IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 February 2024, accepted 7 May 2024, date of publication 13 May 2024, date of current version 20 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3400007

RESEARCH ARTICLE

NLP-Powered Healthcare Insights: A Comparative Analysis for Multi-Labeling Classification With MIMIC-CXR Dataset

EGE ERBERK USLU[1], EMINE SEZER[1], AND ZEKERIYA ANIL GUVEN[2]
[1]Department of Computer Engineering, Faculty of Engineering, Ege University, Bornova, 35040 İzmir, Turkey
[2]Department of Computer Engineering, Faculty of Engineering and Architecture, İzmir Bakırçay University, Menemen, 35665 İzmir, Turkey
Corresponding author: Emine Sezer (emine.sezer@ege.edu.tr)

ABSTRACT The digitization of the healthcare industry has led to a growing number of applications that use machine learning and image processing techniques to improve the diagnostic process. These applications utilize a variety of medical data, including laboratory results, clinical findings, MRI scans, tomographic images, and radiological images. In addition, free-text healthcare documentation, such as well-structured discharge summaries, contains valuable information. Natural Language Processing encompasses the development of automated systems for generating health reports. This process involves using domain-specific knowledge and prior knowledge to extract relevant information from medical records. This article investigates the use of natural language processing techniques for chest X-ray classification. A total of 14 distinct impressions derived from chest radiography findings from the MIMIC-CXR dataset were used in a multi-label classification procedure. Six distinct language models derived from the BERT language model, along with three distinct classification algorithms, were employed to evaluate the effectiveness of the models and the dataset for multi-label categorization. The experimental results showed a successful prediction rate of 80.47% for 14 distinct impressions within the dataset.

INDEX TERMS BERT, chest radiology report, MIMIC-CXR, multi-label classification, natural language processing.

I. INTRODUCTION

Human health is a comprehensive state of well-being that encompasses both the physical and psychological dimensions of the individual. In this context, health services play a special role by facilitating the diagnosis, treatment, and prevention of both chronic and acute diseases and disabilities. An adequate healthcare system has the inherent capacity to increase the overall social welfare in a nation. Achieving this goal requires multifaceted collaboration between experienced healthcare professionals and additional staff to ensure comprehensive healthcare delivery.

It is worth emphasizing that inadequate health services can lead to permanent damage and even death in individuals and that health expenditures may increase due to inappropriate use of medical equipment and unnecessary personnel. Typically, individuals seek the expertise of healthcare professionals to find solutions to their health concerns. Interpretation of radiological images by radiologists is one of the basic medical practices used primarily to make a diagnosis and manage the treatment process. The radiology report, written in the native language, not only evaluates potential abnormalities and describes existing abnormalities, but also determines the extent or stage of chronic disease, compares current radiological imaging with historical data, and provides potential diagnoses and recommendations to guide therapeutic interventions.

The automated generation of patient reports and the extraction of information from health data for diagnostic purposes are crucial innovations that alleviate the workload and time constraints faced by healthcare specialists. In this context, Artificial Intelligence (AI) algorithms have emerged as indispensable tools, enabling the provision of rapid, accurate, and efficient healthcare services. The integration of traditional

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

health information systems with machine learning, semantic web technologies, the Internet of Things, and big data is gaining increasing traction. To address contemporary challenges, a vast array of scientific research and systematic studies encompassing machine learning, deep learning, and image processing methodologies are carried out.

These initiatives have made notable contributions to the advancement of disease diagnosis through radiological imaging and medical reporting. However, radiology image processing research faces challenges, particularly due to the limitations of two-dimensional images and section overlap [1], [2]. Additionally, clinicians value the radiologist's insights embedded in free-text radiology reports, which are integral to the diagnostic process.

The complexity of image features raises the question of whether textual radiological reports can improve diagnostic performance. Natural language processing (NLP), a machine learning approach to processing free text in information technologies, is used to help professionals discover, diagnose, and treat diseases by integrating them with other technologies. Large datasets are essential for developing and testing NLP and deep-learning AI models.

The MIMIC-CXR dataset, used in this study, is preferred because of its high volume, variety, and value. The dataset includes chest radiography images and textual patient histories, and has been anonymized according to the Health Insurance Portability and Accountability Act (HIPAA) and the Personal Health Information (PHI) guidelines [3].

This article presents an AI model trained using textual data extracted from the MIMIC-CXR dataset to predict 14 different chest radiological outcomes to help healthcare specialists in making medicinal decisions. Six different BERT-based language models were used in this study, including BERT [4], BioBERT [5], ClinicalBERT [6], [7], CXR-BERT generalized and specialized editions [8], and S-PubMedBert [9]. In addition, the weights of words from BERT-based language models were tested using three classification techniques: BERT classification, CNN-BERT classification, and BILSTM-BERT classification [10].

The article is structured as follows: the second section presents a literature review, followed by the third section, which describes the methodology used in this study. In the fourth section, the experiments and their results are given. Finally, conclusions and a discussion on future work is presented.

II. LITERATURE REVIEW

NLP is a field of computer science that emerged to understand and analyze the complex structure of human language [11]. The computational linguistics domain is so named because it focuses on analyzing and processing human language, including both written and spoken forms. It is aimed at developing language processing systems that exhibit human-like behavior while performing various tasks and applications [12]. NLP is an intermediary between human users and computer systems through natural languages. In summary, NLP is an academic discipline that investigates the use of computers to understand, process, and evaluate natural language texts and speech expressions [13].

The exponential growth of digital information in texts, including newspapers, websites, emails, social media posts, blogs, and more, is resulting in the production of massive amounts of data reaching millions of terabytes. Effectively managing and extracting insights from this vast data store requires efficient and robust natural language processing (NLP) techniques. NLP covers a wide range of methodologies, including text summarization, sentiment analysis, information extraction, named entity identification, association extraction, social media monitoring, text mining, language translation programs, and question-answering systems. Collectively, these methods facilitate comprehensive language analysis by enabling raw language to be converted into a structured, processable format. Given the inherent complexity of computer science and artificial intelligence, NLP relies on a complex understanding of words and their relationships to extract meaning. Using NLP approaches can significantly reduce the storage space and directory size for texts and documents while increasing the user's expectation of being able to find what they are looking for semantically. Additionally, NLP plays an important role in improving documentation and information retrieval processes [14].

One of the resources containing rich text and information is electronic health records (EHRs). As it is known, professionals document all details concerning the individual's condition in their native language. Thus, patient health reports are essential health information. Facilitating access to health information, especially EHRs, is expected to accelerate diagnosis and treatment, minimize labor and time costs, and improve healthcare [15].

Precise diagnosis is essential for effective treatment, particularly in diseases requiring rapid intervention. Radiology provides valuable insights for diagnosis, staging, treatment planning, and outcome prediction. However, unstructured radiology reports challenge large-scale studies requiring efficient information extraction. NLP offers a promising solution for transforming unstructured reports into a structured format, enabling automated information extraction and analysis. This technology demonstrates potential across various healthcare applications, including diagnostic surveillance, cohorting, quality assessment, and computer vision labeling [16].

The use of chest radiology is prevalent on a global scale as the primary modality for assessing the chest region in medical imaging. Chest radiographs are used in medical research to diagnose both acute and chronic cardiopulmonary disorders, as well as verify the accurate placement of various devices including pacemakers, central lines, chest tubes, and stomach tubes. Additionally, they are utilized to detect both acute and chronic cardiopulmonary problems [17].

The MIMIC-CXR dataset, a publicly available collection of chest X-ray images and reports, has been available since 2019 and has been cited in over 300 publications. In one of the studies conducted using the dataset, NLP was

shown to increase the accuracy of pneumothorax diagnosis [18]. Deep learning has also been used to study multi-label chest X-ray abnormality taxonomies hierarchically [19]. Sidorov et al. [20] developed a method to summarize important information in the results section of radiology reports, improving radiologists' communication with referring physicians. An abstractive summary system for chest radiography report impressions has also been developed and tested, resulting in a significant reduction in radiologists' workload [21].

The CheXpert dataset [22] functions as a standardized measure for the automated interpretation of chest X-rays. Its primary objective is to assess the likelihood of 14 observations derived from radiographs with multiple images. To facilitate the process of labeling, rules-based taggers have been developed in conjunction with the dataset, enabling labelers to accurately identify positive, negative, and ambiguous classifications. CheXpert's annotation process incorporates multiple approaches. DNorm and MetaMap are used to replace automatic word inference, thereby enhancing the precision of the annotations. The Mention Extraction segment examines the statements made by radiologists, while the universal dependency parsing of the report guides categorizing mentions. Sentence segmentation is accomplished through the utilization of the Bllip parser, which is trained on David McClosky's biological model, with the Universal Dependency Graph being computed by Stanford CoreNLP [23]. Uncertainty labeling in CheXpert is carried out through three distinct methods. Visual estimation is used on a scale of 0 to 1, with a value of 1 being assigned to indicate uncertainty. Additionally, a model is utilized to assist with labeling. These methods have been applied to annotate both the ChestX-ray14 and MIMIC-CXR datasets.

Building on the established ability of Convolutional Neural Networks (CNNs) to capture semantic meaning and classify phrases in clinical text, a CNN-based method for sentence-level medical document classification is proposed [24]. This approach aims to identify emergent semantics from medical literature, focusing on 26 categories within Brain and Cancer research. The study utilizes Word2vec embeddings trained on domain-specific data to represent 4,000 sentences. The CNN architecture uses 256 convolutional filters with a consistent size of 5 across all layers. To mitigate overfitting, dropout functions are implemented after the second max pooling layer (0.5 dropout rate) and the fully-connected layer (128 units). A final deep 26-dimensional layer serves to represent the categorization categories. SoftMax activation is used to calculate the output. The CNN-based approach is compared to Sentence Embeddings, Mean Word Embeddings, and Word Embeddings with Bag-of-Words (BOW). The results demonstrate that the CNN strategy outperforms other methods for classification tasks, achieving a minimum 15% accuracy margin.

The deep learning architecture, with Bidirectional Long-Short-Term Memory (LSTM) layers, highlights the potential of Deep Learning (DL) and Word Embeddings in identifying sixteen types of morbidity in healthcare data [25]. This approach enables the utilization of advanced vector data forms like Word Embeddings. The study comprehensively compared DL against TF-IDF with pre-trained Word Embeddings, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) models (GloVe and Word2Vec). The results show that DL surpasses all other methods, regardless of the word embedding used. Moreover, the research also concludes that TF-IDF with SVM outperforms Word Embeddings, attributed to the latter's potential for capturing domain-specific idiosyncrasies.

A text-and-subject feature fusion model improved Chinese medical health query categorization [26]. First, word embedding generates text word vectors. Then, the LSTM model extracts text features. The optimal number of topics is subsampled to optimize classification performance. This work uses one-dimensional convolution to extract and reduce the dimensionality of topic features. Both text and topic features are integrated for text classification. The model's performance was evaluated using two datasets from different online medical Q&A platforms, demonstrating improved recall, accuracy, and F1 value for Chinese medical health query categorization.

Safaya et al. [10] propose a model with two parts: a BERT encoder using 12 self-attention layers to contextualize input text, and a CNN classifier. BERT embeds the content with 64-token inputs. The proposed method then uses the output of the last four BERT layers as channels for each convolution kernel, leveraging their contextual information for better feature extraction. ReLU activation and global max pooling handle the result, which is flattened and fed into a dense layer with Sigmoid activation for binary classification. This model, trained for 10 epochs, outperforms SVM, CNN-Text, and BERT on the development set, achieving the highest macro-averaged F1-Score.

III. CLASSIFICATION METHODOLOGY

The present study investigates the utilization of NLP techniques for chest X-ray classification. Classification is a fundamental machine learning task that involves using a pre-categorized training dataset to categorize new instances. Classification algorithms acquire knowledge of the distribution pattern from the provided training set and subsequently endeavor to accurately classify instances for which the class is unknown while receiving test data.

Classification was conducted on the MIMIC-CXR dataset published by Medical Information Mart for Intensive Care (MIMIC) [27]. The MIMIC-CXR dataset was compiled utilizing chest X-ray images sourced from patients admitted to the emergency department of Beth Israel Deaconess Medical Center between 2011 and 2016. This dataset comprises 227,835 image studies, encompassing 377,110 radiology images obtained from 65,379 distinct patients. Each study is accompanied by semi-structured free-text radiology reports, adhering to pertinent health insurance and data protection regulations. Access to this dataset necessitates user registration,

authentication, and agreement with a stipulated data usage policy. A sample from the dataset is presented in Fig 1.

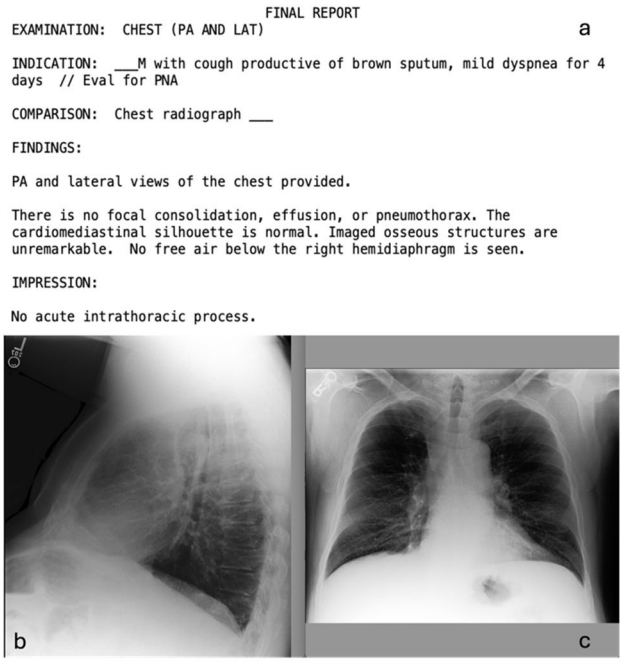

FIGURE 1. A sample study from MIMIC-CXR. (a), the radiology report provides the interpretation of the image. Personal patient information has been removed and replaced with three underscores (___). Two chest X-rays are shown for this study: (b) side view (right image) and (c) front view (left image).

Within the semi-structured reports contained in the MIMIC-CXR dataset, there exist eleven designated fields denoted by the following labels: 'IMPRESSION', 'FINDINGS', 'LAST_PARAGRAPH', 'COMPARISON', 'INDICATION', 'EXAMINATION', 'TECHNIQUE', 'HISTORY', 'NOTIFICATION', 'RECOMMENDATIONS', and 'WET READ'.

During chest radiograph evaluation, radiologists prioritize interpretations that address the specific clinical questions posed by the referring clinician. Radiologists meticulously examine image findings and articulate their detailed interpretations in the 'FINDINGS' section, drawing upon their domain expertise. Subsequently, they encapsulate their overall assessment in the 'IMPRESSION' section. Although some chest radiology reports may omit impressions, others may briefly address a single finding. Regardless, the radiologist's impression plays a crucial role in guiding the clinician's diagnosis and treatment plan. Chest radiography investigates a spectrum of 14 findings, including 13 specific abnormalities and a ''no finding'' class. To address challenges like time constraints and incomplete reporting, direct identification of findings could be proposed to improve diagnostic and therapeutic efficiency. As illustrated in Fig 2, multi-label classification aims to replace the reliance on radiologists' procedural knowledge for impression generation.

In this study, MIMIC-CXR dataset consisting 155,716 radiological reports, with a particular focus on the 'FINDINGS' section was utilized. 14 distinct impressions, comprising 13 positive categories and 1 negative category, derived from chest radiography findings in the MIMC-CXR dataset, were used in a multi-label classification procedure to reveal all the discoveries presented in the reports. Table 1 presents the file counts of findings that are in the radiology reports.

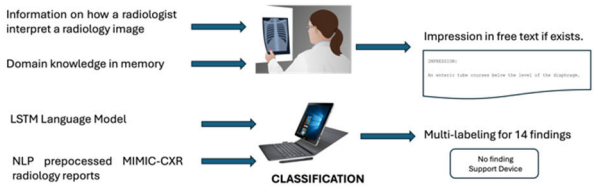

FIGURE 2. Radiologist's generation of impressions from reports and automatic multi-label findings classification.

In NLP studies, prior to submitting the dataset to classification algorithms, it is recommended to perform different pre-processing operations on the text. It is known that changes in sentence structure after tagging and lemmatization can harm the performance of language models [28]. A more detailed structural and semantic analysis of the MIMIC-CXR dataset is presented in [29]. In this study, free text in the 'FINDINGS' section of radiology reports were converted to lowercase in order to improve tokenization efficiency.

TABLE 1. File counts of findings in radiology reports.

Findings	Count
Atelectasis	24,110
Cardiomegaly	19,673
Consolidation	3,899
Edema	14,470
Enlarged Cardiomediastinum	2,911
Fracture	2,704
Lung Lesion	3,894
Lung Opacity	31,672
No Finding	65,899
Pleural Effusion	26,286
Pleural Other	1,037
Pneumonia	10,367
Pneumothorax	5,115
Support Devices	27,924
Total	**155,716**

TABLE 2. Bert classification hyperparameters used in experiments.

Language Model	Max_Length Batch Size	Number of Labels	Epoch	Optimizer	Learning Rate	Weighted F1 Score
bert-base-uncased	(128,32)	14	8	Adam	6x10^-6	0.7921
	(256,64)	14	8	Adam	6x10^-6	0.7960
ClinicalBERT	(128,32)	14	8	Adam	6x10^-6	0.7904
	(256,64)	14	8	Adam	6x10^-6	0.7893
Biobert-v1.1	(128,32)	14	8	Adam	6x10^-6	0.7928
	(256,64)	14	8	Adam	6x10^-6	0.8005
BiomedVLP-CXR-BERT-GENERAL	(128,32)	14	8	Adam	6x10^-6	0.8047
	(256,64)	14	8	Adam	6x10^-6	0.8014
BiomedVLP-CXR-BERT-SPECIAL	(128,32)	14	8	Adam	6x10^-6	0.8031
	(256,64)	14	8	Adam	6x10^-6	0.7954
S-PubMedBert-MS-MARCO- SCIFACT	(128,32)	14	8	Adam	6x10^-6	0.7911
	(256,64)	14	8	Adam	6x10^-6	0.7972

TABLE 3. BERT classification best result on CXR-BERT-GENERAL.

Class	Precision	Recall	F1-Score	Support
Atelectasis	0.744721284	0.74268267	0.74370058	4749
Cardiomegaly	0.787805533	0.734719439	0.760337006	3992
Consolidation	0.655709343	0.501322751	0.568215892	756
Edema	0.874819625	0.827080491	0.850280505	2932
Enlarged Cardiomediastinum	0.611627907	0.460595447	0.525474525	571
Fracture	0.614953271	0.616104869	0.615528531	534
Lung Lesion	0.632	0.717250324	0.671931956	771
Lung Opacity	0.745584989	0.844903064	0.792143067	6396
No Finding	0.873749332	0.863005432	0.86834415	13256
Pleural Effusion	0.785379266	0.936367142	0.854252793	5186
Pleural Other	0.789473684	0.342465753	0.477707006	219
Pneumonia	0.719633308	0.451798561	0.55509723	2085
Pneumothorax	0.896987366	0.91025641	0.903573177	1014
Support Device	0.829298463	0.848643761	0.838859594	5530
micro avg	0.802693463	0.807276364	0.804978391	47991
macro avg	0.754410241	0.699799723	0.716103287	47991
weighted avg	0.802238764	0.807276364	0.801422666	47991
samples avg	0.762484175	0.765653401	0.750970783	47991

Three different classification algorithms, BERT, CNN-BERT, and BILSTM-BERT, were used to evaluate the effectiveness of six language models derived from the BERT language model, trained on the MIMIC-CXR dataset. The algorithms used for multi-label classification were evaluated using ''precision,'' ''recall,'' and ''F-1 score''. The classification process consists of training, verification and testing phases. While 80% of the dataset is allocated to the training phase, the remaining 20% is divided equally between validation and testing.

TABLE 4. CNN-BERT classification hyperparameters used in experiments.

Language Model	Max_Length Batch Size	Number of Labels	Epoch	Optimizer	Learning Rate	Weighted F1 Score
bert-base-uncased	(72,128)	14	8	AdamW	$2e^{-5}$	0.78
ClinicalBERT	(72,128)	14	8	AdamW	$2e^{-5}$	0.7866
Biobert-v1.1	(72,128)	14	8	AdamW	$2e^{-5}$	0.7879
BiomedVLP-CXR- BERTGENERAL	(72,128)	14	8	AdamW	$2e^{-5}$	0.7999
BiomedVLP-CXR- BERT-SPECIAL	(72,128)	14	8	AdamW	$2e^{-5}$	0.7923
S-PubMedBert-MS- MARCO-SCIFACT	(72,128)	14	8	AdamW	$2e^{-5}$	0.7872

TABLE 5. CNN-BERT classification best results on CXR-BERT-GENRERAL.

Class	Precision	Recall	F1-Score	Support
Atelectasis	0.776894294	0.699957859	0.736422079	2373
Cardiomegaly	0.801169591	0.705821741	0.750479321	1941
Consolidation	0.684444444	0.421917808	0.522033898	365
Edema	0.866477273	0.840799449	0.853445261	1451
Enlarged Cardiomediastinum	0.717948718	0.361290323	0.480686695	310
Fracture	0.730994152	0.510204082	0.600961538	245
Lung Lesion	0.704467354	0.590778098	0.642633229	347
Lung Opacity	0.776460125	0.8184375	0.796896394	3200
No Finding	0.881587105	0.858566038	0.869924295	6625
Pleural Effusion	0.836229205	0.87	0.852780396	2600
Pleural Other	0.740740741	0.444444444	0.555555556	90
Pneumonia	0.787148594	0.399185336	0.52972973	982
Pneumothorax	0.93697479	0.864341085	0.899193548	516
Support Device	0.861938891	0.822830393	0.84193073	2777
micro avg	0.832840237	0.779909328	0.805506178	23822
macro avg	0.793105377	0.657755297	0.709476619	23822
weighted avg	0.829818726	0.779909328	0.799895459	23822
samples avg	0.768632802	0.745898239	0.744009206	23822

The next section details the conducted experiments.

IV. CLASSIFICATION EXPERIMENTS

The experiments address the multi-label classification task by using three model architectures: BERT, CNN-BERT, and BILSTM-BERT. The subsequent subsections provide a comprehensive explanation of the procedures involved in utilizing these models.

A. BERT CLASSIFICATION

The classification task was performed on the MIMIC-CXR dataset by taking advantage of the architecture of the BERT model, which is one of the transformers-based models. The analysis of the classification task was implemented on six different language models using the BERT architecture: BERT, ClinicalBERT, BioBERT, CXR-BERT, CXR-BERT-SPECIAL, and S-PubMedBert. To improve the performance of the classification algorithm, the max_length and batch size settings in the training phase of the BERT architecture were tested. Due to VRAM limitations, special parameters for training on the graphics card were tested on the Google Colab platform. NVIDIA MSI 3060 12 GB and A100 40 GB GPUs were used to test LMs with various parameters. The parameters applied for each language model are given in Table 2.

B. CNN-BERT CLASSIFICATION

The CNN-BERT classification method was developed by Safaya et al. [10]. The dataset was processed using BERT, a pre-trained language model with a 72-token max_length. Various language models were tested, including

TABLE 6. Examined BERT-BILSTM classification hyperparameters.

Language Model	Max_Length, Batch Size	Embedding Size	Number of Labels	Epoch	Optimizer	Learning Rate	Weighted F1 Score
bert-base- uncased	(256,128)	300	14	8	Adam	0.001	0.7799
ClinicalBERT	(256,128)	300	14	8	Adam	0.001	0.7787
Biobert-v1.1	(256,128)	300	14	8	Adam	0.001	0.7785
BiomedVLP-CXR-BERT-GENERAL	(256,128)	300	14	8	Adam	0.001	0.7791
BiomedVLP-CXR-BERT-SPECIAL	(256,128)	300	14	8	Adam	0.001	0.7785
S-PubMedBert-MS-MARCO-SCIFACT	(256,128)	300	14	8	Adam	0.001	0.7721

TABLE 7. BERT-BILSTM classification best results on CXR-BERT-GENERAL.

Class	Precision	Recall	F1-Score	Support
Atelectasis	0.7486785199	0.6565528866	0.6995958689	2373
Cardiomegaly	0.7771156138	0.6718186502	0.7206410611	1941
Consolidation	0.5754716981	0.501369863	0.5358711567	365
Edema	0.8768060837	0.794624397	0.8336948662	1451
Enlarged Cardiomediastinum	0.5434782609	0.3225806452	0.4048582996	310
Fracture	0.546875	0.4285714286	0.4805491991	245
Lung Lesion	0.5979643766	0.6772334294	0.6351351351	347
Lung Opacity	0.7564887722	0.810625	0.7826218132	3200
No Finding	0.8528403161	0.8633962264	0.8580858086	6625
Pleural Effusion	0.7993795243	0.8919230769	0.8431194328	2600
Pleural Other	0.8181818182	0.2	0.3214285714	90
Pneumonia	0.7354085603	0.3849287169	0.5053475936	982
Pneumothorax	0.8884540117	0.8798449612	0.8841285297	516
Support Device	0.8223896663	0.8253510983	0.8238677211	2777
micro avg	0.799452935	0.7729409789	0.7859734494	23822
macro avg	0.7385380159	0.6363443128	0.6663532184	23822
weighted avg	0.7960913642	0.7729409789	0.7798846335	23822
samples avg	0.7466735166	0.7377582474	0.728099213	23822

BERT, ClinicalBERT, BioBERT, CXR-BERT, CXR-BERT-SPECIAL, and S-PubMedBert. The model was trained for eight epochs using a learning rate of $2e^{-5}$, AdamW optimizer with 0.9 weight decay, BCE loss function, and batch size 128. The model with the highest weighted F1-score on the development set was retained. Table 4 summarizes the language models with attempted parameters.

According to Table 4, the CXR-BERT-GENERAL model achieved the best multi-label classification results. However, it does not exhibit a significant benefit compared to the

TABLE 8. Comparison of text and image classification for MIMIC-CXR dataset.

Pathology (Finding)	Our study (Text)		(Yarnall, 2020) (Image)	
	Recall	Precision	Recall	Precision
Atelectasis	74.26%	74.47%	76.98%	78.16%
Cardiomegaly	73.47%	78.78%	71.13%	79.13%
Consolidation	50.13%	65.57%	31.42%	89.13%
Edema	82.70%	87.48%	83.02%	82.00%
Enlarged Cardiomediastinum	46.05%	61.16%	83.42%	72.49%
Fracture	61.61%	61.49%	75.69%	60.12%
Lung Lesion	71.72%	63.20%	60.29%	65.00%
Lung Opacity	84.49%	74.55%	72.96%	71.50%
No Findings	86.30%	87.37%	-	-
Pleural Effusion	93.63%	78.53%	81.37%	85.00%
Pleural Other	34.24%	78.94%	53.10%	73.17%
Pneumonia	45.17%	71.96%	61.38%	69.25%
Pneumothorax	91.02%	89.69%	74.17%	77.61%
Support Device	84.86%	82.92%	-	-

other options. The evaluation of the CNN-Bert Classification on CXR-BERT-GENERAL language model is presented in Table 5.

C. BERT-BILSTM CLASSIFICATION

The experimental study presented in this subsection demonstrates the classification procedure, which is widely acknowledged as one of the functions performed by the Bidirectional Long Short-Term Memory (BILSTM) framework. For the multi-label classsification, the language models, BERT, ClinicalBERT, BioBERT, CXR-BERT, CXR-BERT-SPECIAL, and PubMedBert, were used to perform tokenization of the words present in the sentences contained within the dataset. The vectorization process was applied to the tokenized words, and subsequently, the interrelationships among these vectors were analyzed based on the numerical values assigned by the tokenizer. The tokenize operation has been configured with a maximum length of 512. Furthermore, a batch size of 128 and 8 epochs was utilized during the training stage, with a learning rate of 0.001. The dimensionality of the embedding utilized in BILSTM is established as 300. The LSTM model from Safaya et al. [10] was used. Table 6 summarizes the utilized parameters and language models.

Following the completion of the study and the interpretation of the data, it was determined that the BERT-base-uncased language model generated the most productive results according to Table 6. However, it does not exhibit a significant benefit compared to the other options. The evaluation of the BERT-BILSTM classification on Bert-base-uncased language model is presented in Table 7.

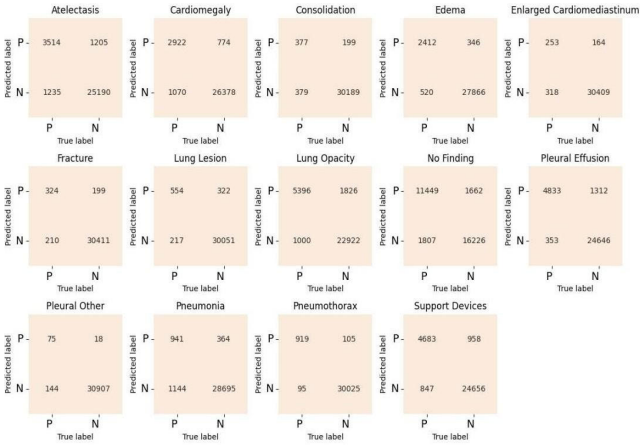

FIGURE 3. Confusion matrix of CXR-BERT-GENERAL.

V. RESULTS

This study aimed to evaluate the performance of three classification models: BERT, CNN-BERT, and BILSTM-BERT on a set of 14 findings from the MIMIC-CXR dataset. Each model was applied to six distinct language models: BERT, ClinicalBERT, BioBERT, CXR-BERT, CXR-BERT-SPECIAL, and S-PubMedBert. Among them, the BERT classification technique achieved a slightly better weighted F-1 score of 0.8047 with the CXR-BERT-GENERAL language model. The confusion matrix for the model is shown in Fig 3. However, it was discovered within the confines of the research that the categorization models did not exhibit superiority over one another as seen in Fig 4.

A comparative analysis was performed to evaluate the text and image classification of MIMIC-CXR. The results, including the recall and precision metrics for each pathology,

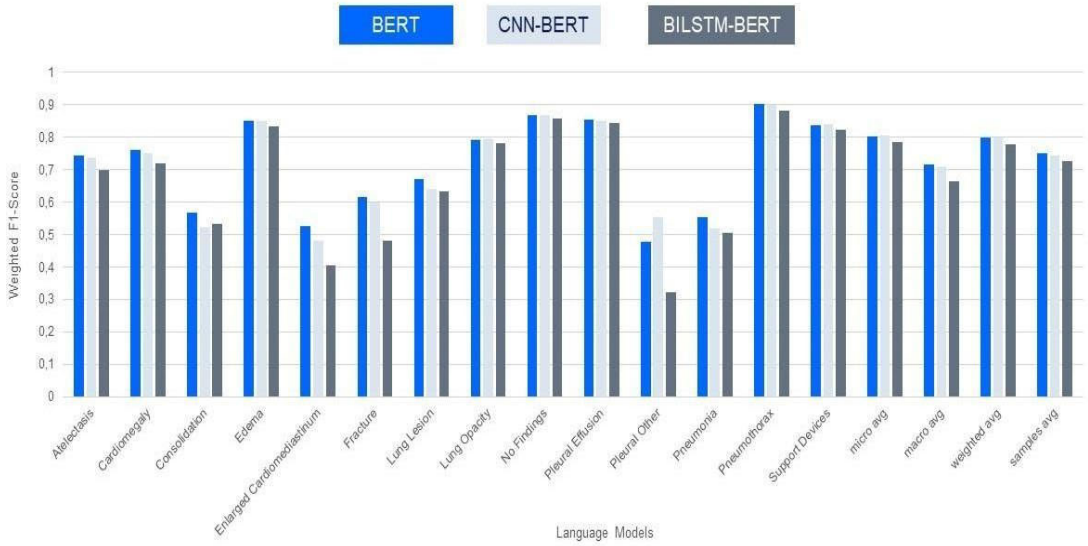

FIGURE 4. F1-Score distributions of classifications methods.

are shown in Table 8. Hence, a clear difference between image classification [30] and text classification is not evident when their respective results are compared assessed. The classification of 'Enlarged Cardiomediastinum' and 'Pneumonia' is more clearly delineated within the domain of image classification. Conversely, for diagnoses related to 'Pneumothorax,' 'Pleural Effusion,' 'Lung Opacity,' and 'Consolidation,' text classification achieved significantly higher precision than image classification.

Testing with two different batch sizes and maximum lengths revealed variability in BERT classification training times across configurations. The average duration for a (256,64) configuration was 2.30 hours, while models with (128,32) varied by a standard deviation of 7 hours. The overall training procedure spanned six language models and required an estimated 22 hours. Compared to this, the average CNN-BERT training session with six language models lasted 1 hour and 15 minutes, totaling 7 hours and 10 minutes. LSTM-BERT was the fastest, with single training sessions requiring only 15 minutes, and a six-model run completing in 1.30 hours.

VI. CONCLUSION AND FUTURE WORK

This study aimed to evaluate the performance of three classification models; BERT, CNN-BERT, and LSTM-BERT, for multi-label categorizing 14 medical findings from the MIMIC-CXR dataset. While all models demonstrated promising results, the BERT model achieved a slightly higher weighted F-1 score of 0.8047 when paired with the CXR-BERT-GENERAL language model. However, it's important to note that no statistically significant differences were observed in overall performance between the classification models. This suggests that other factors, such as the specific dataset or task, may play a more significant role in determining the optimal classification approach.

Moreover, no clear advantage of any certain language model over another was seen. The BERT classification task included trials on individual language models with varying batch sizes and maximum durations, leading to disparate results. Differences in the duration of the training process were noted across different experimental setups. On average, the training time for a setup size (256, 64) was determined to be 2.30 hours. Nevertheless, it was noted that the training periods for configurations (128, 32) varied and may last as long as 7 hours. The training consisted of six distinct language models and required around 22 hours.

Additionally, the average time of each training session for the CNN-BERT classification was 75 minutes. The period under consideration consisted of six distinct iterations and lasted for seven hours and ten minutes. The LSTM-BERT classification model required a training time of around 15 minutes for each.

This study compared text and image classification methods on the MIMIC-CXR dataset, focusing on recall and accuracy for individual pathologies. While disentangling their distinct outputs proved challenging, some findings emerged. 'Enlarged Cardiomediastinum' and 'Pneumonia' exhibited clearer delineation with image classification, while 'Pneumothorax', 'Pleural Effusion', 'Lung Opacity', and 'Consolidation' showed greater discernibility with text classification.

While this research utilized all available records, potential bias could be mitigated by future studies focusing on a single report per patient. In addition, considering the temporal aspect of medical data, future work could analyze sequence of records, historical findings, and discovery dates to assess their influence on outcomes. This could involve expanding the study's range by incorporating text analysis of radiological reports and even including the images themselves.

Furthermore, automatic report generation based on radiological images holds promise for enhanced automation in healthcare. Integrating such generated reports into existing classification models could significantly streamline the medical imaging workflow, enabling healthcare professionals to focus on higher-level tasks and ultimately benefit patient care.

It is conceivable that the predictive precision of the MIMIC-CXR dataset could be enhanced by using modern tools beyond BERT; however, several essential factors need to be considered. Generative AI techniques, such as Conditional Generative Adversarial Networks (CGANs) and Variational Autoencoders (VAEs), have shown promise in tasks like image generation and data augmentation, potentially improving model performance on datasets like MIMIC-CXR. Synthetic data generation is particularly beneficial for datasets with limited labeled data, such as MIMIC-CXR. Synthetic data can support the model's ability to generalize and reduce overfitting. By data augmentation, generative models allow the model to learn from a broader range of examples and enhance its resilience to variations in real data.

Two generative AI studies in chest radiology were analyzed. While the first study solely produced synthetic images, the second generated both synthetic images and corresponding reports. In [31], a Turing test was employed to assess radiologists' ability to discern the authenticity of the generated images. Conversely, [32] focused on evaluating the accuracy of the produced dataset. Both studies identified common limitations associated with synthetic data: its time-consuming and complex generation, potential shortcomings in capturing the full complexity and diversity of real-world images, and potential impact on the model's generalizability to real-world data.

Implementing generative models effectively requires significant expertise and meticulous tuning of hyperparameters. Additionally, training these models can be computationally expensive, especially for large datasets like MIMIC-CXR. Furthermore, the efficacy of generative models is heavily dependent on the specific task and dataset. While they have shown promise in certain domains, there is no guarantee of substantial improvements in all cases. Beyond the choice of algorithm, the model architecture can significantly impact accuracy. Exploring different architectures, such as convolutional neural networks (CNNs) specifically designed for medical image analysis, could be beneficial. Additionally, the quality and preprocessing of the data used for training are crucial for achieving high accuracy. Techniques like noise reduction, normalization, and data augmentation can significantly boost the model's performance. Moreover, optimizing the training process through methods like hyperparameter tuning, learning rate adjustment, and regularization can potentially lead to improvements in accuracy.

ACKNOWLEDGMENT

The authors are grateful to Associate Professor Dr. Ezgi Guler and Specialist Dr. Akın Cinkooglu, Department of Radiology, Faculty of Medicine, Ege University, for their invaluable insights into chest radiology reports. Their expertise has significantly contributed to our understanding and practice in this field.

REFERENCES

[1] J. T. P. D. Hallinan, M. Feng, D. Ng, S. Y. Sia, V. T. Y. Tiong, P. Jagmohan, A. Makmur, and Y. L. Thian, "Detection of pneumothorax with deep learning models: Learning from radiologist labels vs natural language processing model generated labels," *Academic Radiol.*, vol. 29, no. 9, pp. 1350–1358, Sep. 2022.

[2] A. Névéol, T. M. Deserno, S. J. Darmoni, M. O. Güld, and A. R. Aronson, "Natural language processing versus content-based image analysis for medical document retrieval," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 1, pp. 123–134, Sep. 2008.

[3] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, Dec. 2019.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[6] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," 2019, *arXiv:1904.05342*.

[7] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323*.

[8] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the most of text semantics to improve biomedical vision–language processing," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, 2022, pp. 1–21.

[9] P. Deka, A. Jurek-Loughrey, and P. Deepak, "Improved methods to aid unsupervised evidence-based fact checking for online heath news," *J. Data Intell.*, vol. 3, no. 4, pp. 474–504, Nov. 2022.

[10] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media," in *Proc. 14th Workshop Semantic Eval.*, 2020, pp. 2054–2059.

[11] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.

[12] S. R. Joseph, H. Hlomani, K. Letsholo, F. Kaniwa, and K. Sedimo, "Natural language processing: A review," *Int. J. Res. Eng. Appl. Sci.*, vol. 6, no. 3, pp. 207–210, Mar. 2016.

[13] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2020, *arXiv:2003.01200*.

[14] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A survey of text representation and embedding techniques in NLP," *IEEE Access*, vol. 11, pp. 36120–36146, 2023, doi: 10.1109/ACCESS.2023.3266377.

[15] B. G. Patra et al., "Extracting social determinants of health from electronic health records using natural language processing: A systematic review," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 12, pp. 2716–2727, Nov. 2021, doi: 10.1093/jamia/ocab170.

[16] H. Yeo, "A machine learning based natural language question and answering system for healthcare data search using complex queries," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Seattle, WA, USA, Dec. 2018, pp. 2467–2474, doi: 10.1109/BIGDATA.2018.8622448.

[17] P. Rajpurkar et al., "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002686, doi: 10.1371/journal.pmed.1002686.

[18] H. Liu, T. Christiansen, W. A. Baumgartner, and K. Verspoor, "BioLemmatizer: A lemmatization tool for morphological processing of biomedical text," *J. Biomed. Semantics*, vol. 3, no. 1, pp. 1–29, Apr. 2012.

[19] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. 215–220, Jun. 2000, doi: 10.1161/01.cir.101.23.e215.
[20] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, "Syntactic N-grams as machine learning features for natural language processing," *Expert Syst. Appl.*, vol. 41, no. 3, pp. 853–860, Feb. 2014.
[21] X. Cai, S. Liu, J. Han, L. Yang, Z. Liu, and T. Liu, "ChestXRayBERT: A pretrained language model for chest radiology report summarization," *IEEE Trans. Multimedia*, vol. 25, pp. 845–855, 2023, doi: 10.1109/TMM.2021.3132724.
[22] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 590–597.
[23] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, Jun. 2014, pp. 55–60, doi: 10.3115/v1/P14-5010.
[24] M. Hughes, L. Irene, K. Spyros, and S. Toyotaro, "Medical text classification using convolutional neural networks," in *Studies in Health Technology and Informatics*, vol. 235, 2017, pp. 246–250.
[25] D. Dessi, R. Helaoui, V. Kumar, D. R. Recupero, and D. Riboni, "TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study," 2021, *arXiv:2105.09632*.
[26] S. Mao, L.-L. Zhang, and Z.-G. Guan, "An LSTM&Topic-CNN model for classification of online Chinese medical questions," *IEEE Access*, vol. 9, pp. 52580–52589, 2021, doi: 10.1109/ACCESS.2021.3070375.
[27] *MIMIC—Medical Information Mart for Intensive Care*. Accessed: Feb. 23, 2024. [Online]. Available: https://mimic.mit.edu/
[28] A. Davies, J. Jiang, and C. Zhai, "Competence-based analysis of language models," 2023, *arXiv:2303.00333*.
[29] E. E. Uslu, E. Sezer, and Z. A. Guven, "Semantic and structural analysis of MIMIC-CXR radiography reports with NLP methods," *Politeknik Dergisi*, p. 1, Feb. 2024, doi: 10.2339/politeknik.1395811.
[30] J. Yarnall, "X-ray classification using deep learning and the MIMIC-CXR dataset," Dept. Comput. Sci., Villanova Univ., Villanova, PA, USA, Order 28030044, 2020.
[31] Y. Myong, D. Yoon, B. S. Kim, Y. G. Kim, Y. Sim, S. Lee, J. Yoon, M. Cho, and S. Kim, "Evaluating diagnostic content of AI-generated chest radiography: A multi-center visual Turing test," *PLoS ONE*, vol. 18, no. 4, Apr. 2023, Art. no. e0279349, doi: 10.1371/journal.pone.0279349.
[32] J. Shentu and N. Al Moubayed, "CXR-IRGen: An integrated vision and language model for the generation of clinically accurate chest X-ray image-report pairs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5212–5221.

EGE ERBERK USLU received the B.S. degree in computer engineering from Yasar University, İzmir, Turkey, in 2021, and the M.S. degree in computer engineering from Ege University, İzmir, in 2023, where he is currently pursuing the Ph.D. degree in computer engineering.

Since 2022, he has been a Research Assistant with the Computer Engineering Department, Ege University. His research interests include ad-hoc networks, wireless sensor networks, the Internet of Things, artificial intelligence, and security.

EMINE SEZER received the B.S. degree in electrical-electronics engineering from Dokuz Eylul University, İzmir, Turkey, in 2000, and the M.S. and Ph.D. degrees in computer engineering from Ege University, İzmir, in 2004 and 2014, respectively.

From 2001 to 2017, she was a Research Assistant with the Computer Engineering Department, Ege University, where she has been an Assistant Professor, since 2017. Her research interests include semantic web, ontology modeling, knowledge graph, health information systems, machine learning, natural language processing, and LLM.

ZEKERIYA ANIL GUVEN was born in Samsun, Turkey, in 1992. He received the B.S. degree in computer engineering from Kocaeli University, Kocaeli, Turkey, in 2015, the M.S. degree in computer engineering from Yıldız Technical University, İstanbul, Turkey, in 2018, and the Ph.D. degree in computer engineering from Ege University, İzmir, Turkey, in 2022.

From 2018 to 2022, he was a Research Assistant with the Computer Engineering Department, Ege University. He is currently an Assistant Professor with the Department of Computer Engineering, İzmir Bakırçay University. In 2018, he focused on creating new approaches within the question-answering and semantic domain. During his Ph.D. studies, he got familiar with machine learning and language models. He has published a good number of articles on topics related to how to apply language models to question-answering. His research interests include NLP, ML, data mining and LLM, sentiment analysis, and question answering.