
Input Layers (1024)
256 Hidden Layers
128 Hidden Layers
Output Layers
Pulmonary Disease
No Disease

Figure 2. *Proposed DL-based NLP framework for predicting pulmonary diseases from diagnostic notes.*

3.1 Basic text pre-processing

To cleanse the data and make them ready to feed into the models, we pass the radiology report findings from both the cohorts through a sequence of text pre-processing stages.

1. **Punctuation removal:** Punctuation symbols, such as periods, commas, colons and semicolons, do not contribute any significance to the text and their inclusion may impede the processing of textual data. When analysing radiology reports, eliminating punctuation guarantees that the diagnostic conclusions are properly handled. For instance, the existence of unnecessary punctuation within the report may perplex the model in identifying the accurate disease condition.
   For example, consider the following sentence from a pulmonary disease report: “Patient presented with shortness of breath, cough, and wheezing”. After the removal of punctuation, the sentence becomes: “Patient presented with shortness of breath cough and wheezing”. Notice how the commas have been removed from the sentence, making it easier to process for machine learning models. Punctuation removal also helps reduce the dimensionality of the text data, making it more manageable for downstream tasks such as feature extraction and classification.
2. **Stop word removal:** Stop words are commonly used words that do not provide much context to the text data and can be safely removed from the dataset without losing important information. In clinical notes such as pulmonary disease reports, stop word removal can be an essential step to remove unnecessary words that do not contribute to the diagnosis or prediction of the disease. As an illustration, when analysing a pulmonary disease report, certain common stop words such as “a”, “an”, “the”, “and”, “in”, “on” and “of” may occur frequently. These words do not provide any specific information regarding the illness and may even introduce noise into the data. By eliminating these words, the data can become more concise and meaningful.
   Consider the following sentence from a pulmonary disease report: “The patient has a history of smoking, which may have contributed to the development of chronic obstructive pulmonary disease (COPD)”. In this sentence, the stop words are “the”, “has”, “a”, “of”, “which”, “may”, “have”, “to”, “the”, and “of”. Removing these words would result in the sentence: "patient history smoking contributed development chronic obstructive pulmonary disease (COPD)." The remaining words convey the same meaning as the original sentence and can be more efficiently processed by NLP models.
3. **Stemming:** Stemming is a technique for standardizing text that involves reducing words to their root or base form. This method aids in normalizing the textual data by consolidating various forms of words, such as plurals and verb tenses, into a solitary representation. In the context of radiology

reports, stemming can help identify related words and reduce the feature space of the data, which in turn can enhance the proposed DL model's performance.

To illustrate, in a diagnostic note on predicting pulmonary disease, the term “effusions” could be simplified to “effusion”, making it possible to detect all occurrences of effusion in the text, regardless of whether the word is in its singular or plural form. Likewise, the term “fibrotic” could be reduced to “fibros”, making it possible to detect all instances of fibrosis in the text, whether the word is used as an adjective or a noun.

4. **Tokenization:** The act of dividing a document or sentence into smaller units called tokens is known as tokenization. When dealing with radiology reports, tokens could signify either single words or phrases, and they can be beneficial in detecting patterns and connections within the textual information.

   Let's take an instance of a sentence from a radiology report: “A nodule in the right lung was detected on the chest X-ray”. After tokenization, the sentence may be broken down into the following tokens: [“A”, “nodule”, “in”, “right”, “lung”, “was”, “detected”, “on”, “the”, “chest”, “X-ray”]. In this sentence, each token corresponds to an individual word. Tokenization plays a crucial role in recognizing significant patterns and connections within the textual information, as well as extracting useful information that can be utilized in tasks such as disease prognosis or image retrieval.

In this research, the first step involves eliminating the punctuation, and then frequently used words, known as stop words, are removed from the corpus. Afterwards, the text is subjected to a standardization process known as stemming, wherein the words are transformed into their root or base form. Lastly, the raw text is fragmented into smaller segments known as tokens.

3.2 Medical text modelling

We propose a knowledge based-medical word embedding (KB-MWE) technique that generates the learned representations of the medical words from the report by jointly learning from the radiology cohort and the knowledge base. We also discuss standard text modelling techniques employed to compare the performance of the proposed framework. We use the GloVe word embedding technique as a base for our proposed knowledge-based medical word embedding (KB-MWE) technique. Therefore, we will first review the GloVe model before presenting the proposed KB-MWE model.

1. **Global vectors (GloVe):** GloVe (Pennington et al., 2014) is an unsupervised learning technique introduced by researchers from Stanford University focusing on producing word vectors by learning from global word-word co-occurrence matrices obtained from the given corpus.

   With respect to medical text modelling, the Glove model comprises of the following steps:

   – Given a radiology report cohort RC, the radiology word co-occurrence statistics are gathered and arranged in a matrix format known as a word co-occurrence matrix (Y^R). Each entry Y_{ij}^R in the co-occurrence matrix indicates how often the context radiology word, $\widetilde{rw}_j$ appears in the target radiology word, rw_i. The radiology cohort RC is scanned and for each target word rw_i, the context word $\widetilde{rw}_j$ is defined by a window size before and after the target word. Smaller weights are allocated for more distant words.

   – The soft constraints for each radiology word pair are defined as follows:

$$rv_i^T \cdot \widetilde{rv}_j + b_i + b_j = log(Y_{ij}^R) \qquad (1)$$

   Where b_i and $\widetilde{b_j}$ depict the scalar bias terms with respect to the radiology target and context words, $\widetilde{rv}_j$ and rv_i denote the word vectors of a context radiology word $\widetilde{rw}_j$ and a target radiology word rw_i, respectively.

- We define a weighted least square cost function J_{RC}, which focuses on minimizing the difference between the dot product of two radiology word embeddings (i.e., $\widetilde{rv}_j$ and rv_i) and the logarithm of total occurrences of the radiology word vectors obtained from Y^R.

$$J_{RC} = \sum_{i,j=1}^{Voc} f(Y_{ij}^R)(M^{RC} + b_i + \widetilde{b_j} - log(Y_{ij}^R))^2 \quad (2)$$

  Where Voc represents the vocabulary comprising of the word collection in the radiology report cohort, $M^{RC} = rv_i^T \cdot \widetilde{rv}_j$ indicates the inner dot product of radiology target and context word vectors comprising rw_i and $\widetilde{rw}_j$. The weighted function $f(Y_{ij}^R)$ aids in the prevention of learning from commonly occurring pairs of words. The weighted function is given as follows:

$$f(Y_{ij}^R) = \begin{cases} \left(\frac{Y_{ij}^R}{Y_{max}^R}\right)^\alpha & if\ Y_{ij}^R < Y_{max}^R \\ 1 & otherwise \end{cases} \quad (3)$$

  The efficacy of the GloVe model depends on the cut-off; therefore, we initialize α to 3/4 and Y_{max}^R to 100.

2. **Knowledge-based medical word embedding (KB-MWE):** In the GloVe technique, the learning of word vectors is performed only on the radiology corpus and does not utilize any existing knowledge base. Therefore, unseen or rare clinical words are not captured when the cohort size is small in number. Radiology cohorts available today are very small, restricted to private health centres or specific to a particular domain. This produces a challenge for word embedding models in learning word vectors from the limited vocabulary size. To solve the above issue, we proposed a medical text modelling strategy using a DL model that incorporates an existing knowledge base to acquire and learn the infrequent word embeddings. Considering the existing radiology knowledge base RKB, following are the steps involved in the proposed KB-MWE:

- We acquired a radiology knowledge base, which comprises word vectors learnt on 4.5 million Stanford diagnostic notes (Zhang et al., 2018). There are no fixed rules to include any particular RKB. Any radiology knowledge base with a semantic connection between radiology words can be considered a radiology knowledge base.

$$M^{RKB} = rkv_i^T \cdot \widetilde{rkv}_j \quad (4)$$

  Where M^{RKB} denotes the scalar dot product of the radiology knowledge vectors (i.e., $rkv_i, \widetilde{rkv}_j \in \mathbb{R}^{RKB}$) corresponding to the radiology words (i.e., rw_i and $\widetilde{rw}_j$) in a radiology report cohort RC. The cost function of the KB-MWE (J_{RKB}) can be derived as follows:

$$J_{RKB} = \sum_{i,j=1}^{Voc} f(Y_{ij}^R)(M^{RKB} + b_i + \widetilde{b_j} - log(Y_{ij}^R))^2 \quad (5)$$

- We utilize a radiology knowledge base that consists of word embeddings derived from 4.5 million Stanford radiology reports as the weight matrix for embeddings. This knowledge base comprises l-dimension radiology knowledge vectors, $rkv_{j:l}^1$, $rkv_{j:l}^2$, ... , $rkv_{j:l}^r = rkv_{j:l}^{i:r}$, (where j is a set from 1 to l) for all the radiology words r. By conducting matrix multiplication between the one-hot vector hv for each radiology term obtained from the vocabulary and the radiology knowledge vectors retrieved from the embedding weight matrix, we obtain the corresponding word vectors for each radiology term. Let r_1, r_2, ..., r_m denote the unique radiology terms (i.e., vocabulary) extracted from the radiology cohort RC. We generate the one-hot vector for each term in the vocabulary (i.e., $hv_{j:m}$). The radiology knowledge vectors produced for each radiology word in the corpus are given as follows:

$$\widetilde{rkv}_{j:l}^{i:r} \leftarrow hv_{j:m} \times rkv_{j:l}^{i:r} \tag{6}$$

3. **Bag of words (BoW):** BoW (Sivic et al., 2009) is the primary word embedding technique in which the sequence of words is converted into a bag of words. In the BoW strategy, the word occurrence count is calculated, disregarding the grammar. A unique set of words is extracted from the radiology cohort, and the frequency of each word is calculated to create a vocabulary. The major drawback with the above technique is that the word count will provide information about the occurrences of the word in a cohort ignoring the context of the word.

4. **Term frequency — inverse document frequency (TF-IDF):** TF-IDF (Sammut & Webb, 2011) provides a statistical evaluation of how relevant a radiology word is to a cohort. The TF-IDF comprises two steps: firstly, term frequency is calculated by counting the word occurrences in a radiology cohort. The inverse document frequency will diminish the weights provided on common words. TF-IDF is based on BoW; consequently, it does not capture the semantics of the medical word in a cohort.

5. **FastText:** FastText (Bojanowski et al., 2017) is a type of word2vec model, where an n-gram (i.e., sequences of adjacent characters) of characters represents each radiology word. By doing so, the model can comprehend the meanings of shorter words and grasp the prefixes and suffixes associated with radiology terminology. After the n-gram representation of the words, the skip-gram strategy is applied to learn the word embeddings. A major drawback of this method is the higher memory requirement as the model deals with the character of words.

6. **Continuous bag of words (CBOW):** CBOW (Wang et al., 2017) is an unsupervised word2vec-based model which takes radiology context words as input and predicts the radiology target word. The lambda and softmax layers are utilised to learn the word embeddings by a backpropagation strategy. We update weights in the embedding layer with each epoch using the gradient descent technique.

7. **Skip-gram:** Skip-gram (Sivic et al., 2009) model predicts the radiology context word given the target word. A positive and negative input sample is created and fed as input to the model. The model utilizes these samples to gain an understanding of the context and develop semantic embeddings for every radiology term. The radiology target and context word pair given as input is merged to compute the dot product of the word embeddings. These embeddings are then ingested through the sigmoid layer that provides either 0 or 1 as an output. The obtained output is compared with the original label, and the loss is calculated by backpropagating for every epoch.

3.3 Deep neural network-based prediction model

We apply a fully connected DNN-based prediction model to predict pulmonary abnormalities. A DNN consists of multiple layers with connected nodes named neurons inspired by biological neural networks. A DNN comprises several connected modules between the input and output layers. The textual encoded features produced from the medical text modelling layers are given as input to DNN layers for predicting the pulmonary abnormality from radiology free-text reports. A four-layer DNN network is proposed where the ReLU non-linear function is employed in the penultimate layers, and the sigmoid activation function is applied in the final prediction layer. The dropout mechanism of 0.2 is utilized to prevent any overfitting problems. The primary operation of the DNN on the text encoding feature is shown below:

$$z'_{i+1} = f(W_i \cdot z'_i + b_i), \tag{7}$$

Here, f() indicates the ReLU and the sigmoid activation function, z'_i denotes discriminative textual features extracted from the text modelling strategies, w_i indicates the weighted array and b_i represents the scalar bias.

4 Experimental Setup and Results

This section provides a comprehensive explanation of the cohort selection process for the radiology report dataset. Additionally, it highlights the evaluation metrics that were employed to measure the effectiveness of the proposed framework. Finally, a detailed analysis of the results obtained and diagnostic outcomes is presented.

4.1 Radiology report cohort selection

We considered two radiology report cohorts to comprehensively evaluate the proposed framework for unstructured radiology free-text abnormality prediction. A comprehensive description of the two radiology cohorts is presented in Table 1. The Indiana University (IU) dataset is a publicly available radiology cohort containing chest X-rays and associated radiology reports. The dataset is widely used for cross-modal retrieval tasks (Liu et al., 2019), and there is limited work with respect to classification and prediction tasks. The IU dataset with 3996 reports consists of findings, indication, impression and medical subject heading (MeSH) sections. We discarded reports with an incomplete findings/impression section and selected 3638 radiology reports. The MeSH encodings are extracted from the report, containing labels indicating whether the disease is present or not. To check the generalization ability of the proposed framework, we applied it to the cohort obtained from the KMC Hospital in Mangalore, India. The necessary steps were taken to ensure the privacy and security of the patient data used. The cohort collected from the private medical hospital was delinked and de-identified so that there would not be any direct link or any identification to the patient. The study adheres to all applicable laws and regulations regarding the use and protection of patient data, including obtaining necessary ethical approvals and informed consent from patients. The Institutional Ethical Committee permitted the use of the de-identified radiology cohort for research purposes. Expert radiologists were involved in annotating the radiology free text into normal and abnormal cases. The rationales behind selecting the two diagnostic report cohorts collected are as follows:

- The datasets are obtained from two different geographical locations: the publicly available IU dataset (USA) and the KMC hospital dataset (India), see Additional Information and Declarations section. Pulmonary abnormalities may have different prevalence rates and patterns in various parts of the world. Testing the model on datasets from different regions can help evaluate the performance of the proposed framework in identifying abnormalities in multiple populations. Therefore, selecting datasets from other geographic areas can be a validation strategy to assess the effectiveness of the proposed model in different contexts.
- Testing the proposed model on datasets from different geographic regions can provide valuable insights into the generalizability and robustness of the model across diverse populations.
- Furthermore, since the number of unstructured radiology reports available for research purposes is limited, conducting this benchmark study can contribute to enhancing the medical research domain.

Table 1. *Radiology free-text report corpus characteristics.*

Cohort description	IU dataset	KMC dataset
Total number of diagnostic free-text reports	3,996	502
Total number of cases selected	3,638	502
Total number of sentences	11,541	3,649
Total number of words	91,171	17,198

Cohort description	IU dataset	KMC dataset
Total vocabulary size	1,568	393
Total number of training/validation samples	3,264	451
Total number of test samples	374	51
Total percentage of cases with no diseases	38%	52%
Total percentage of cases with pulmonary diseases	62%	48%

4.2 Evaluation metrics

We use six assessment measures, namely accuracy, precision, recall, F1-score, the Matthews correlation coefficient (MCC) and the area under receiver operating characteristic curve (AUROC), for analysing the performance of the proposed word embedding model with standard NLP models (Powers et al., 2011).

1. **Accuracy:** It calculates the performance of the model by evaluating the percentage of correct predictions. When predicting pulmonary disease, accuracy aids in assessing the model's capacity to accurately identify the presence or absence of the disease in radiology reports. Accuracy measures the number of correct predictions made by the model in relation to the total number of predictions made.
2. **Precision:** This metric measures the ratio of correct pulmonary disease predictions to the total number of positive predictions generated by the DL framework. Precision is a crucial metric in pulmonary disease prediction as it helps in evaluating the proportion of true positive cases among all cases predicted as positive.
3. **Recall:** This metric calculates the ratio of correctly predicted pulmonary disease cases to the total number of positive cases present in the diagnostic cohorts. In the context of pulmonary disease prediction, recall is significant as it measures the model's capability to correctly detect all the classes of abnormalities present in the data.
4. **F1-score:** This metric is the harmonic average of precision and recall, which achieves a balance between the two evaluation measures. The F1-score is useful for evaluating the overall performance of the pulmonary disease prediction model in correctly predicting both positive (disease) and negative (no disease) cases. This evaluation measure also aids in assessing the model's performance when dealing with imbalanced classes.
5. **MCC:** This metric quantifies the correlation between the predicted and actual binary classifications and is particularly valuable when dealing with datasets that have imbalanced classes. MCC is a useful metric for evaluating the performance of a model in correctly predicting both positive and negative cases in the context of pulmonary disease prediction, particularly when the dataset is imbalanced.
6. **AUROC:** This metric evaluates the model's capability to distinguish between positive (disease) and negative (no disease) cases, providing a measure of its overall performance. AUROC is a useful metric in evaluating the ability of the DL model to distinguish between cases with pulmonary disease and those without, thereby measuring its overall discriminatory performance.

To evaluate the performance of the proposed word embedding model for pulmonary disease prediction, it is essential to consider the above six evaluation metrics. These metrics provide a comprehensive assessment of the model's ability to accurately predict both positive and negative cases, considering the imbalanced nature of the dataset and the model's discriminatory power. Overall, these metrics are crucial in determining the effectiveness of the model in pulmonary disease prediction.

5 Results and Discussion

In order to comprehensively examine the proposed word embedding model, we compared its performance with the BoW, TF-IDF, FastText, CBOW, Skip-gram and GloVe embedding models. To compare the effectiveness of the proposed model with the standard NLP techniques, we trained each of the six techniques on the private medical institution dataset from scratch. We used a server with the following specifications for our experimental analysis: NVIDIA M40 server, 3TB HD, 128GB RAM and Ubuntu server OS. We used a TensorFlow DL framework to implement the proposed model and state-of-the-art text modelling models. The DL framework was trained for 100 epochs and 10-cross fold validation with the early stopping strategy. The quantitative benchmarked results on the Indiana University and the KMC cohort is showcased in Tables 2 and 3.

Table 2. *Benchmarked performance analysis results of the proposed DL-based NLP technique with standard text modelling techniques on the diagnostic clinical free-text cohort collected from the publicly available Indiana University dataset.*

Model	Accuracy	Precision	Recall	F1-score	MCC	AUROC
BoW	87.32%	0.8553	0.8150	0.8690	0.7253	0.8899
TF-IDF	87.32%	0.8776	0.8026	0.8714	0.7336	0.8899
FastText	89.30%	0.8791	0.8608	0.8916	0.7931	0.9152
CBOW	89.37%	0.8956	0.8535	0.8916	0.7924	0.9221
Skip-gram	89.95%	0.8982	0.8601	0.8978	0.7924	0.9260
GloVe	87.27%	0.8741	0.8727	0.8729	0.7232	0.9333
Proposed KB-MWE	**90.40%**	**0.9080**	**0.9040**	**0.9059**	**0.7939**	**0.9555**

Table 3. *Benchmarked performance analysis results of the proposed DL-based NLP technique with standard text modelling techniques on the diagnostic clinical free-text cohort collected from KMC hospital.*

Model	Accuracy	Precision	Recall	F1-score	MCC	AUROC
BoW	86.05%	0.9027	0.8358	0.8526	0.7418	0.8610
TF-IDF	87.05%	0.9340	0.8396	0.8689	0.7768	0.8795
FastText	92.84%	0.9271	0.8974	0.9192	0.8701	0.9295
CBOW	92.72%	0.8764	0.8888	0.8822	0.8608	0.9294
Skip-gram	92.72%	0.9236	0.9214	0.9343	0.8608	0.9208
GloVe	92.53%	0.9270	0.9250	0.9290	0.8500	0.9630
Proposed KB-MWE	**94.13%**	**0.9475**	**0.9413**	**0.9443**	**0.8827**	**0.9651**

The outcomes show that the proposed DL framework with KB-MWE outperforms the standard DL strategies. The proposed KB-MWE has a staggering improvement in the accuracy of 90.40% and 94.13% on the IU and KMC datasets, respectively, showcasing the model's prediction performance compared to the other standard models. Our proposed KB-MWE achieved 3% better precision compared to the CBOW and Skip-gram models, proving its ability to predict abnormal classes correctly. The increase in recall denotes lesser chances of abnormal classes being unpredicted. The superior F1-score and MCC of the KB-MWE indicate that the proposed model can accurately predict pulmonary disease despite a class imbalance problem. The proposed model obtained an AUROC of 0.9555 and 0.9651 for the IU and KMC radiology cohorts, depicting that the model can accurately predict the normal and abnormal class. A graphic visualization depicting the performance evaluation of the proposed KB-MWE with the standard NLP strategies on the IU and KMC radiology cohorts is shown in Figures 3 and 4. The analysis shows that the knowledge base incorporated for the proposed medical text modelling technique significantly affected the performance by learning unseen or rare medical words. Therefore, the presented KB-MWE can be incorporated when there is a low data condition while training the DL frameworks.

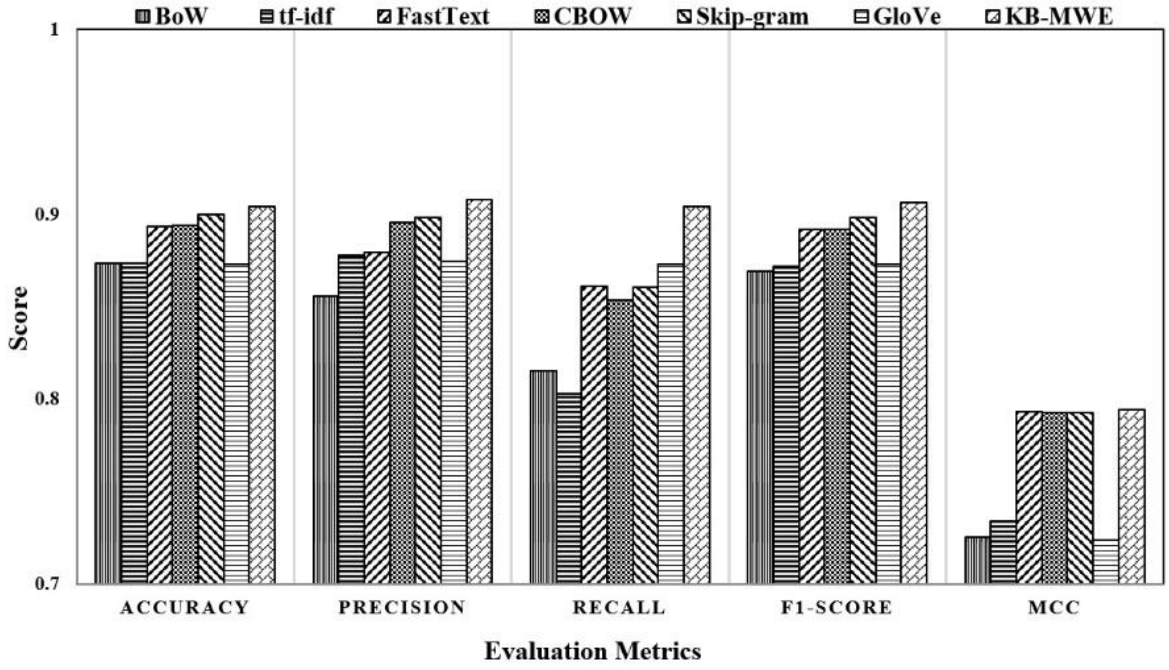


Figure 3. *Performance analysis of KB-MWE with state-of-the art NLP models on IU cohort.*

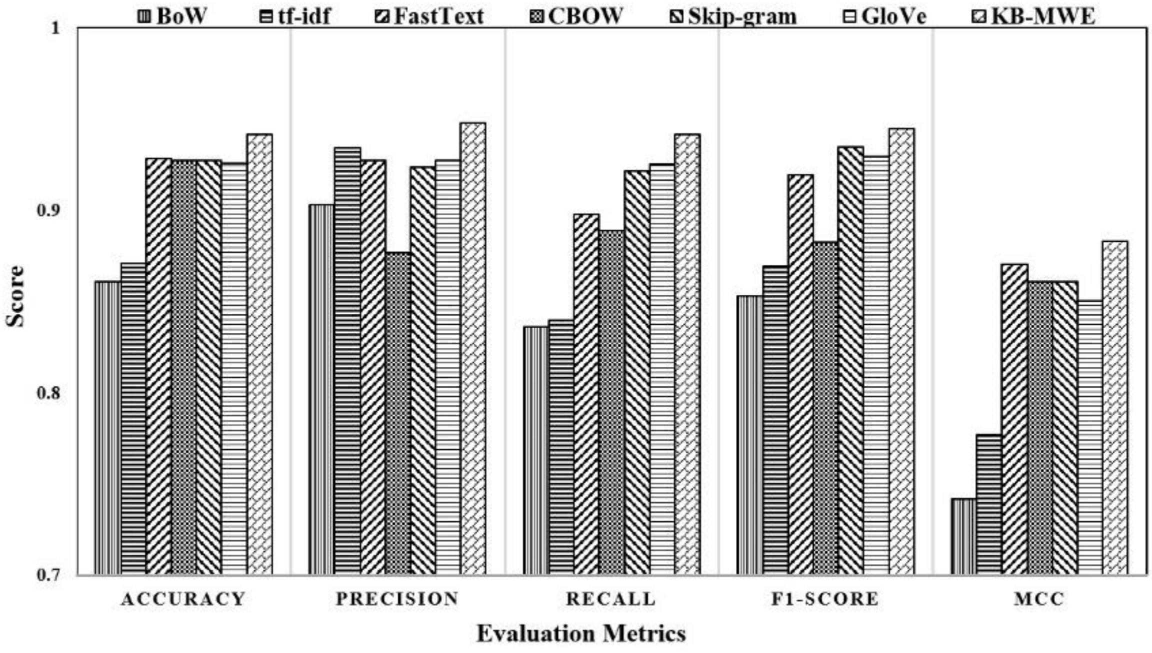
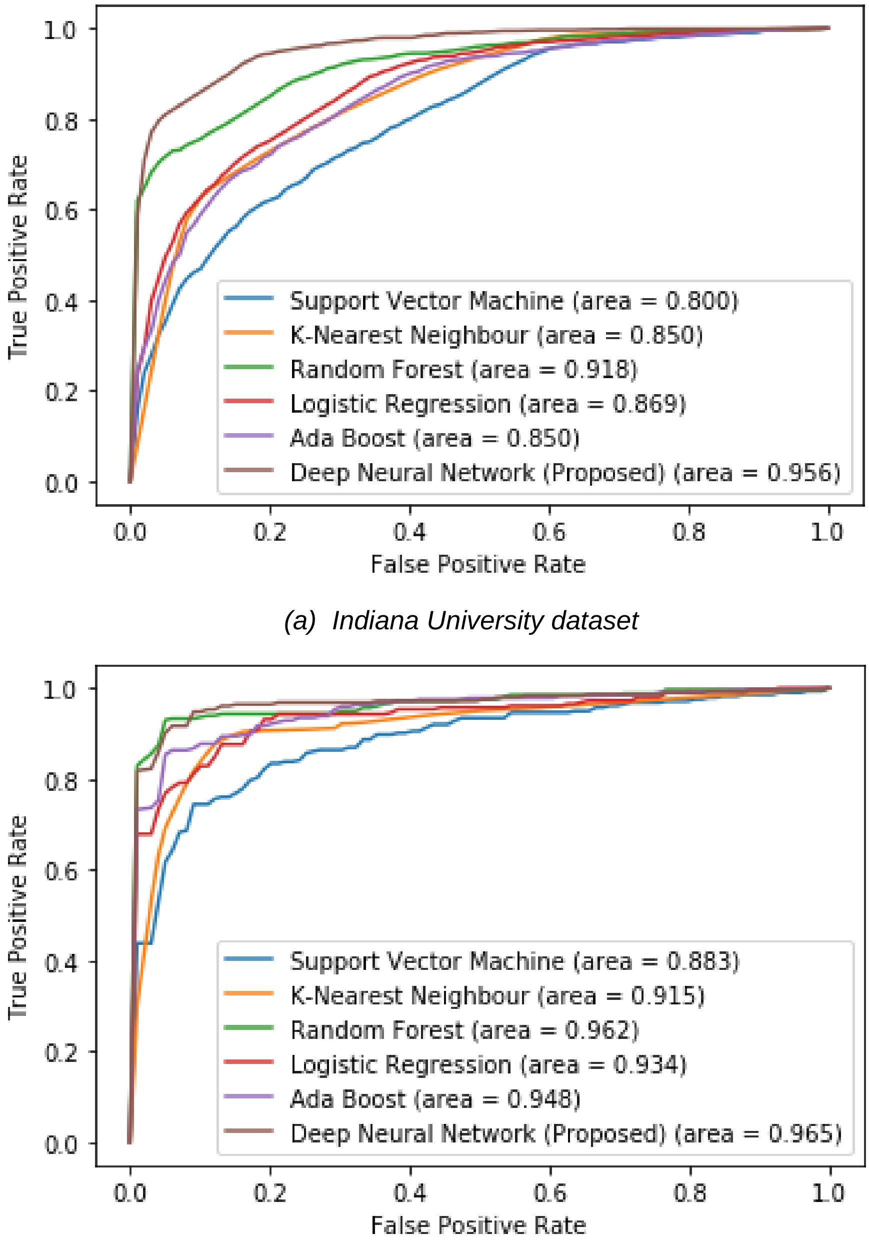


Figure 4. *Performance analysis of KB-MWE with state-of-the art NLP models on KMC cohort.*

We also compared the word embeddings extracted from the KB-MWE medical text modelling technique with various state-of-the-art ML classifiers such as SVM, k-nearest neighbour, random forest, logistic regression and Ada boost. The DNN performed better with an AUROC of 0.956 compared with other classifiers. Figure 5 presents the AUROC performance comparison of the proposed DNN model with the standard ML techniques.



(a) Indiana University dataset

(b) KMC hospital dataset

Figure 5. *Comparing AUROC performance of proposed DL model with state-of-art machine learning techniques.*

6 Conclusion and Future Research

To conclude, we proposed a knowledge-based medical word embedding technique with a DL framework to predict pulmonary diseases in radiology free-text reports. We compared the proposed word embedding technique with standard NLP models and found that the proposed KB-MWE achieves superior performance. We also compared the performance of the DNN classifier with the standard machine learning-based classifier and saw that the DNN classifier generated a better AUROC than ML classifiers. We observe that the increase in performance of the KB-MWE is due to the incorporation of the radiology knowledge base, which aids the prediction accuracy even though the cohort used for training is small in number. Therefore, the proposed model can be incorporated when data are scarce, as is common in the medical domain since the cohorts are specific to a private institution or are restricted to certain domains. In future avenues, we would like to build a DL model that can utilize valuable information extracted from radiology free-text reports to improve the radiology imaging feature classification.

Additional Information and Declarations

Acknowledgments: We would like to thank the Department of Information Technology, National Institute of Technology Karnataka, Surathkal, for providing us with the resources for carrying out this research. We would like to thank KMC, Mangalore, for technical assistance.

Funding: This research received no specific funding.

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: S.S.: Conceptualization, Methodology, Investigation, Software, Validation, Writing – Original draft preparation, Data Curation. A.V.S.: Conceptualization, Methodology, Supervision, Validation, Writing – Reviewing and Editing. A.M.: Validation, Writing – Reviewing and Editing.

Institutional Review Board Statement: Ethical review and approval were waived by the Institutional Ethics Committee, Kasturba Medical College, Mangalore, because this study was performed on the de-identified/de-linked chest X-rays (ref. IEC KMC MLR 01-2020/80 dated 15 January 2020).

Informed Consent Statement: Patient consent was waived since the chest X-rays utilized for this study were de-identified/de-linked and there was no direct or indirect contact with patients.

Data Availability: The study was performed on two datasets: (a) the Indiana University dataset that is publicly available (https://openi.nlm.nih.gov/faq, accessed on 2 April 2023) and (b) the KMC hospital dataset, which is the data collected from the private hospital and is available upon reasonable request from the corresponding author.

References

Bayrak, Ş., Yucel, E., & Takci, H. (2022). Epilepsy radiology reports classification using deep learning networks. *Computers, Materials & Continua*, 70(2), 3589–3607. https://doi.org/10.32604/cmc.2022.018742

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistic*, 5, 135–146.

Chapman, B. E., Lee, S., Kang, H. P., & Chapman, W. W. (2011). Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. *Journal of Biomedical Informatics*, 44(5), 728–737. https://doi.org/10.1016/j.jbi.2011.03.011

Chen, M. C., Ball, R. L., Yang, L., Moradzadeh, N., Chapman, B. E., Larson, D. B., Langlotz, C. P., Amrhein, T. J., & Lungren, M. P. (2018). Deep learning to classify radiology Free-Text reports. *Radiology*, 286(3), 845–852. https://doi.org/10.1148/radiol.2017171115

Dahl, F. A., Rama, T., Hurlen, P., Brekke, P., Husby, H., Gundersen, T., Nytrø, Ø., & Øvrelid, L. (2021). Neural classification of Norwegian radiology reports: using NLP to detect findings in CT-scans of children. *BMC Medical Informatics and Decision Making*, 21(1), Article number 84. https://doi.org/10.1186/s12911-021-01451-8

Hassanpour, S., Bay, G. H., & Langlotz, C. P. (2017). Characterization of change and significance for clinical findings in radiology reports through natural language processing. *Journal of Digital Imaging*, 30(3), 314–322. https://doi.org/10.1007/s10278-016-9931-8

Liu, G., Hsu, T.H., McDermott, M.B.A., Boag, W., Weng, W., Szolovits, P., & Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. In *Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR* (pp. 249–269). https://proceedings.mlr.press/v106/liu19a.html

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations, ICLR 2013*. arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

Nakamura, Y., Hanaoka, S., Nomura, Y., Nakao, T., Miki, S., Watadani, T., Yoshikawa, T., Hayashi, N., & Abe, O. (2021). Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. *BMC Medical Informatics and Decision Making*, 21(1), Article number 262. https://doi.org/10.1186/s12911-021-01623-6

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Pons, E., Braun, L., Hunink, M. G. M., & Kors, J. A. (2016). Natural Language Processing in Radiology: A Systematic review. *Radiology*, 279(2), 329–343. https://doi.org/10.1148/radiol.16142770

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

Sammut, C., & Webb, G.I. (2011). TF–IDF. In: Sammut, C., Webb, G.I. (eds) *Encyclopedia of Machine Learning* (pp. 986–987). Springer. https://doi.org/10.1007/978-0-387-30164-8_832

Shetty, S., Ananthanarayana, V.S., Mahale, A. (2020). Medical knowledge-based deep learning framework for disease prediction on unstructured radiology free-text reports under low data condition. In *Proceedings of the 21st EANN (Engineering Applications of Neural Networks) 2020 Conference,* (pp. 352–364). Springer. https://doi.org/10.1007/978-3-030-48791-1_27

Shetty, S., Ananthanarayana, V. S., & Mahale, A. (2022). Comprehensive Review of Multimodal Medical data Analysis: open issues and future research Directions. *Acta Informatica Pragensia*, 11(3), 423–457. https://doi.org/10.18267/j.aip.202

Shetty, S., Ananthanarayana, V. S., & Mahale, A. (2023). Multimodal medical tensor fusion network-based DL framework for abnormality prediction from the radiology CXRs and clinical text reports. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-14940-x

Shin, B., Chokshi, F.H., Lee, T., & Choi, J. D. (2017). Classification of radiology reports using neural attention models. *arXiv*:1708.06828. https://doi.org/10.48550/arXiv.1708.06828

Sippo, D. A., Warden, G. I., Andriole, K. P., Lacson, R., Ikuta, I., Birdwell, R. L., & Khorasani, R. (2013). Automated Extraction of BI-RADS Final Assessment Categories from Radiology Reports with Natural Language Processing. *Journal of Digital Imaging*, 26(5), 989–994. https://doi.org/10.1007/s10278-013-9616-5

Sivic, J., & Zisserman, A. (2009). Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4), 591–606. https://doi.org/10.1109/tpami.2008.111

Wang, Q., Xu, J., Chen, H., & He, B. (2017). Two improved continuous bag-of-word models. In *2017 International Joint Conference on Neural Networks (IJCNN).* (pp. 2851–2856). IEEE. https://doi.org/10.1109/IJCNN.2017.7966208

Zhang, Y., Ding, D. Y., Qian, T., Manning, C. D., & Langlotz, C. P. (2018). Learning to summarize radiology findings. In *EMNLP 2018 Workshop on Health Text Mining and Information Analysis*. ACL. https://aclanthology.org/W18-5623.pdf

Editorial record: The article has been peer-reviewed. First submission received on 27 February 2023. Revisions received on 6 April 2023 and 13 April 2023. Accepted for publication on 13 April 2023. The editor in charge of coordinating the peer-review of this manuscript and approving it for publication was Stanislav Vojir.

Acta Informatica Pragensia is published by Prague University of Economics and Business, Czech Republic.

ISSN: 1805-4951