Interpretation of Mammogram and Chest X-Ray Reports Using Deep Neural Networks - Preliminary Results

Hojjat Salehinejad, Shahrokh Valaee, Senior Member, IEEE, Aren Mnatzakanian, Tim Dowdell, Joseph Barfett, and Errol Colak

Abstract—Radiology reports are an important means of communication between radiologists and other physicians. These reports express a radiologist’s interpretation of a medical imaging examination and are critical in establishing a diagnosis and formulating a treatment plan. In this paper, we propose a bi-directional convolutional neural network (Bi-CNN) model for the interpretation and classification of mammograms based on breast density and chest radiographic radiology reports based on the basis of chest pathology. The proposed approach helps to organize databases of radiology reports, retrieve them expeditiously, and evaluate the radiology report that could be used in an auditing system to decrease incorrect diagnoses. Our study revealed that the proposed Bi-CNN outperforms the random forest and the support vector machine methods.

Index Terms—Breast density, chest radiograph, convolutional neural networks, mammography, radiology reports classification.

I. INTRODUCTION

Radiology reports are an important means of communication between radiologists and other physicians [1]. These reports express a radiologist’s interpretation of a medical imaging examination and are critical in establishing a diagnosis and formulating a treatment plan.

Radiology is among the medical specialties with the highest rate of malpractice claims [2]. These claims can arise from a failure to communicate important findings, a failure of perception, lack of knowledge, and misjudgment. A failure to detect an abnormality on a medical imaging examination can lead to significant medical consequences for patients such as a delayed diagnosis. With a delayed or incorrect diagnosis, patients can present later with worsening symptoms and more advanced disease that may require more aggressive treatment or may be untreatable.

In this paper, we propose an auditing system for radiologists that has two main components: a natural language processing (NLP) model to process and interpret a radiology report and a machine vision model that interprets the medical imaging examination. This auditing system reviews the radiologist’s report and compares it with the interpretation of a machine vision model. The proposed system would notify the radiologist if there is a discrepancy between the two interpretations.

Many investigators have aimed to develop machine vision models for the interpretation of medical imaging examinations, such as chest radiographs [3], mammograms [4], and head computerized tomography (CT) scans [5]. However, fewer attempts have been made for the NLP of radiology reports [6, 7, 8]. The focus of this paper is on the design and performance evaluation of the NLP component.

We propose a bi-directional convolutional neural network (Bi-CNN) for the NLP of radiology reports. The Bi-CNN will be trained independently for two distinct report data types: breast mammograms and chest radiographs. In particular, this model will classify the degree of breast density and the type of thoracic pathology based on the radiology report content. Our proposed NLP model differs from a keyword search algorithm. It is capable of interpreting and classifying a radiology report. The proposed Bi-CNN has two independent input channels, where the order of non-padded input to one channel is the reverse of the non-padded input to the other channel. Performance of the proposed model is compared with a single channel convolutional neural network (CNN), random forest (RF), and support vector machine (SVM) models and a comparative study is conducted. We will share our datasets of anonymized patient radiology reports, labeled by radiologists, to encourage machine learning research in medicine.

II. RELATED WORK

The NLP models used for the classification and interpretation of radiology reports can be categorized into rule-based and machine learning models. Rule-based systems are usually built upon a series of “if-then” rules, which require an exhaustive search through documents. Such rules are typically designed by human experts. By contrast machine learning models do not require explicit rules. Instead, they try to train themselves iteratively and extract features from data.

A. Rule-based Approaches

A radiology report mining system generally has three main components. These include a medical finding extractor, a report and image retriever, and a text-assisted image feature extractor [9]. For example, a set of hand-crafted semantic rules are presented in [9] for mining brain CT radiology reports. A natural language parser for medical reports is proposed in [10] that uses four-gram and higher order systems to assign a stability metric of a reference word within a
2D Input Data

32x32 14x14x64

vector space representation of words. It uses global matrix
in this space [20]. The GloVe approach provides a global
Vectors with closer contexts are positioned in closer proximity
model assigns a word vector to each unique word in the corpus.
from a large corpus of text with reduced dimensionality. This
space representations of words. It produces a vector space
maximum entropy, and support vector machines [19].
being one example, using tools such as Naive Bayes classifiers,
for understanding medical reports in other languages, Korean
the report. Machine learning models have also been developed
method distinguishes the words that have a major impact in
generates less noise and is more stable. The max-pooling
paring with the window-based neural networks, this model
design to extract features from word representations. Com-
annotations [17]. The model proposed in [18] uses a recurrent
is capable of learning long term dependencies between the
annotations [17]. The model proposed in [18] uses a recurrent
design to extract features from word representations. Com-
paring with the window-based neural networks, this model
creates less noise and is more stable. The max-pooling
method distinguishes the words that have a major impact in
the report. Machine learning models have also been developed
for understanding medical reports in other languages, Korean
being one example, using tools such as Naive Bayes classifiers,
maximum entropy, and support vector machines [19].

C. Learning Vector Representations of Words

Word2vec is a well-known model used for learning vector
space representations of words. It produces a vector space
from a large corpus of text with reduced dimensionality. This
model assigns a word vector to each unique word in the corpus.
Vectors with closer contexts are positioned in closer proximity
in this space [20]. The GloVe approach provides a global
vector space representation of words. It uses global matrix
factorization and local context windowing methods to con-
struct a global log-bilinear regression model [21]. The word-
word co-occurrence matrix of the words is a sparse matrix by
nature. However, GloVe only uses the non-zero elements for
training and does not consider individual context windows in a
large corpus [21]. CNNs have been shown to work well on n-
gram representations of data [22]. For example, a convolution
layer can perform feature extraction from various n values of a
n-gram model and perform personality detection from
documents [22]. Dynamic k-Max pooling can operate as a
global operation over linear sequences. Such dynamic CNNs
can perform semantic modelling of sentences [23]. The model
receives variable dimension size input sentences and induces
a feature graph over the sentence that is capable of explicitly
capturing short and long-range relations [23].

CNN models are also utilized for learning representations,
opinion sentiment understanding, and analysis of products.
For example, a CNN with a single output softmax layer can
classify a vector representation of text with high accuracy [24].
The design of output layers in CNNs varies depending on the
application of the model. A collaboration of RNNs and CNNs
for feature extraction of sentences has been examined, where a
CNN learns features from an input sentence and then a gated
RNN model discourses the information [25]. In such system, a
bi-directional long short term memory (Bi-LSTM) RNNs can
sequentially learn words from question-answer sentences. The
trained network can select an answer sentence for a question
and present the corresponding likelihood for correctness [26].

III. A BRIEF REVIEW ON CONVOLUTIONAL NEURAL
NETWORKS

As NLP for radiology reports is an interdisciplinary research
effort, a brief review on CNN is provided in this section. A
CNN is a machine learning model inspired by the visual cortex
of cats [27]. A CNN has at least one layer of convolution
and sub-sampling, in which the number of layers can increase
in depth, generally followed by a fully connected multi-
layer perceptron (MLP) network. A consecutive arrangement
of convolutional layers followed by sub-sampling build a
pyramid-shape model, where the number of feature maps
increases as the spatial resolution decreases. Instead of hand-
designing feature extractors, the convolutional layers extract
features from raw data and the MLP network classifies the
features [28]. A CNN typically has three main pieces which
are local receptive fields, shared weights, and spatial and/or
temporal sub-sampling.
A. Local Receptive Fields

Local receptive fields are made from artificial neurons, which observe and extract features from data such as edges in images. Let us consider an image as input to a CNN such as in Figure 1. A rectangular kernel with size $M \times N$ scans the input matrix at every single element (i.e., a pixel) and performs convolution such as

$$h_{u,v} = \sigma \left( \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (I_{u+m,v+n} \cdot w_{m,n}) + b_{u,v} \right)$$

where the output is the state of the neurons at element $I_{u,v}$, $\sigma(\cdot)$ is a non-linear function, $w$ is the shared weights matrix, and $b$ is the bias term.

B. Shared Weights

The term “shared weights” refers to the weight matrix $w$ which appears repeatedly in the convolution operation for every element of $I$. Weight sharing reduces the number of free parameters of the model and hence improves generalization of the model [28].

For a general machine vision task, such as classification of natural images, the convolution kernel uses the correlation among spatial or temporal elements of the image to extract local features. Since the image is stationary (i.e., statistics of one part of the image are similar to any other part), the learned shared weights using one sub-region of the image can also be used at another sub-region to extract different features.

C. Sub-sampling

Passing the whole bag of extracted features after the convolution operation to a classifier is computationally expensive [29]. The feature pooling task (i.e. sub-sampling) generalizes the network by reducing the resolution of the dimensionality of intermediate representations (i.e. feature maps) as well as the sensitivity of the output to shifts and distortions [28]. The two most popular subsampling methods are mean-pooling and max-pooling. By dividing the feature map into a number of non-overlapping rectangle-shape sub-regions, the mean-pooling computes the average of features sitting inside a sub-region as the pooled feature of that sub-region. Max-pooling performs the same task as mean-pooling but with a maximum operator. A study on these two operations is provided in [29].

IV. The Proposed Method

An auditing system for radiology is presented in Figure 2. The main components of the system are a machine vision model to interpret the input image and a NLP component to interpret the corresponding radiology report. The system notifies the radiologist if there is a discrepancy between the two components of the system. In this paper, our focus is on the design and evaluation of the NLP component.

A. Preprocessing

In general, radiology reports include major sections such as “Indication”, “Findings”, “Impression”, and the name of the reporting radiologist. However, there are variations in report formatting due to “radiologist style” or institutional requirements. Despite brute-force search approaches, the deep learning models have the advantage of requiring the least amount of data cleaning and preprocessing due to their natural adaptation to the data and non-linear feature extraction ability.

In the preprocessing step, the model extracts the “Findings” and “Impression” sections of the report and performs tokenization and string cleaning. Some of the major tasks include converting upper-case characters to lower-case, removing unnecessary punctuation and symbols, and separating the remainders from the attached string. We add every unique word to a vocabulary dictionary. Each vocabulary has a unique associated index which represents it in the sentence vector. For example, given the sentence “The breasts show scattered fibroglandular tissue. There are scattered fibroglandular densities in breasts”, the entire vocabulary dictionary would consist of 10 words (breasts, show, scattered, fibroglandular, tissue, densities, the, there, are, in), which is represented as a list of integers such as $[1, ..., 10]$. A vector to the length of the longest report in terms of words count in the dataset represents the report. Padding fills up the gap for shorter reports.

In the character embedding step, each integer is mapped to a high dimensional (i.e., $D$) vector with a uniform random distribution. In such a high dimensional space, due to the “curse of dimensionality”, the vectors are considered independent [30]. Character embedding vector generation for medical vocabularies based on the correlation among the vectors (i.e., similar to Word2Vec) is an interesting topic for further investigation.

B. Bi-CNN Model Architecture

An input report with $N$ words can be represented as a sequence $X = [x_1, ..., x_N]$, where each word is a vector
Fig. 3: Proposed bi-directional convolutional neural network for the interpretation and classification of mammograms based on breast density. For the chest reports case, the labels of the output layer change accordingly but the model architecture is identical.

\[
\begin{align*}
x_n & \in \mathbb{R}^D, & \text{[32]} \quad \text{As Figure 3 shows, the proposed architecture has two input channels followed by two independent convolution layers. Both channels have an identical design, except the order of non-padded input to one channel is the reverse of the non-padded input to the other channel. In the example in Figure 4, the dependencies between feature vectors (x_3, x_4, x_5) are visible to the kernels in both channels. However, the dependencies are not visible to the kernels in Channel 1. Other examples are the feature vectors (x_0, 0, 0) in Channel 2 and feature vectors (x_4, x_5, 0) and (x_5, 0, 0) in Channel 1, which are not visible in the corresponding other channel. Each channel in the proposed model in Figure 3 has three filters with varying window sizes } K \in \{3, 4, 5\} \text{ that slide across the input layer } \mathbb{R}^D. \text{ The filters extract features from the input layer to construct feature maps of size } (N-K+1) \times F, \text{ where } F \text{ is the number of feature maps for each filter. Each feature map } h_{N-K+1} \text{ has its own shared weight } W_{K \times D} \text{ and bias } b_{N-K+1}. \text{ The value of a hidden neuron } m \text{ is }
\end{align*}
\]

\[
h_m = \sigma(\sum_{k=1}^{K} \sum_{d=1}^{D} x_{m+k-1,d} \cdot w_{k,d} + b_m). \quad (2)
\]

The window is } x_{m+m+K-1} \text{ and } \sigma(\cdot) \text{ is a rectified linear unit (ReLU) activation function defined as }

\[
\sigma(z) = \max(0, z) \quad (3)
\]

where } z \in \mathbb{R}. \text{ The max pooling } [33] \text{ extracts the most significant feature from the feature map of a filter as }

\[
\hat{h}_l = \max\{h_1, \ldots, h_{N-K+1}\}. \quad (4)
\]

The output of the max-pooling layer contains the max-pooled features } h = [h_1, \ldots, h_L] \text{ where the length of the feature vectors is } L = C \cdot F \text{ and } C \text{ is the number of classes. For example, for the breast fat density classification with five classes we have } L = 5F \text{ (see Figure 3)}. \text{ The features from the input channels are concatenated and passed to a fully connected perceptron network. The output of the softmax layer is the probability distribution over all the labels } P \text{ (e.g., for mammograms the breast density classes). The value of output unit } p \text{ is }

\[
y_p = \phi(L \sum_{l=1}^{L} (\hat{h}_{l} \cdot r_{l,p}) \cdot w_{l,p} + b_p) \quad (5)
\]

where } w_{l,p}^{h_0} \text{ is the weight of connection from the hidden unit } l \text{ in the hidden layer } h \text{ to the output unit } p \text{ in the output layer } a, b_p \text{ is the bias of the output unit } p, \text{ and } \phi(z) \text{ is the softmax activation function defined as }

\[
\phi(z_p) = \frac{e^{z_p}}{\sum_{j=1}^{P} e^{z_j}} \quad \text{for } p = [1, \ldots, P]. \quad (6)
\]

C. Training

Adam optimizer is a first-order gradient-based optimization method with integrated momentum functionality [34]. During training, the momentum helps to diminish the fluctuations in weight changes over consecutive iterations. The drop-out regularization method randomly drops units along with their connections from the concatenated layer (i.e. the input layer to classifier) to the output layer using a binary } mask \text{ vector } r_{1 \times L} \text{ with Bernoulli random distribution [35].}
D. Evaluation Scheme

Since we are dealing with discrete categories, we use cross-entropy between a predicted value \( y^{(i)} \) from the network and the real label \( t^{(i)} \) to measure the loss of networks such as

\[
\hat{L}(Y, T) = -\frac{1}{R} \sum_{i=1}^{R} t^{(i)} \ln(y^{(i)}) + (1 - t^{(i)}) \ln(1 - y^{(i)})
\]

(7)

where \( R \) is the number of reports, \( Y \) is the set of network predictions, and \( T \) is the set of targets to predict. Adding the weight decay term (i.e., \( L_2 \) regularization) to the loss function helps the network to avoid over-fitting while training such as

\[
\mathcal{L}(Y, T) = \hat{L}(Y, T) + \eta \|W^{ho}\|_2
\]

(8)

where \( W^{ho} \) is the weight matrix of connections between the hidden layer \( h \) and the output layer \( o \) and \( \eta \) is the regularization control parameter.

| Dataset | NR | VS | ANS Mean | ANS Median | ANW Mean | ANW Median | ASL Mean | ASL Median |
|---------|----|----|----------|------------|----------|------------|----------|------------|
| MRD    | 4,080 | 1,091  | 1.21    | 1.20   | 30.39         | 27.00     | 10.44    | 9.80        | 6.22       |
| CXRD   | 1,030  | 772     | 2.01    | 2.00   | 1.07           | 1.46      | 9.21     | 8.21        | 7.60       |

Fig. 5: Four classes of breast fat density. a) Fatty breast tissue; b) Scattered density; c) Heterogeneously dense; d) Extremely dense.

Fig. 6: The chest radiographs for eleven categories. a) Normal; b) Cardiomegalic; c) Consolidation; d) Pulmonary Edema; e) Lung Nodules; f) Lung Mass; g) Pleural Effusion; h) Widened Mediastinum; i) Vertebra Fractures; j) Clavicular Fracture; k) Pneumothorax.

V. The Datasets

Our institutional review board approved this single-center retrospective study with a waiver for informed consent. A search of our Radiology Information System (RIS) (Syngo; Siemens Medical Solutions USA Inc, Malvern, PA) was performed for mammography and chest radiograph reports using Montage Search and Analytics (Montage Healthcare Solutions, Philadelphia, PA). This search identified 4,080 mammography reports and 1,030 chest radiographic reports (see Table I). These reports were exported from the RIS and removed of any patient identifying information.

A. Mammogram Reports Dataset

The mammography reports dataset (MRD) contained 4,080 reports. Breast density was classified in each report according to the American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) classification system [36] (see Figure [5]). Breast density reflects the relative composition of fat and fibro-glandular tissue. “Fatty breast tissue” refers to breast with less than 25% fibro-glandular tissue. Fibro-glandular tissue accounts for 25 – 50% of breasts with “scattered density”, 51 – 75% of “heterogeneously dense” breasts, and greater than 75% of breasts considered “extremely dense”. Descriptive statistics of this dataset are presented in Table II while sample reports are presented in Table III.

B. Chest Radiographs Reports Dataset

The chest radiograph (x-ray) dataset (CXRD) consisted of 1,030 reports. These reports were classified into 11 categories which includes normal examinations and ten pathologic states (see Figure [6]). Descriptive statistics of this dataset are presented in Table IV while sample reports are presented in Table V. This dataset has more categories, fewer reports per category, and is less imbalanced when compared with the MRD.

VI. Experiments and Results Analysis

In this section, we compare performance of the proposed Bi-CNN model with the random forest (RF), support vector machine (SVM), and CNN models. The experiments were conducted on both the mammogram and chest radiograph datasets.

The CNN and Bi-CNN models are implemented in TensorFlow [37] and the RF and SVM models are implemented using the classifiers in scikit-learn [38]. The experiments are conducted on a DevBox with an Intel Core i7-5930K 6 Core 3.5GHz desktop processor, 64 GB DDR4 RAM, and two TITAN X GPUs with 12GB of memory per GPU.
TABLE II: A general consensus for the breast density descriptors for five categories of breast density.

| Category                | Label   | Count | Fat density level | Some General Consensus for the Breast Density Descriptors                                                                 |
|-------------------------|---------|-------|-------------------|-------------------------------------------------------------------------------------------------------------------------|
| Fatty breast tissue     | 0       | 342   | <25%              | “fat”, “fatty” - includes variations of “mainly fatty”, “predominantly fatty”, “predominantly fat”                       |
| Scattered density       | 1       | 1546  | <50%              | “scattered”                                                                                                             |
| Heterogeneously dense   | 2       | 1510  | >50%              | “heterogenous” and “heterogeneously”                                                                                   |
| Extremely dense         | 3       | 436   | >75%              | “dense”, “extremely dense”, “very dense”                                                                               |
| Uncertain               | 4       | 246   |                   | “mild dense”, “mildly dense” - overlap in describing categories with label 1 and 2                                        |

TABLE III: A sample of radiology reports with associated categories for mammogram reports dataset (MRD).

Sample from MRD                                                                 Category
The breast parenchyma is heterogeneously dense. No suspicious masses, calcifications or architectural distortion seen on either side.  
The breasts show scattered fibroglandular densities. There are no dominant nodules or suspicious calcification seen in either breast.  
The breast parenchyma is extremely dense. No suspicious masses, calcifications or architectural distortion seen on either side.  
Postoperative changes are again noted in the left breast. The breasts are almost entirely fatty bilaterally.  
No suspicious calcification, dominant mass or architectural distortion. No interval change.  

TABLE IV: A general consensus for the chest radiographs descriptors for 11 categories.

| Category              | Label | Count | Some General Consensus for the Chest Radiographs Descriptors |
|-----------------------|-------|-------|-------------------------------------------------------------|
| Normal                | 0     | 116   | no acute findings; unremarkable study; the cardiopericardial silhouette and hilar anatomy is within normal limits |
| Cardiomegaly          | 1     | 106   | mild cardiomegaly/pericardial effusion; enlarged cardiac silhouette; cardiopericardial silhouette is enlarged |
| Consolidation         | 2     | 81    | mild cardiomegaly/pericardial effusion; enlarged cardiac silhouette; cardiopericardial silhouette is enlarged |
| Pulmonary Edema       | 3     | 104   | mild to moderate interstitial pulmonary edema; airspace edema; right pulmonary edema |
| Lung Nodules          | 4     | 118   | calcified granulomas present; nodules have developed; multiple faint bilateral pulmonary nodules |
| Lung Mass             | 5     | 89    | bilateral pulmonary masses; mass in the right upper lobe; middle mediastinal mass |
| Pleural Effusion      | 6     | 144   | persistent bilateral pleural effusions; bilateral pleural effusions; persistent loculated right pleural effusion |
| Widened Mediastinum   | 7     | 50    | widening of the superior mediastinum; mediastinum appears widened; mediastinum is slightly widened |
| Vertebral Fractures   | 8     | 67    | vertebral fracture; several old vertebral compression injuries; thoracoolumbar vertebral compression injuries |
| Clavicular Fracture   | 9     | 73    | plated fracture of the right clavicle; fracture deformity of the right lateral clavicle; right lateral clavicle fracture |
| Pneumothorax          | 10    | 82    | partial right upper lobe collapse; chronic collapse of the right middle lobe; right apical pneumothorax |

TABLE V: A sample of radiology reports with associated categories for chest radiographs dataset (CXRD).

Sample from CXRD                                                                 Category
Cardiome-diastinal contours are within normal limits. The hila are unremarkable. The lungs are clear.  
No pleural effusion or pneumothorax. The osseous structures are within normal limits. No displaced fracture apparent.  
The lungs remain overinflated and show mild chronic parenchymal changes.  
Inhomogeneous airspace consolidation has developed in the basal segments of the left lower lobe. The pleural spaces are clear.  
The left pulmonary nodular density superimposed on the posterior seventh rib is smaller. There is mild stable bilateral upper lobe pleuropulmonary scarring and a stable right upper lobe nodule. The remainder of the lungs is clear.  
The lungs are overinflated but clear.  
The pleural spaces, mediastinum and diaphragm appear normal.  

A. Parameters Setting

Unless stated, the CNN and Bi-CNN models are trained with a mini-batch size of 64, drop-out probability of 0.5, filter sizes {3, 4, 5} and 120 feature maps per filter size. The number of training iterations is set to 50 and the initial learning rate for Adam optimizer is set to 0.001. An exponential decay adaptive learning rate is applied. The weights at output layers are initialized using the Xavier method, the weights in the convolutional layer are selected based on normal distribution with standard deviation of 0.1, and biases are set to 0.1. The activation function before the max-pooling layer is ReLU. The $L_2$ regularization is set to $1.0 \times 10^{-4}$ and early-stopping is applied.

For all the experiments, 70%, 15%, and 15% of data is allocated for training, validation, and test, respectively. The data is shuffled before splitting. In each experiment, the model is cross-validated over 30 independent experiments and the results after statistical testing are reported.

B. Performance Evaluation and Analysis

We experimented with various configurations of the RF [41], SVM [42], and CNN [24] models.
| Model | MRD | CXRD | MRD | CXRD |
|-------|-----|------|-----|------|
| RF (1-gram) | 69.73 | 67.32 | 69.73 | 67.32 |
| RF (2-gram) | 73.76 | 69.27 | 73.76 | 69.27 |
| RF (3-gram) | 80.53 | 77.59 | 80.53 | 77.59 |
| SVM (sigmoid) | 82.74 | 80.24 | 82.74 | 80.24 |
| SVM (poly) | 82.97 | 81.53 | 82.97 | 81.53 |
| CNN (1-kernel) | 85.72 | 85.27 | 85.72 | 85.27 |
| CNN (2-kernel) | 86.52 | 86.01 | 86.52 | 86.01 |
| CNN (3-kernel) | 86.91 | 86.05 | 86.91 | 86.05 |
| Bi-CNN (1-kernel) | 89.60 | 89.16 | 89.60 | 89.16 |
| Bi-CNN (2-kernel) | 90.88 | 89.91 | 90.88 | 89.91 |
| Bi-CNN (3-kernel) | 92.94 | 91.34 | 92.94 | 91.34 |

- **RF**: Instead of using a bag-of-words technique, which uses the frequency of the words in a query as the features, a n-gram model with a lower boundary of 1 and upper boundary \( n \in \{1, 2, 3\} \) performs the feature extraction of reports. We use an implementation of RF [41] which combines classifiers by averaging their probabilistic predictions. The number of estimators is set to 10.

- **SVM**: Similar to RF, the SVM model uses n-gram feature extraction. The SVM model is deployed for two different kernels, a “sigmoid” and a “poly” with degree three.

- **CNN**: The CNN implementation proposed in [24] for text classification and later used in [8] for the classification of radiology head CT reports is used. This model has a single input channel and the order of input words is similar to the report. The model has a convolution layer followed by a max-pooling layer and a “softmax” classifier. The studies are for kernel sizes \( k \in \{1, 2, 3\} \).

- **Bi-CNN**: The proposed Bi-CNN with two input channels. The settings are similar to CNN.

The experiments are for breast fat density classification and chest radiograph classifications based on basis of chest pathology using the MRD and CXRD, respectively. The results of performance comparisons between the RF, SVM, CNN, and Bi-CNN models are presented in Table [VI]. As \( n \) increases in n-gram for RF, it considers the dependency between a greater number of words (i.e., \( n \)). The RF with 3-gram model has better performance than the 2-gram and 1-gram RF models. The SVM with polynomial of degree three (i.e., “poly”) and “sigmoid” kernels almost have the same performance. The number of kernels in the CNN models have a minor impact on accuracy. However, the CNN models with 3-kernel have better performance than models with 1-kernel and 2-kernel. However, the proposed Bi-CNN with 3-kernel has the best performance compared to the other models.

The convergence of CNN and Bi-CNN for the validation of datasets is presented in Figure [7]. For large values of the learning rate (i.e., \( \lambda = 0.1 \)), the models converge rapidly to a local solution. However, for smaller learning rates, the models converge more slowly but with greater stability towards a better solution. As the results in Table [VII] show, the CNN model converges with a lower accuracy for three different learning rates \( \lambda \in \{0.1, 0.01, 0.001\} \). For \( \lambda = 0.001 \), the best performance is achieved by Bi-CNN for both MRD and CXRD. One of the advantages of Bi-CNN is adding diversity into the model through the integration of two different representations of feature vectors into the model. This diversity helps the CNN avoid convergence to local solutions with low accuracy.

**VII. Conclusion**

Radiology reports contain a radiologist’s interpretation of a medical imaging examination. Vast numbers of such reports are digitally stored in health care facilities. Text mining and knowledge extraction from such data may potentially be used for double checking radiologist interpretations as part of an audit system. For instance, these methods can be used to audit utilization of limited health care resources and to enhance the study of patient treatment plans over time.

In this paper, we have proposed a Bi-CNN for the interpretation of radiology reports. We have collected and anonymized two datasets for our experiments: a mammogram reports dataset and a chest radiograph reports dataset. We intend to release these anonymized datasets for the broader research community to use. The Bi-CNN has two input channels where one channel represents the features of the report (including zero-paddings) and the other channel represents the reverse non-padded features of the report. The combination of forward and reverse order of feature vectors represents more information about dependencies between feature vectors. Our comparative studies demonstrate that the Bi-CNN model outperforms the CNN, RF, and SVM methods.

**Acknowledgement**

The authors thank the support of NVIDIA Corporation with the donation of the Titan X GPUs used for this research.

**References**

[1] “The royal children’s hospital of melbourne.” [Online]. Available: http://www.rch.org.au/.

[2] A. Pinto and L. Brunese, “Spectrum of diagnostic errors in radiology,” *World Journal of Radiology*, vol. 2, no. 10, pp. 377–383, 2010.

[3] M. Cicero, A. Bilbily, E. Colak, T. Dowdell, B. Gray, K. Perampaladas, and J. Barfett, “Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs,” *Investigative radiology*, vol. 52, no. 5, pp. 281–287, 2017.

[4] J. Wang, H. Ding, F. Azamian, B. Zhou, C. Iribarren, S. Molloi, and P. Baldi, “Detecting cardiovascular disease from mammograms with deep learning,” *IEEE transactions on medical imaging*, 2017.

[5] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, “Brain tumor segmentation with deep neural networks,” *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[6] D. B. Aronow, F. Fangfang, and W. B. Croft, “Ad hoc classification of radiology reports,” *Journal of the American Medical Informatics Association*, vol. 6, no. 5, pp. 393–411, 1999.

[7] A. Wilcox, “Automated classification of medical text reports,” Ph.D. dissertation, Columbia University New York, 2000.
[8] B. Shin, F. H. Chokshi, T. Lee, and J. D. Choi, “Classification of radiology reports using neural attention models,” in Neural Networks (IJCNN), 2017 International Joint Conference on. IEEE, 2017, pp. 4363–4370.

[9] T. Gong, C. L. Tan, T. Y. Leong, C. K. Lee, B. C. Pang, C. T. Lim, Q. Tian, S. Tang, and Z. Zhang, “Text mining in radiology reports,” in Data Mining, 2005. ICDM’05. Eighth IEEE International Conference on. IEEE, 2008, pp. 815–820.

[10] R. K. Taira, V. Bashyam, and H. Kangarloo, “A field theoretical approach to medical natural language processing,” IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 4, pp. 364–375, 2007.

[11] G. F. Cooper, G. F. Cooper, P. Hanbury, B. E. Chapman, L. H. Harrison, and M. M. Wagner, “Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders,” Journal of the American Medical Informatics Association, vol. 10, no. 5, pp. 494–503, 2003.

[12] P. A. Dang, M. K. Kalra, M. A. Blake, T. J. Schultz, M. Stout, E. F. Halpem, and K. J. Dreyer, “Use of radcube for extraction of finding trends in a large radiology practice,” Journal of digital imaging, vol. 22, no. 6, pp. 629–629, 2009.

[13] G. Zuccon, A. S. Wagholikar, A. N. Nguyen, L. Butt, K. Chu, S. Martin, and J. Greenslade, “Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology,” AMIA Summits on Translational Science Proceedings, vol. 2013, p. 300, 2013.

[14] H.-C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers, “Interleaved text/image deep mining on a very large-scale radiology database,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1090–1099.

[15] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, “Learning to read chest x-rays: recurrent neural cascade model for automated image annotation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2497–2506.

[16] H. Salehinejad, “Learning over long time lags,” arXiv preprint arXiv:1602.04335, 2016.

[17] S. Tai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” in AAAI, vol. 333, 2015, pp. 2267–2273.

[18] H.-S. Oh, J.-B. Kim, and S.-H. Myaeng, “Extracting targets and attributes of medical findings from radiology reports in a mixture of languages,” in Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM, 2011, pp. 550–552.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in neural information processing systems, 2013, pp. 3111–3119.

[20] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation.” in EMNLP, vol. 14, 2014, pp. 1532–1543.

[21] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, “Deep learning-based document modeling for personality detection from text,” IEEE Intelligent Systems, vol. 32, no. 2, pp. 74–79, 2017.

[22] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” arXiv preprint arXiv:1404.2188, 2014.

[23] Y. Kim, “Convolutional neural networks for sentence classification,” arXiv preprint arXiv:1408.5882, 2014.

[24] Y. Ren and D. Ji, “Neural networks for deceptive opinion spam detection: An empirical study,” Information Sciences, vol. 385, pp. 213–224, 2017.

[25] D. Wang and E. Nyberg, “A long short-term memory model for answer sentence selection in question answering,” in ACL, 2015.

[26] D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” The Journal of physiology, vol. 195, no. 1, pp. 215–243, 1968.

[27] Y. LeCun, Y. Bengio et al., “Convolutional networks for images, speech, and time series,” The handbook of brain theory and neural networks, vol. 3361, no. 10, p. 1995, 1995.

[28] Y.-L. Bourane, J. Ponce, and Y. LeCun, “A theoretical analysis of feature pooling in visual recognition,” in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 111–118.

[29] R. Weber, H.-J. Schek, and S. Blott, “A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces,” in VLDB, vol. 98, 1998, p. 194–205.

[30] P. Domingos, “A few useful things to know about machine learning,” Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012.

[31] H. Salehinejad, J. Barfett, P. Aarabi, S. Valae, E. Colak, B. Gray, and T. Dowdell, “A convolutional neural network for search term detection-a draft.” arXiv preprint arXiv:1708.02238, 2017.

[32] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” Journal of Machine Learning Research, vol. 12, no. Aug, pp. 2493–2537, 2011.

[33] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[34] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] R. TAo, “American college of radiology breast imaging reporting and data system bi-rads,” Reston, VA, 1998.

[36] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin et al., “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” arXiv preprint arXiv:1603.04467, 2016.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., “Scikit-learn: Machine learning in python,” Journal of Machine Learning Research, vol. 12, no. Oct, pp. 2825–2830, 2011.

[38] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.

[39] L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[40] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” IEEE Intelligent Systems and their applications, vol. 13, no. 4, pp. 18–28, 1998.