Attention Mechanisms in Clinical Text Classification: A Comparative Evaluation

Christoph S. Metzner, Shang Gao, Drahomira Herrmannova, Elia Lima-Walton, and Heidi A. Hanson

***Abstract*—Attention mechanisms are now a mainstay architecture in neural networks and improve the performance of biomedical text classification tasks. In particular, models that perform automated medical encoding of clinical documents make extensive use of the label-wise attention mechanism. A label-wise attention mechanism increases a model's discriminatory ability by using label-specific reference information. This information can either be implicitly learned during training or explicitly provided through embedded textual code descriptions or information on the code hierarchy; however, contemporary studies arbitrarily select the type of label-specific reference information. To address this shortcoming, we evaluated label-wise attention initialized with either implicit or explicit label-specific reference information against two common baseline methods—target-attention and text-encoder architecture-specific methods—to generate document embeddings across four text-encoder architectures—a convolutional neural network, two recurrent neural networks, and a transformer. We also present an extension of label-wise attention that can embed the information on the code hierarchy. We performed our experiments on the MIMIC III dataset, which is a standard dataset in the clinical text classification domain. Our experiments showed that using pretrained reference information and the hierarchical design helped improve classification performance. These performance improvements had less impact on larger datasets and label spaces across all text-encoder architectures. In our analysis, we used an attention mechanism's energy scores to explain the perceived differences in performance and interpretability between the text-encoder architectures and types of label-attention.**

***Index Terms*—Attention, natural language processing, neural networks, text classification, transformer.**

I. Introduction

ROUGHLY 80% of the information stored in EHRs is unstructured data and comes as free-form text written by healthcare professionals [1]. However, the manual retrieval and classification of this information are infeasible because these activities are time-consuming, cost-intensive, error-prone, and require significant domain knowledge [2]. To overcome these barriers and automate the classification of clinical documents, the biomedical informatics community leverages machine learning, natural language processing (NLP), and deep learning-based methods. Particularly, contemporary NLP models use the popular domain-agnostic building block located at specific positions in the architecture of deep neural networks called attention.

Attention mechanisms dynamically manage the perceived information stream by using soft weights to indicate the relevance of a token to a task within a text input sequence [3]. The introduction of the Transformer [4] further increased the popularity of attention mechanisms because it enabled state-of-the-art performance across multiple NLP tasks while relying solely on multiple self-attention layers. As a result, current state-of-the-art clinical document classification models incorporate different types of attention mechanisms. For example, self- and target-attention have been deployed by the HiSAN model to classify cancer pathology reports [5], whereas models that perform the automated encoding of hospital discharge summaries primarily utilize label-wise attention [6], [7], [8].

Label-wise attention's popularity in extreme multi-label scenarios stems from its ability to incorporate label-specific auxiliary information [9], [10]. Auxiliary information can be either implicitly learned through randomly initialized embeddings or explicitly provided by using embeddings of textual code descriptions or information on the code hierarchy [11]. Selecting the appropriate type of auxiliary information for the task and dataset at hand is nontrivial; however, in most studies, this selection is often arbitrary. Furthermore, given the often-existing linguistic mismatch between the formally defined code descriptions and the informally written clinical documents [12], it is important to quantify the differences between the outputs of different types of text-encoder architectures with differently initialized label-attention mechanisms.

Manuscript received 22 May 2023; revised 4 December 2023 and 12 January 2024; accepted 16 January 2024. Date of publication 19 January 2024; date of current version 5 April 2024. This work was supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE), in part by the National Cancer Institute (NCI) of the National Institutes of Health, in part by Argonne National Laboratory under Grant DE-AC02-06-CH11357, in part by Lawrence Livermore National Laboratory under Grant DEAC52-07NA27344, in part by Los Alamos National Laboratory under Grant DE-AC5206NA25396, in part by Oak Ridge National Laboratory (ORNL) under Grant DE-AC05-00OR22725 performed under the auspices of DOE, and in part by the UT-Battelle LCC under Grant DE-ACO5-000R22725 with the DOE. *(Corresponding author: Christoph S. Metzner.)*

Christoph S. Metzner and Heidi A. Hanson are with the ORNL, Oak Ridge, TN 37830 USA (e-mail: metznercs@ornl.gov; hansonha@ornl.gov).

Shang Gao is with the Casetext Inc., San Francisco, CA 94107 USA (e-mail: shang@casetext.com).

Drahomira Herrmannova and Elia Lima-Walton are with the Elsevier, Philadelphia, PA 19103 USA (e-mail: d.herrmannova@elsevier.com; e.lima@elsevier.com).

Digital Object Identifier 10.1109/JBHI.2024.3355951

In this work, we evaluate label-attention mechanisms that incorporate either implicit or explicit auxiliary information across multiple text-encoder architectures — convolutional neural networks (CNN), recurrent neural networks (RNN), and transformers — in the multi-label classification setting using hospital discharge summary notes. We also compare the performance of our label-attention mechanisms against target-attention and other common methods for generating document embeddings. Our contributions are as follows:

- We perform the first comprehensive comparative study on the label-attention mechanism in clinical text classification with a focus on improving the automated encoding of clinical documents with medical codes (i.e., ICD-9).
- We examine the effect of different attention mechanisms on CNN-, RNN-, and transformer-based text-encoder architecture-specific document embeddings.
- We show that initializing an attention mechanism's reference information with explicit label-specific auxiliary information improves classification performance.
- We are the first to demonstrate a performance improvement in clinical document classification by incorporating information on code hierarchy solely via the attention mechanism with minimal increase in compute time and without requiring an additional neural network.

II. Related Work

Attention mechanisms are universally used in contemporary deep learning models and significantly improve performance in a wide range of NLP tasks, including aspect-level sentiment analysis [13], opinion mining [14], neural machine translation [15], and text classification [16]. Despite the ubiquity of attention mechanisms in deep learning, only a few studies in the NLP domain have empirically compared their effects on the performance of deep neural networks. Kardakis et al. [13] investigated global-, self-, and hierarchical-attention mechanisms in RNNs by performing sentiment analysis of movie reviews that showed an average increase in accuracy through attention by two points. Jain et al. [17] and Feucht et al. [18] analyzed a single type of attention across multiple encoder architectures to predict ICU-related tasks (e.g., readmission, medical encoding of notes) by using MIMIC-III clinical notes. Jain et al. [17] tested target-attention, which utilizes a single query that contains the reference information about the entire label space, to find important parts within documents and reported accuracy improvements in CNNs and Bi-LSTMs. Feucht et al. [18] investigated how the performance of different transformers (e.g., BERT, Hierarchical-BERT, Longformer) is impacted by similar label-attention mechanisms with pretrained embeddings of textual code descriptions. Although their work is similar to ours, the underlying relationships between the different types of text-encoder architectures and attention mechanisms remain unexplored.

A noticeable trend in deep learning models used to perform automated medical encoding of clinical documents is the reliance on an attention mechanism that can incorporate label-specific information. Early work by Shi et al. [12] highlighted the benefits of incorporating textual code descriptions in their attention mechanism to assign ICD-9 codes to sentence-level diagnosis descriptions. Their attention mechanism creates a single weighted document representation that contains the similarities between a diagnosis description and all considered ICD-9 code description embeddings. Mullenbach et al. [6] argued that using a single document representation is insufficient to assign multiple medical codes to a clinical note and proposed a label-attention mechanism instead. Label-attention generates one document representation for each label in the label space to capture locations in the text that are semantically relevant to the specific label by using queries with label-specific reference information. Contemporary state-of-the-art models utilize label-attention as a critical component in their architecture but differ in their approach to initializing the query matrix randomly or pretrained. Randomly initialized queries are learned during training [6], [7], [19], whereas pretrained queries use label description embeddings as a starting point and are fine-tuned during training [6], [9]. For example, Liu et al. [8] proposed the hierarchical label-wise attention transformer model that uses randomly initialized label-attention on token- and chunk-level data to extract the most informative features from a clinical document, thereby achieving state-of-the-art performance on MIMIC-III-50. Despite label-attention's widespread use, the different architectures of the proposed models prevent a fair evaluation of both query initialization strategies. Therefore, we perform the first comparative evaluation of label-attention by following the implementation of Mullenbach et al. [6].

Medical encoding systems have an intrinsic hierarchical structure that provides information about the relationships between high-level categories and low-level codes. A large body of work on the subject contains diverse methods to exploit these hierarchical relationships to improve model performance [11]. Some studies use label co-occurrence or label correlation during the training process to initialize specific parts of the neural network by introducing bias and exploiting the parent-child relationships between the labels [20], [21]. Others use graph neural networks (GNN) to learn the hierarchical relationships of the label space [22], [23]. Xu et al. [24] combined label correlation and a GNN to learn a precise representation of the hierarchy. Shen et al. [25] instead used weak supervision to create a hierarchical structure of the label space. In particular, they approached the problem from the perspective of a layman trying to classify an unknown set of documents. They argue that laymen would first classify the documents with high-level categories before selecting more specific low-level labels. This way, knowledge about high-level categories will enrich the available information for predicting the low-level labels associated with a document. We adopted their idea of a layman approach to this problem in our implementation of the hierarchical label-attention mechanism and show that it is possible to incorporate the hierarchical structure in a simple attention mechanism.

III. Materials and Methods

A. General Document Classification Attention Model

Many deep learning models incorporate different attention mechanisms within different parts of their architectures, thereby making it difficult to perform a direct and fair comparison

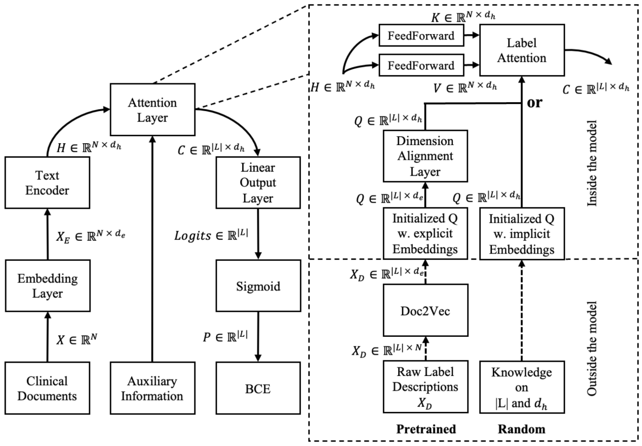
$K \in \mathbb{R}^{N \times d_h}$
FeedForward
FeedForward
Label Attention
$H \in \mathbb{R}^{N \times d_h}$
$V \in \mathbb{R}^{N \times d_h}$
$C \in \mathbb{R}^{|L| \times d_h}$
or
$Q \in \mathbb{R}^{|L| \times d_h}$
Dimension Alignment Layer
$Q \in \mathbb{R}^{|L| \times d_e}$
$Q \in \mathbb{R}^{|L| \times d_h}$
Initialized Q w. explicit Embeddings
Initialized Q w. implicit Embeddings
Inside the model
$X_D \in \mathbb{R}^{|L| \times d_e}$
Doc2Vec
$X_D \in \mathbb{R}^{|L| \times N}$
Raw Label Descriptions X_D
Knowledge on |L| and d_h
Outside the model
Pretrained
Random
Attention Layer
$H \in \mathbb{R}^{N \times d_h}$
$C \in \mathbb{R}^{|L| \times d_h}$
Text Encoder
Linear Output Layer
$X_E \in \mathbb{R}^{N \times d_e}$
$Logits \in \mathbb{R}^{|L|}$
Embedding Layer
Sigmoid
$X \in \mathbb{R}^{N}$
$P \in \mathbb{R}^{|L|}$
Clinical Documents
Auxiliary Information
BCE

Fig. 1. Structure of the general document classification attention model. The text-encoder segment represents a CNN-, RNN-, or transformer-based architecture. The label-attention layer utilizes implicit (i.e., random embeddings) or explicit (e.g., embedded textual code descriptions) auxiliary information. The multi-label classification models use a sigmoid activation function after the output layer and the binary-cross-entropy objective function.

of the impact of attention mechanisms on a model's performance [26]. With this fair comparison in mind, we designed a general attention-based document classification model (Fig. 1) that consists of the following layers: i) an embedding layer; ii) a text-encoder architecture; iii) an attention layer; and iv) a linear output layer. We describe each layer in more detail in the following sections.

B. Embedding Layer

The first layer of each model is a pretrained word-embedding layer that takes a text sequence as input, $X = [x_1, x_2, \ldots, x_N]$, where N is the number of words. The embedding layer transforms each word to its corresponding dense vector representation with dimension d_e, which results in a word-embedding matrix, $X_E \in \mathbb{R}^{N \times d_e}$. For the CNN- and RNN-based models, we initialize the word-embedding layer with pretrained embeddings of each word in the vocabulary generated with Word2Vec [27]; the vocabularies and word embeddings were created from the training documents of each dataset. In contrast, the pretrained transformer model comes with a byte-pair-encoding vocabulary based on the corpus associated with the model during pretraining [28].

C. Base Text-Encoder Architectures

We consider four base text-encoder architectures in this study: i) CNN, ii) bi-directional long short-term memory network (Bi-LSTM), iii) gated recurrent unit network (Bi-GRU), and (iv) transformer-based Clinical-Longformer (CLF). We limited the scope of our study to these four text-encoder architectures as they represent the most-common and fundamental architectures in the contemporary clinical text classification literature [29]; we selected the CLF as the representative of the transformer class as it was pretrained on clinical text and can process up to 4096 tokens, which is important for longer clinical documents.

Generally, each encoder type takes the word-embedding matrix, X_E, as the input and learns a latent document representation with dimension d_h, which results in a latent document matrix, H. A dropout layer with probability p is connected in series to prevent overfitting [30]. The model passes the dropout layer's output to an attention mechanism or other common method used to generate document embeddings.

1) Convolutional Neural Network (CNN): The first encoder architecture is a shallow CNN that follows the implementation by Kim [31]. The CNN passes the learned word-embedding matrix X_E to three parallel 1D convolution layers with d_h filters and window sizes of three, four, and five tokens, thereby learning multiple n-grams that contain relevant task-specific information. Following the ReLU activation, the three outputs are padded and concatenated, which results in the latent document matrix $H \in \mathbb{R}^{N \times 3d_h}$.

2) Recurrent Neural Network (RNN): The next text-encoder architectures are two popular RNNs: the Bi-LSTM [32] and the less complex Bi-GRU [33]. Each bi-directional, two-layered RNN model takes the word-embedding matrix X_E as the input and learns a latent document matrix $H \in \mathbb{R}^{N \times 2d_h}$ with $2d_h$ representing the bi-directional pass over the size of the hidden states.

3) Transformer-Based Clinical Longformer (CLF): The final text-encoder architecture is the transformer-based CLF [34]. The CLF takes the word-embedding matrix X_E as the input and learns a latent document representation $H \in \mathbb{R}^{N \times d_h}$ with hidden dimension $d_h = 768$.

D. Attention Mechanism

In text classification, attention mechanisms guide a neural network's prediction process by assigning attention weights to each token in the input text sequence [35]. Modern attention mechanisms usually follow the design proposed by Vaswani et al. [4], which utilizes three inputs—the key-value pair matrices and the query matrix. The keys, K, and values, V, are two learned sequences that encode the key features of the latent document representation, H, whereas the queries, Q, are a reference to the information that the model is searching for within the input [36]. The mechanism uses K and Q to find relationships between the input text sequence and the reference information. We investigated if the type of initialization of Q, either random or with pretrained embeddings, influences model performance. Specifically, we evaluated the popular label-attention mechanism [6] in the single- and multi-head setting and proposed a new hierarchical adaption of the label-attention mechanism that can incorporate the hierarchical relationships of the label space.

1) Label-Attention: Label-attention enables the model to learn a single document embedding for each label in the label space, $|L|$, thereby increasing a model's discriminatory ability [6]. In our implementation, we pass the latent document representation H into two separate position-wise feed-forward layers with the weights $W_K \in \mathbb{R}^{d_h \times d_h}$ and $W_V \in \mathbb{R}^{d_h \times d_h}$, biases b_K and b_V, kernel size and stride of one, and an Exponential

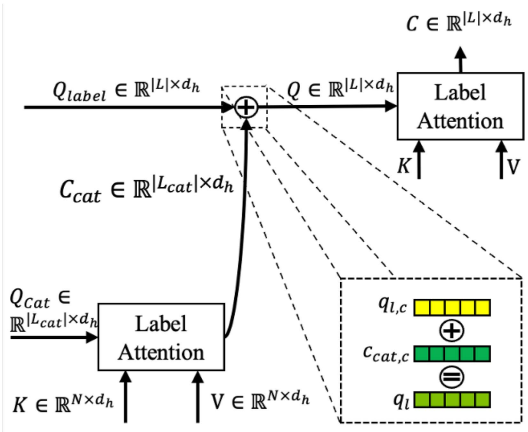
$C \in \mathbb{R}^{|L|\times d_h}$
$Q_{label} \in \mathbb{R}^{|L|\times d_h}$
$Q \in \mathbb{R}^{|L|\times d_h}$
Label Attention
K
V
$C_{cat} \in \mathbb{R}^{|L_{cat}|\times d_h}$
$Q_{Cat} \in \mathbb{R}^{|L_{cat}|\times d_h}$
Label Attention
$K \in \mathbb{R}^{N\times d_h}$
$V \in \mathbb{R}^{N\times d_h}$
$q_{l,c}$
$c_{cat,c}$
q_l

Fig. 2. Flow of our hierarchical label-attention mechanism. Both K and V are generated from the encoder output by using a feed-forward layer with ELU activation. Both Q and Q_{cat} are initialized with random or pretrained embeddings. The label query vector $q_{l,c}$ is enriched with information on the code hierarchy by adding the parent category context vector $c_{cat,c}$.

Linear Unit (ELU) activation function to generate the key-value pair matrices $K \in \mathbb{R}^{N\times d_h}$ (1) and $V \in \mathbb{R}^{N\times d_h}$ (2). The query matrix $Q \in \mathbb{R}^{|L|\times d_h}$ is a trainable embedding matrix initialized either randomly (random label-attention) or with pretrained embeddings of textual code descriptions (pretrained label-attention). Textual code descriptions vary in length; however Q consists of fixed length embeddings. The Doc2Vec algorithm [37] is a natural solution to this issue, as it generates fixed-length embeddings with size d_e; each description represents its own document tagged with its respective medical code as the label. Each element of Q is a query and represents the reference information for a specific label. Label-attention uses the three inputs, K, V, and Q, to compute the label-specific context vector (3). We used a 1D convolution layer as a dimension alignment layer to map the resulting pretrained query embeddings with size d_e to d_h to avoid dimension mismatch between K and Q. The task-specific textual code descriptions were retrieved from ICD-9 [38].

$$K = ELU(FeedForward(H, W_K) + b_K) \quad (1)$$

$$V = ELU(FeedForward(H, W_V) + b_V) \quad (2)$$

$$C(Q, K, V) = softmax\left(\frac{Q \cdot K^{\top}}{\sqrt{d_h}}\right) V \quad (3)$$

The scaled matrix product (i.e., compatibility function) uses K and Q to compute the energy scores, $E \in \mathbb{R}^{|L|\times N}$, which indicate the strength and direction of the association between the hidden representation of a token and the label-specific reference information in Q. The softmax operation (i.e., distribution function) retrieves the label-specific attention scores, $A \in \mathbb{R}^{|L|\times N}$, and the attention scores are subsequently applied to V to generate the label context matrix, $C \in \mathbb{R}^{|L|\times d_h}$, for a given document.

2) Hierarchical Label-Attention: This study proposes an extension of label-attention (Fig. 2) that can incorporate the code hierarchy of medical encoding systems. Hierarchical label-attention allows us to exploit the interdependence between the less specific but easier-to-predict high-level and low-level classifications. We argue that performing label-attention on two hierarchy levels enriches the downstream information and will lead to improved model performance. The hierarchical label-attention first generates context vectors, $C_{Cat} \in \mathbb{R}^{|L_{Cat}|\times d_h}$, for all high-level categories, $|L_{Cat}|$, associated with the label space by using label-attention described in Section III-D1 and the three input matrices, K, V, and, $Q_{Cat} \in \mathbb{R}^{|L_{Cat}|\times d_h}$. Q_{Cat} contains random or pretrained reference information for the high-level categories. Subsequently, we enrich the low-level label query matrix $Q \in \mathbb{R}^{|L|\times d_h}$ with the high-level category information in C_{Cat} (4). Specifically, we add the context vector $c_{cat,c}$ of the high-level parent category c to each related low-level label query $q_{l,c} \in L_c$, where L_c refers to all labels associated with category c, to obtain an enriched label query q_l for label l for all task-specific categories L_{cat}:

$$q_l(l_c, c) = q_{l,c} + c_{cat,c}$$
$$\forall 1 <= q_{l,c} <= L_c \quad \text{and} \quad \forall 1 <= c <= L_{cat}. \quad (4)$$

We retrieved the high-level categories directly from the medical encoding system of each classification task. For the ICD-9 codes, we used the high-level chapters as the categories (e.g., *Infectious and Parasitic Diseases*) [38]. Our hierarchical label-attention design was inspired by a proposed attention-via-attention mechanism [26].

3) Multi-Head Label-Attention: We implemented a multi-head variant of the label-attention and hierarchical label-attention mechanisms. Multi-head attention [4] uses h attention heads to linearly project h different parts of the Q, K, and V matrices in parallel (5). The output is then concatenated and projected by using $W_O \in \mathbb{R}^{(h\times d_v)\times d_h}$.

$$Multihead(Q, K, V) = Concat(head_1, \ldots, head_h)W_O$$
$$\text{where} \quad head_i = Attention(QW_{i,Q}, KW_{i,K}, VW_{i,V}) \quad (5)$$

The query, key, and value projection matrices have the shape of $W_{i,Q} \in \mathbb{R}^{d_h\times d_k}$, $W_{i,K} \in \mathbb{R}^{d_h\times d_k}$, $W_{i,V} \in \mathbb{R}^{d_h\times d_v}$, and $d_k = d_v = d_h/h$.

4) Baseline Methods: We tested the validity of the label-attention mechanism against two baseline methods: an architecture-specific baseline method for generating document embeddings and an alternative-attention mechanism called *target-attention*. For the first baseline, we substituted label-attention with a max-pooling operation over the logits after the output layer for CNN-based models. We used the average hidden state space over all time steps as the input to the output layer for both RNNs. Lastly, the transformer-typical baseline method utilizes the hidden representation of the first token of the encoded sequence as the input to the classification layer. Target-attention deploys a query matrix, $Q_T \in \mathbb{R}^{1\times d_h}$, with a single element that attempts to capture the task-specific reference information on the label space and generates only a single context vector for

TABLE I
STATISTICS FOR MIMIC-III

Dataset	N_{total}	N_{train}	N_{val}	N_{test}	$\overline{N_d}(N_{d,\sigma})$	N_L	N_{Cat}	$\overline{N_L}$
MIMIC-III-Full	52,722	47,719	1,631	3,372	1,861 (949)	8,907	37	15.9
MIMIC-III-50	11,368	8,066	1,573	1,729	1,989 (968)	50	14	5.8

$\overline{N_d}$, (N_{d},σ) is the average number (standard deviation) of tokens per document, N_L is the total number of unique labels, N_{cat} is the total number of categories, and $\overline{N_L}$ is the average number of labels per document.

prediction, $C_T \in \mathbb{R}^{1\times d_h}$ (6) [5].

$$C_T(Q_T, K, V) = softmax\left(\frac{Q_T K^\top}{\sqrt{d_h}}\right) V \tag{6}$$

E. Output Layer

Given the generated context vector c_l, we compute the prediction probability of label l by using projection weights w_l and bias b_l of a linear layer, with $W \in \mathbb{R}^{|L|\times d_h}$ and $B \in \mathbb{R}^{|L|\times 1}$ (7), followed by the sigmoid activation function.

$$\hat{y}_l = act(w_l^\top c_l + b_l) \tag{7}$$

F. Training

We trained the model parameters by using the Adam optimizer and minimizing the the binary-cross-entropy loss for the multi-label experiments. We used a linear learning rate scheduler to stabilize training during the first five epochs. We implemented an early stopping mechanism to interrupt the training process once the performance on the validation dataset stops improving after ten epochs for the MIMIC-III notes [6]. The tunable model hyperparameters (e.g., batch size, learning rate) are shown in Section IV-B.

G. Dataset: MIMIC-III

MIMIC-III is a publicly available database that contains de-identified EHRs (e.g., clinical notes, vital signs, laboratory measurements) for over 40,000 patients; more detailed information on the clinical aspects of the dataset can be found in Johnson et al. [39]. Each unique patient ID is associated with at least one hospital admission ID. Every hospital admission ID was annotated by trained human encoders with a subset of the ICD-9 diagnosis and procedure codes. We formulated this problem as a multi-label document classification task. Therefore, our models aimed to predict a subset of labels, $y_i \in \{0, 1\}^l$, associated with document i from the entire label set $|L|$ of the set of N documents $D = \{(x_i, y_i)\}_{i=1}^N$ by utilizing only the discharge summary note and potentially available addenda associated with the respective hospital admission ID. We performed experiments on the entire dataset for the full label set (MIMIC-III-Full) and on a smaller subset that contained documents associated with the 50 most frequent ICD-9 codes (MIMIC-III-50). We split the data into training, testing, and validation sets by following the splits proposed by Mullenbach et al. [6]. We tokenized each sequence, lower-cased all tokens, and removed non-alphabetic tokens or tokens that occurred less than three times in the corpus. We set the maximum document length to 3,000 tokens for the CNNs and RNNs and 4,096 tokens for the CLF to account for the differences in information content provided by word-tokens used by CNNs and RNNs and subword-tokens used by the CLF. Table I presents relevant statistics for this dataset.

TABLE II
FOR MIMIC-III WE DENOTE THE SELECTED HYPERPARAMETERS FOR THE ARCHITECTURE-SPECIFIC METHODS FOR GENERATING DOCUMENT EMBEDDINGS AS B, TARGET-, AND LABEL-ATTENTION AS T AND L

CNN	
Filter Size	100^L, 300, 500^B, 1000^T
Word/Label Embedding Dimension	$100^{B,L}$, 200, 300^T
Batch Size	$16^{T,L}$, 32^B, 64, 128
Dropout %	0, 0.15^B, 0.3^T, 0.5^L
BiGRU	
Hidden Size	128, 256^L, $512^{B,T}$
Word/Label Embedding Dimension	$100^{T,L}$, 200, 300^B
Batch Size	$16^{B,L}$, 32, 64, 128^T
Dropout %	0, 0.15^T, 0.3^B, 0.5^L
BiLSTM	
Hidden Size	128, 256, $512^{B,T,L}$
Word/Label Embedding Dimension	100^B, 200^L, 300^T
Batch Size	$16^{B,T,L}$, 32, 64, 128
Dropout %	0, $0.15^{B,T}$, 0.3, 0.5^L
CLF	
Word/Label Embedding Dimension	$100^{B,L}$, 200, 300^T
Batch Size	$4^{T,L}$, 6^B, 8, 16
Dropout %	0^L, 0.15, 0.3, $0.5^{B,T}$
Learning Rate	1e-$5^{B,T}$, 2e-5, 5e-5^L

IV. EXPERIMENTS

A. Evaluation Metrics

The model performance was assessed with evaluation metrics commonly featured in the document classification literature [7]. We report micro- and macro-averaged F1-scores, micro- and macro-averaged scores for the area-under-curve (AUC), and precision at k for k in [5, 8, 15] for the classification experiments.

B. Hyperparameter Optimization

We performed hyperparameter optimization for all text-encoder architectures on the validation split of MIMIC-III-50 by using a hill-climbing strategy to alleviate the computational burden; we utilized the found hyperparameters for the MIMIC-III-Full models. Table II lists and indicates the optimal hyperparameters for each architecture and dataset. The explored hyperparameters are based on previous work [40]. We selected a learning rate of $1e^{-4}$ for the CNN, BiGRU, and BiLSTM [40] and tuned the learning rate for the CLF with values provided by [34].

V. RESULTS

A. Results: MIMIC-III – Single-Head Label-Attention

Table III shows the performance scores for our single-head label-attention experiments on MIMIC-III-Full and MIMIC-III-50 and compares them with the performances of current benchmark models taken from the original studies. For MIMIC-

TABLE III

PERFORMANCE RESULTS ON MIMIC-III-FULL AND MIMIC-III-50 FOR THE ARCHITECTURE-SPECIFIC BASELINE (B), TARGET-ATTENTION MECHANISM (T), RANDOM (R), PRETRAINED (P), HIERARCHICAL RANDOM (HR), AND HIERARCHICAL PRETRAINED (HP) LABEL-ATTENTION MECHANISM ACROSS MULTIPLE TEXT-ENCODER ARCHITECTURES

Model	Attention Type	MIMIC-III-Full						MIMIC-III-50				
		AUC		F1		P@k		AUC		F1		P@k
		Macro	Micro	Macro	Micro	8	15	Macro	Micro	Macro	Micro	5
CAML [6]	R	0.895	0.986	0.088	0.539	0.709	0.561	0.875	0.909	0.532	0.614	0.609
DR-CAML [6]	P	0.897	0.985	0.086	0.529	0.690	0.548	0.884	0.916	0.576	0.633	0.618
MultiResCNN [41]	R	0.910	0.986	0.085	0.552	0.734	0.584	0.899	0.928	0.606	0.670	0.641
LAAT [7]	R	0.919	0.988	0.990	0.575	0.738	0.591	0.925	0.946	0.666	0.715	0.675
JointLAAT [7]	R	0.921	0.988	**0.107**	0.575	0.735	0.590	0.925	0.946	0.661	0.716	0.671
HyperCore [9]	P	**0.930**	**0.989**	0.090	0.551	0.722	0.579	0.895	0.929	0.609	0.663	0.632
D2SBERT [42]	R	—	—	—	—	—	—	—	—	0.629	0.686	—
EffectiveCAN [19]	R	0.921	**0.989**	0.105	0.581	0.755	0.604	0.920	0.945	0.668	0.717	0.664
HiLAT + ClinicalplusXLNet [8]	R	—	—	—	—	—	—	**0.927**	**0.950**	**0.690**	**0.735**	**0.681**
PLM-ICD* [43], [44]	R	0.926	**0.989**	0.104	**0.598**	**0.771**	**0.613**	0.917	0.938	0.663	0.705	0.657
CNN	B	0.827	0.977	0.021	0.379	0.619	0.476	0.836	0.888	0.454	0.558	0.571
	T	0.864	0.978	0.030	0.350	0.601	0.455	0.853	0.888	0.448	0.539	0.575
	R	0.862	0.979	0.039	0.463	0.642	0.491	0.891	0.922	0.565	0.640	0.628
	P	0.876	0.982	0.049	0.481	0.658	0.509	0.885	0.917	0.542	0.623	0.621
	HR	0.846	0.972	0.015	0.318	0.544	0.410	0.873	0.907	0.479	0.573	0.601
	HP	0.882	0.981	0.045	0.443	0.629	0.481	0.881	0.916	0.534	0.625	0.620
Bi-GRU	B	0.883	0.982	0.033	0.435	0.661	0.503	0.872	0.906	0.490	0.586	0.603
	T	0.883	0.982	0.038	0.438	0.647	0.493	0.855	0.892	0.452	0.552	0.581
	R	0.890	0.984	0.066	0.531	0.702	0.548	0.871	0.907	0.519	0.622	0.606
	P	0.887	0.984	0.068	0.535	0.706	0.551	0.879	0.915	0.552	0.643	0.624
	HR	0.898	0.984	0.063	0.526	0.699	0.545	0.870	0.906	0.526	0.614	0.608
	HP	0.903	0.985	0.068	0.533	0.705	0.550	0.880	0.916	0.568	0.654	0.626
Bi-LSTM	B	0.879	0.982	0.026	0.377	0.614	0.464	0.837	0.879	0.377	0.479	0.529
	T	0.876	0.981	0.032	0.381	0.599	0.451	0.849	0.883	0.450	0.540	0.560
	R	0.893	0.985	0.067	0.538	0.712	0.557	0.873	0.906	0.495	0.590	0.596
	P	0.892	0.984	0.066	0.538	0.711	0.555	0.880	0.914	0.546	0.634	0.618
	HR	0.901	0.985	0.061	0.522	0.699	0.545	0.866	0.901	0.494	0.584	0.594
	HP	0.911	0.986	0.067	0.537	0.709	0.555	0.880	0.914	0.546	0.634	0.618
CLF	B	0.867	0.977	0.025	0.383	0.612	0.462	0.840	0.879	0.437	0.541	0.565
	T	0.879	0.980	0.030	0.407	0.613	0.462	0.888	0.917	0.558	0.635	0.628
	R	0.903	0.985	0.057	0.524	0.724	0.571	0.898	0.925	0.575	0.644	0.640
	P	0.915	0.987	0.069	0.550	***0.741***	***0.589***	***0.905***	***0.930***	***0.597***	***0.664***	***0.646***
	HR	0.865	0.978	0.024	0.359	0.611	0.466	0.888	0.915	0.565	0.635	0.624
	HP	**0.930**	***0.988***	***0.070***	***0.552***	0.738	0.586	0.896	0.924	***0.597***	0.660	0.637

* Original authors [43] only provided results for MIMIC-III-Full; for completion showing reproduced results of MIMIC-III-50 [44].
Models: Convolutional Neural Network (CNN), bi-directional gated recurrent unit neural network (Bi-GRU), bi-directional long short-term memory neural network (Bi-LSTM), and transformer-based Clinical Longformer (CLF).

We show performance results for previous and current state-of-the-art models directly taken from their work. Best overall performances are highlighted in bold; best performances of our models are bold and italicized.

III-Full, the results indicate that label-attention substantially improves classification performance over both baselines across all four text-encoder architectures. However, hierarchical random label-attention underperforms both baselines for the CNN- and CLF-based models. Models using pretrained reference information performed on average 0.014 points better than their randomly initialized counterparts on the macro-averaged F1-score across all text-encoder architectures. As expected, both RNNs are less sensitive in performance to different query matrix initializations than the CNN- and CLF-based models. Unexpectedly, incorporating code hierarchy via our hierarchical label-attention did not significantly boost model performance. Hierarchical pretrained label-attention performed better than hierarchical random label-attention. Overall, CLF experienced the largest boost in performance with label-attention, followed by the Bi-LSTM and Bi-GRU, and the CNN benefited the least from label-attention. The CLF with hierarchical pretrained label-attention achieved the best performance across all our models with macro- and micro-averaged F1-scores of 0.070 and 0.552, respectively.

For MIMIC-III-50, the results suggest that random, pretrained, and hierarchical pretrained label-attention perform significantly better than the baseline methods across all text-encoder architectures except for the CLF-based model with random hierarchical label-attention. In contrast to MIMIC-III-Full, the initialization of the reference information for the smaller dataset can have a significant impact on model performance and suggests that selecting random or pretrained embeddings depending on the given text-encoder architecture optimizes performance. Context-based text-encoder architectures (e.g., Bi-GRU, Bi-LSTM, and CLF) achieved their best performances when using label-attention and hierarchical label-attention with pretrained embeddings. In contrast, CNN-based models had their best result when using random label-attention because of potential synergies with learning multiple n-gram patterns. Hierarchical pretrained label-attention achieved equal or better performance than pretrained label-attention across all text-encoder architectures, and the performance difference for the Bi-GRU is significant. Models that use hierarchical random label-attention performed worse than the remaining types of label-attention.

TABLE IV

FREQUENCY COUNT OF LABELS THAT ACHIEVE THEIR BEST PERFORMANCE WITH EITHER RANDOM (R), PRETRAINED (P), HIERARCHICAL RANDOM (HR), OR HIERARCHICAL PRETRAINED (HP) LABEL-ATTENTION MECHANISMS BROKEN DOWN BY FREQUENCY QUARTILE ACROSS ALL TEXT-ENCODER ARCHITECTURES FOR MIMIC-III-FULL

Encoder	Label	Label Frequency Count			
	Attention	Q2	Q3	Q4	Total
CNN	R	2	19	217	238
	P	2	79	629	710
	HR	0	0	5	5
	HP	10	166	424	600
Bi-GRU	R	5	74	364	443
	P	8	79	451	538
	HR	4	89	289	382
	HP	5	65	376	446
Bi-LSTM	R	7	84	433	524
	P	6	81	459	546
	HR	2	52	198	252
	HP	0	86	429	515
CLF	R	0	55	189	244
	P	1	85	641	727
	HR	0	5	17	22
	HP	0	98	639	737
Number of Correctly Predicted Labels:		30	452	1,752	2,234
Total Number of Predicted Labels:		2,437	2,312	2,248	8,907
Label Frequency Range:		2–5	6–27	28–20,046	

The counts only consider labels for which at least one attention type achieved a nonzero performance across all models; no label in Q1 achieved nonzero performance, and thus Q1 is omitted. Different model combinations may correctly classify a different subset of labels.

Notably, the Bi-GRU performed significantly better than the Bi-LSTM for the RNN-specific baseline method by using the averaged hidden states over all time steps as the input for the classification layer. Overall, CLF-based models achieved the highest performance, followed by the Bi-GRU, and last with similar performance the CNN and Bi-LSTM. The CLF-based model that uses pretrained label-attention scored the best performance with a macro- and micro-averaged F1-score of 0.597 and 0.664, respectively. The performance discrepancies between our best models and the current state-of-the-art model stem from the differences in the applied text-encoder architectures. For example, the HiLAT [8] first splits each clinical document into smaller equally long chunks allowing it to use a transformer-based text-encoder architecture with dense self-attention to encode each chunk, whereas our best CLF-model is designed to encode the entire sequence at once using sparse self-attention. Sparse self-attention is more likely to underperform compared to dense self-attention but provides computational benefits like memory usage. By reducing the effective document, the HiLAT is able to learn more fine-grained semantic information from the clinical document, whereas, our CLF-based model makes long-range connections, potentially neglecting local semantic information. We re-emphasize that for the purpose of this study, we focus on more fundamental architectures, which is why we chose the more straightforward CLF as our Transformer baseline rather than a more complicated architecture such as the HiLAT.

For MIMIC-III-Full, we counted the number of times a specific type of label-attention achieved the highest performance for all labels in the label space. We then examined how the strategy used to initialize label-attention relates to the text-encoder architecture. Table IV summarizes the retrieved counts per frequency quartile for all labels for which at least one attention type achieved a nonzero performance; no label in Q1 achieved nonzero performance across all evaluated models, and thus Q1 is omitted. The results indicate that i) all models experience difficulties in predicting minority classes (e.g., only 1% of labels in Q2 were predicted correctly), ii) different text-encoder architecture and label-attention type combinations can learn to correctly classify a different subset of labels, iii) hierarchical pretrained label-attention helps CNN- and CLF-based models classify minority classes better (e.g., Q2, Q3), and iv) the majority of labels achieve their best performance with models utilizing either hierarchical pretrained or pretrained label-attention mechanisms.

B. Results: MIMIC-III – Multi-Head Label-Attention

We performed extensive experiments on a multi-head version of the label-attention mechanism. Table V lists the micro- and macro-averaged F1-scores of our experiments for both MIMIC-III datasets. For both datasets, all text-encoder architectures that used multi-head label-attention experienced significant performance improvements over their single-head versions. Significant performance improvements were predominantly achieved by using multi-head attention with either two or four attention heads. Additionally, hierarchical random label-attention experienced the strongest performance increase from using multi-head attention compared with the remaining types of label-attention. We believe this significant performance increase stems from each attention head attending to its own set of features of the input, thereby allowing the model to learn more important parts more easily. For MIMIC-III-Full, multi-head label-attention significantly improved the macro-average scores across all text-encoder architectures with the strongest performance of 0.092 achieved by the Bi-GRU using random multi-head label-attention. This indicates that multi-head label-attention can improve the classification of minority classes, which is beneficial for the identification of rare diseases. Further, the micro-averaged F1-scores for the Bi-GRU-based models were unaffected by the implementation of multi-head attention, whereas the Bi-LSTM-, CNN-, and CLF-based models experienced significant performance improvements. The CLF-based model that used hierarchical pretrained multi-head label-attention performed slightly better than the single-head version with a micro-averaged score of 0.554. For MIMIC-III-50, the Bi-LSTM benefited tremendously from using multi-head attention and showed significant improvements over the single-head version for random, hierarchical random, and hierarchical-pretrained label-attention across all numbers of attention heads. Multi-head attention had less impact on the performance of CLF-based models compared with the remaining three encoders. Our results suggest that the ideal number of attention heads depends on the architecture used, as well as trade-off considerations between improvement gains and compute requirements.

Our results on MIMIC-III revealed the following: i) label-attention generally improves classification performance over architecture-specific baselines and target-attention methods, ii) using label-attention with pretrained embeddings

TABLE V
MULTI-HEAD LABEL-ATTENTION RESULTS ON MIMIC-III

Dataset	Encoder	Heads	Random		Pretrained		Hierarchical Random		Hierarchical Pretrained	
			Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
MIMIC-III-Full	CNN	2	0.087**	0.510**	0.082**	0.522**	0.079**	0.509**	0.081**	0.516**
		4	0.004	0.187	0.078**	0.522**	0.077**	0.509**	0.081**	0.514**
		6	0.027	0.313	0.027	0.394	0.072**	0.500**	0.070*	0.510**
	Bi-GRU	2	**0.092****	0.522	0.063	0.506	0.087**	0.520	0.088**	0.521
		4	0.087*	0.523	0.073*	0.530	0.077**	0.509	0.080**	0.518
		8	0.083*	0.517	0.072*	0.526	0.064*	0.509	0.074**	0.522
	Bi-LSTM	2	0.080**	0.542*	0.079**	0.548**	0.087**	**0.554****	0.081**	0.544*
		4	0.082**	0.540*	0.068*	0.539*	0.084**	0.547**	0.074*	0.540*
		8	0.081*	0.535	0.068*	0.534	0.064*	0.526*	0.068*	0.532
	CLF	2	0.064*	0.535*	0.066	0.548	0.050**	0.513**	0.072*	**0.554***
		4	0.063*	0.534*	0.066	0.548	0.065**	0.535**	0.071*	0.553*
		8	0.035	0.460	0.051	0.523	0.072**	0.547**	0.075*	0.544
MIMIC-III-50	CNN	2	0.566*	0.646*	0.556**	0.638*	0.556**	0.639**	0.550**	0.634**
		4	0.558	0.638	0.549*	0.631*	0.563**	0.641**	0.542*	0.628*
		6	0.553	0.636	0.542	0.626*	0.537**	0.630**	0.528	0.621
		10	0.545	0.631	0.534	0.620	0.539**	0.630**	0.536*	0.625
	Bi-GRU	2	0.559**	0.637**	0.547	0.635	0.581**	0.650**	0.567	0.645
		4	0.550**	0.638**	0.554*	0.643	0.580**	0.646**	0.577*	0.652
		8	0.530*	0.620	0.565*	0.645*	0.573**	0.644**	0.562	0.646
		16	0.521*	0.612	0.554*	0.639	0.555**	0.635**	0.555	0.638
	Bi-LSTM	2	0.558**	0.642**	0.549*	0.640	0.558**	0.646**	0.563**	0.651**
		4	0.542**	0.628**	0.564**	0.649**	0.564**	0.647**	0.557**	0.645**
		8	0.543**	0.630**	0.559*	0.644*	0.553**	0.640**	0.553**	0.644**
		16	0.534**	0.617**	0.543	0.635*	0.543**	0.627**	0.547*	0.638*
	CLF	2	0.580*	0.645*	0.592	0.664	0.575*	0.640*	**0.600***	**0.669***
		4	0.569	0.643	0.588	0.656	0.582*	0.649*	0.593*	0.661*
		8	0.561	0.640	0.585	0.651	0.573*	0.639*	0.579	0.653
		16	0.557	0.631	0.582	0.652	0.570*	0.638*	0.571	0.646

Best performance across all models is highlighted in bold, * indicates improved performance over the single-head label-attention counterpart; ** indicates significant improvement.

is preferable over random embeddings for contextual-based text-encoders (e.g., RNNs and transformers), iii) using the transformer-based CLF achieves the best performance at the cost of a significant increase in compute time (i.e., GPU-usage), iv) the Bi-GRU outperforms the Bi-LSTM on aggregate, v) incorporating code hierarchy via label-attention can significantly improve model performance for biomedical text classification tasks and for the Bi-GRU and CLF in particular, and vi) multi-head label-attention has a strong impact on the classification performance, of the CNNs and RNNs but only a marginal affect on CLF-based models.

VI. DISCUSSION

A. Comparison of Attention and Energy Scores

Modern attention mechanisms consist of two primary operations: i) the matrix multiplication between the input sequence in K and the reference information in Q, which results in the energy scores, and ii) the softmax operation, which produces the attention scores by normalizing the raw energy scores. Both energy and attention scores are essential to direct a model's focus to the most relevant tokens associated with a medical concept; however, both sets of scores differ in their interpretation depending on the nature of the desired analysis.

The primary difference between energy scores and attention scores lies in the interpretation of the output of the deployed mathematical operations. The matrix multiplication allows us to interpret the magnitude and sign of the output as the strength and direction of the association between the input in K and the label-specific reference information in Q. In this way, the output, or energy score, provides a global view of how related the token is to a medical concept regardless of its current document or corpus. In contrast, the softmax function output can be interpreted as pseudo-probabilities that indicate the local importance of a token for predicting a positive association relative to all other words in a document. The final attention score is therefore influenced by the relevance of all other tokens and the total length of a document [45]. While attention scores help the model discriminate between signal and noise in a specific document, the softmax operation retrieves the underlying global association between a token and the label-specific reference information initially captured by the energy score and, ultimately, with the associated medical concept.

Consider the following two energy score sequences with three tokens $E1 = [-1.0, -2.0, -3.0]$ and $E2 = [3.0, 2.0, 1.0]$. After applying the softmax operation, we get identical attention scores $A1 = A2 = [0.665, 0.245, 0.090]$ for both energy score sequences despite having different initial energy scores. This example illustrates that a token may appear relevant locally within the context of a document but may possess a negative underlying global association with the medical concept that the token is trying to explain. This ambiguity is critical because the medical profession's acceptance of neural networks as a diagnostic tool depends on the traceability of a model's reasoning for its predictions [46].

Lastly, our work explores how different initialization strategies of the reference information in Q impact model performance and a model's motivation behind its predictions. These differences in Q across the label-attention types can be directly observed in the generated energy scores but not in the attention scores, making energy scores the ideal candidate for understanding the effect of different attention architectures and initialization strategies. However, we want to emphasize that in clinical practice the combination of both sets of scores are essential for sufficiently interpreting a model's predictions.

B. Random vs. Pretrained Label Reference Information

This study investigates the impact of an attention mechanism's initialization strategy on classification performance and potential differences between the initialization of an attention mechanism's query matrix with random or pretrained embeddings across various text-encoder architectures. Our results in Section V indicate observable differences between both strategies for smaller datasets such as MIMIC-III-50, whereas the differences become negligible for larger datasets.

Randomly initialized models try to learn label-specific reference information from the available samples during training. This method relies on the training corpus having a diverse set of documents to provide information about the entire label space; however, the label spaces of clinical text classification tasks are often dominated by a small but frequently occurring set of labels which inhibits the learning of representative features for infrequent labels. This imbalanced representation of learned features can sometimes be addressed by incorporating prior knowledge about the label space, such as the initialization of an attention mechanism's query matrix with pretrained embeddings of textual code descriptions. The purpose of using these pretrained embeddings is to prime the model with information relevant to each label and to push model performance. For example, if a document has the ICD-9 code *427.31 - Atrial Fibrillation* assigned to it, then we expect the most relevant tokens or phrases to be similar to *atrial*, *fibrillation*, or a combination of both. We think the encodings of relevant tokens will be closer to the label-specific query vector in the embedding space, thereby resulting in larger energy scores. Thus, we hypothesize that the choice of text-encoder architecture and type of label-attention directly affect the resulting energy scores on a token level. To investigate this hypothesis, we retrieved the energy score distributions from the test corpus of MIMIC-III-50 for all label-attention types across the four text-encoder architectures (Fig. 3).

We can infer from Fig. 3 the following differences between the energy score distributions for the label-attention mechanisms: i) RNN- and CLF-based models are less sensitive in their energy score distributions to differently initialized query matrices (i.e., reference information) compared with the CNN-based model. We think using the contextual-based encoding of clinical documents allows RNNs and transformers to generalize better and produce more similar token-vector representations, and this increases robustness for the computation of the energy scores. In contrast, CNNs learn embeddings from multiple n-grams disconnected from the context of the document, and this results in a more diverse set of token-vector representations that affect the computation of the energy scores. ii) The majority of energy scores are negative across all label-attention and text-encoder architecture combinations, except for CNN when using random label-attention. This is reasonable because only a minority of tokens are indicative of medical classifications in clinical text classification tasks. iii) Label-attention mechanisms with pretrained embeddings have larger extreme values and a larger spread for their energy scores compared with their respective randomly initialized counterparts. This is an artifact of using the dot product between a token representation and the embedded, label-specific reference information used to calculate the energy score for each token. Essentially, the model can more confidently distinguish between relevant tokens and noise, thereby improving its discriminatory ability.

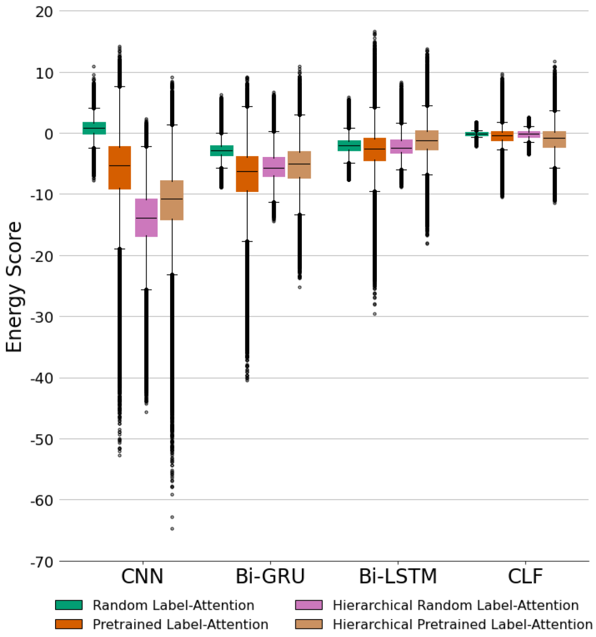

Fig. 3. Differences between the energy score distributions for all types of label-attention and text-encoder architectures extracted from the MIMIC-III-50 test documents. The dots are outliers and represent token-specific energy scores that lie beyond $\pm 1.5IQR$.

As mentioned above, prior knowledge can be used to address the label imbalance issues commonly found in clinical text classification tasks to improve the prediction performance of rare classes. We retrieved the model performance broken down by label frequency quartile to investigate if pretrained embeddings can improve prediction performance on minority labels for MIMIC-III-Full. The results in Table VI show that models overall and across all text-encoder architectures have improved performance on minority labels (i.e., Q2 and Q3) when using hierarchical pretrained or pretrained label-attention compared against their respective randomly initialized counterpart. This indicates that using pretrained embeddings in an attention mechanism helps the model learn meaningful features used to search for tokens or phrases relevant to the model.

TABLE VI

MODEL PERFORMANCE PER QUARTILE WITH RANDOM (R), PRETRAINED (P), HIERARCHICAL RANDOM (HR), OR HIERARCHICAL PRETRAINED (HP) LABEL-ATTENTION ORGANIZED BY FREQUENCY QUARTILE AVERAGED AND ACROSS ALL TEXT-ENCODER ARCHITECTURES FOR MIMIC-III-FULL

Encoder	Label Attention	Micro F1-Score Q2	Q3	Q4
Average	R	0.004	0.048	0.532
	P	0.006	0.058	0.545
	HR	0.002	0.030	0.447
	HP	0.005	0.061	0.536
CNN	R	0.002	0.023	0.485
	P	0.002	0.049	0.504
	HR	0	0.002	0.331
	HP	0.006	0.061	0.471
Bi-GRU	R	0.005	0.068	0.547
	P	0.012	0.072	0.552
	HR	0.002	0.058	0.543
	HP	0.011	0.072	0.550
Bi-LSTM	R	0.008	0.062	0.555
	P	0.009	0.060	0.555
	HR	0.004	0.055	0.539
	HP	0.004	0.060	0.554
CLF	R	0	0.037	0.541
	P	0.001	0.050	0.567
	HR	0	0.006	0.373
	HP	0	0.054	0.569

No model was able to predict labels in Q1, and thus Q1 is omitted.

TABLE VII

MODEL INTERPRETABILITY WITH LABEL-ATTENTION FOR TWO ICD-9 CODES

276.2: *Acidosis*

E:		1.02	1.30	1.20	2.39	3.04	4.81	6.32
A:		0.0008	0.0011	0.0010	0.0033	0.0063	0.0372	0.1687
T:	. . .	*patient*	*continued*	*to*	*have*	*raising*	***lactic***	***acidosis.***

E:	4.40	3.35	2.38	1.97	
A:	0.0246	0.0086	0.0033	0.0022	
T:	*PT*	*was*	*ultimately*	*found*	. . .

518.81: *Acute Respiratory Failure*

E:		3.69	3.85	4.19	5.43	6.35	6.61	5.95
A:		0.0029	0.0034	0.0048	0.0168	0.0419	0.0544	0.0280
T:	. . .	*on*	*floors*	*for*	***hypoxic***	***respiratory***	***failure.***	*This*

E:	6.08	6.20	5.37	4.46	3.49	2.65	2.77	
A:	0.0321	0.0361	0.0157	0.0063	0.0024	0.0010	0.0012	
T:	***respiratory***	***failure***	***was***	*quickly*	*reversed*	*with*	*diuresis*	. . .

The example text (T) snippets were taken from two evaluated test documents using the Bi-GRU with hierarchical pretrained label-attention; phrases marked as important are highlighted in bold. The numbers above each word represent their respective energy (E) or attention (A) scores.

TABLE VIII

EXPERT EVALUATION OF THE MOST-RELEVANT LABEL-SPECIFIC KEY PHRASE PER FREQUENCY QUARTILE FOR THE RANDOM (R), PRETRAINED (P), HIERARCHICAL-RANDOM (HR), AND HIERARCHICAL-PRETRAINED (HP) LABEL-ATTENTION MECHANISMS

		ICD-9 Codes*																
		Q1				Q2				Q3				Q4				
Encoder	Attention	93.90	45.13	V15.82	287.5	507.0	276.1	V58.61	36.15	38.91	276.2	530.81	39.61	599.0	518.81	38.93	401.9	Total
		Label Frequency →																
CNN	R	✓							✓			✓	✓	✓	✓		✓	7
	P				✓		✓		✓	✓	✓		✓	✓		✓		8
	HR				✓		✓		✓	✓	✓		✓	✓		✓		8
	HP				✓		✓		✓	✓	✓		✓	✓		✓	✓	9
Bi-GRU	R		✓		✓		✓	✓	✓	✓	✓		✓	✓				9
	P		✓		✓	✓	✓	✓	✓		✓		✓	✓	✓		✓	11
	HR			✓			✓		✓		✓		✓	✓	✓		✓	8
	HP		✓		✓	✓	✓	✓	✓		✓		✓	✓			✓	10
Bi-LSTM	R			✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓		11
	P				✓	✓	✓	✓	✓		✓		✓	✓	✓		✓	10
	HR			✓	✓		✓		✓	✓	✓	✓	✓	✓		✓		10
	HP				✓	✓	✓	✓	✓		✓		✓	✓	✓		✓	10
CLF	R	✓							✓			✓	✓	✓	✓	✓	✓	8
	P	✓		✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	13
	HR				✓	✓	✓		✓		✓	✓	✓	✓	✓		✓	10
	HP	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	15

*The top two most- and least-frequent ICD-9 codes per frequency quartile were extracted from the MIMIC-III-50 test corpus.

Results are presented across four text-encoder architectures. A ✓ indicates a strong association of a key phrase with an ICD-9 code as deemed by the medical expert.

C. Model Interpretability: Phrase-Level Evaluation

Showcasing a model's ability to provide interpretability helps build trust in clinicians in AI-driven clinical support systems and improve patient outcomes [47]. To investigate the general capability of models that use label-attention to provide such evidence and potential differences between the various label-attention types on a global level, we extracted the most relevant phrase per code across all MIMIC-III-50 test documents. Specifically, for each model, we extracted the key phrase, kp, with the largest average energy score out of a pool of phrases with different sizes from the entire set of label-specific energy score sequences, E, (8):

$$kp = \max\left(\frac{1}{s}\sum_{i=j}^{j+s-1} e_i\right) \qquad i : [0, N], \tag{8}$$

where e_i is the energy score of the ith token in the sequence with N tokens, and s is the phrase size with either i) three, four, or five tokens for the RNNs/CNNs or ii) four, six, or eight tokens for the CLFs. The phrase size differences account for the reduced information provided by the CLFs' subword-tokens.

Table VII provides examples for extracted text snippets containing highly relevant phrases including their word-specific energy scores for the prediction of ICD-9 codes, e.g., *lactic acidosis* for the ICD-9 code *276.2 Acidosis*. A clinician can use the energy scores in tandem with the respective attention scores to interpret the prediction of a medical code by answering the following two questions: i) How strong is the global relationship between a word and the medical concept? ii) How important is the same word relative to all other words in the clinical document? The first question can be answered using the energy scores where high energy scores indicate that particular word is universally associated with a medical concept across all documents in the corpus, e.g., the model is able to associate *hypoxic*, a term describing the deficiency of oxygen in tissue, with the ICD-9 code *518.81 Acute Respiratory Failure*. The latter question can be answered using the attention scores where larger scores represent higher importance within that particular document. See Section VI-A for a more detailed discussion on the differences between energy and attention scores.

A medical professional evaluated a total of 256 extracted phrases for their associations (3 - Strong Association, 2 - Somewhat Associated, and 1 - No Association) with the respective ICD-9 codes. The results of this evaluation are shown in Table VIII and indicate differences in evidence-based interpretability between the text-encoder architectures and between the label-attention mechanisms. On the text-encoder architecture level, we see that context-based text-encoder architectures are more likely to provide immediate evidence across a larger subset of ICD-9 codes compared with the n-gram-based CNN; this difference is more prominent for infrequent codes associated

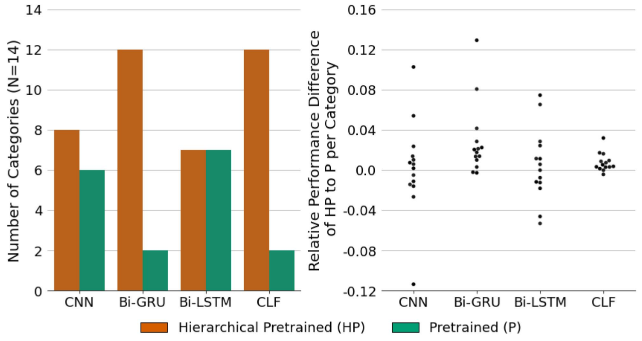

Fig. 4. Achievable performance improvements per category of hierarchical pretrained label-attention (HP) over pretrained label-attention (P) for CNN, Bi-GRU, Bi-LSTM, and CLF. Aggregated number of categories achieving their best performance with hierarchical pretrained or pretrained label-attention, respectively (left). The relative performance difference is computed by subtracting P from HP (i.e., each point above 0 represents cases in which HP performs better. Points below 0 represent cases in which P performs better). Each statistic was computed as the aggregated, macro-averaged F1-score of all associated labels with their parent category (right).

with Q1. We think that the CNNs' design causes the model to put emphasis on phrases that help the model to better discriminate between the codes but are not meaningful to humans. For the CNN, Bi-GRU, and the CLF, the pretrained label-attention mechanisms are more likely to provide evidence for a larger set of medical codes compared with their respective randomly initialized counterparts. We think that initializing attention mechanisms with pretrained embeddings helps the model look for phrases similar to the code description. For example, code 36.15 *Single internal mammary-coronary artery bypass* models often assigned the highest importance to phrases such as *coronary artery disease* for the CNN or *coronary artery bypass graft* for the CLF. However, the results indicate that both pretrained label-attention mechanisms assign high relevance to specific phrases across all labels. For example, the phrase *potassium chloride meq* was indicated as highly relevant across all labels while only being positively associated with a subset of codes. The retrieval of the same phrase as important across labels with opposing associations may indicate that label-attention is prone to propagating similar reference information across the different label-specific queries.

D. Attention to Code Hierarchy

This study proposes an extension of label-attention that can incorporate the hierarchical structure of a medical encoding system (e.g., ICD-9) without needing to train an auxiliary network. Contemporary work often deploys auxiliary networks (e.g., GNNs) to learn the code hierarchy and infuse the primary model with information on the hierarchical relationships between the labels to improve predictions [22]. In contrast, our hierarchical label-attention method incorporates hierarchical relationships by learning a set of features associated with high-level categories for each document that are then propagated downstream to enrich the low-level label reference information (i.e., each label query is enriched with the features of its parent category). We hypothesize that our category-wise incorporation of hierarchical information will lead to performance improvement on a category level. To test this hypothesis, we retrieved the label-specific performance and computed an aggregated, macro-averaged F1-score for all 14 categories associated with the MIMIC-III-50 classification task. The results are shown in Fig. 4. For the CNN, Bi-GRU, and CLF, the majority of categories exhibit improved performance. In particular, the Bi-GRU- and CLF-based models strongly benefited from deploying hierarchical label-attention. We showed that enriching the label-specific reference information with high-level categorical information leads to improved classification performance overall (see Section V) and on a category level. The analysis also demonstrates that the performance gains are equally distributed across all categories.

VII. Conclusion

In this work, we compared two strategies to initialize the reference information of the label-attention's query matrix across four text-encoder architectures, CNN, Bi-GRU, Bi-LSTM, and the transformer-based CLF, on the two common MIMIC-III clinical text classification tasks. We compared these methods against common methods to generate document embeddings and against target-attention. We showed that utilizing pretrained reference information in an attention mechanism's query matrix improves classification performance across smaller label spaces, but the performance increase is negligible with an increasing number of labels and documents. Additionally, we demonstrated that performance improvements are achievable by incorporating the hierarchical structure of medical encoding systems via the attention mechanism without requiring a second external network. Furthermore, we showed that label-attention mechanisms can provide highly relevant phrases allowing clinicians to interpret the prediction of a specific medical code. We expect our research to be helpful for future work that explores label-attention and pretrained reference information to maximize the potential of deep-NLP models for the automated classification of clinical documents with medical codes. The source code is available at https://github.com/cmetzner93/attention_mechanisms.

Acknowledgment

Publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). This research used resources of the Oak Ridge Leadership Computing Facility and the Compute and Data Environment for Science at ORNL.

References

[1] I. Li et al., "Neural natural language processing for unstructured data in electronic health records: A review," *Comput. Sci. Rev.*, vol. 46, 2022, Art. no. 100511.

[2] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh, "A systematic literature review of automated clinical coding and classification systems," *J. Amer. Med. Informat. Assoc.*, vol. 17, no. 6, pp. 646–651, 2010.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016, *arXiv:1409.0473*.

[4] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[5] S. Gao et al., “Classifying cancer pathology reports with hierarchical self-attention networks,” *Artif. Intell. Med.*, vol. 101, 2019, Art. no. 101726.

[6] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, “Explainable prediction of medical codes from clinical text,” in *Proc. Conf. North Amer. Chap. Assoc. Comput. Linguistics: Human Lang. Technol.*, New Orleans, Louisiana, vol. 17, pp. 1101–1111, 2018, doi: 10.18653/v1/N18-1100.

[7] T. Vu, D. Q. Nguyen, and A. Nguyen, “A label attention model for icd coding from clinical text,” in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, doi: 10.24963/ijcai.2020/461.

[8] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett, and L. Jorm, “Hierarchical label-wise attention transformer model for explainable icd coding,” *J. Biomed. Inform.*, vol. 133, 2022, Art. no. 104161.

[9] P. Cao, Y. Chen, K. Liu, J. Zhao, S. Liu, and W. Chong, “Hypercore: Hyperbolic and co-graph representation for automatic ICD coding,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3105–3114.

[10] I. Chalkidis, M. Fergadiotis, S. Kotitsas, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “An empirical study on large-scale multi-label text classification including few and zero-shot labels,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 7503–7515.

[11] S. Ji, W. Sun, H. Dong, H. Wu, and P. Marttinen, “A unified review of deep learning for automated medical coding,” 2023, *arXiv:2201.02797*.

[12] P. Xie and E. Xing, “A neural architecture for automated ICD coding,” in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, doi: 10.18653/v1/p18-1098.

[13] S. Kardakis, I. Perikos, F. Grivokostopoulou, and I. Hatzilygeroudis, “Examining attention mechanisms in deep learning models for sentiment analysis,” *Appl. Sci.*, vol. 11, no. 9, 2021, Art. no. 3883.

[14] W. Quan, Z. Chen, J. Gao, and X. T. Hu, “Comparative study of CNN and LSTM based attention neural networks for aspect-level opinion mining,” in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 2141–2150.

[15] O. Firat, K. Cho, and Y. Bengio, “Multi-way, multilingual neural machine translation with a shared attention mechanism,” 2016, *arXiv:1601.01073*.

[16] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2016, pp. 1480–1489.

[17] S. Jain, R. Mohammadi, and B. C. Wallace, “An analysis of attention over clinical notes for predictive tasks,” 2019, *arXiv:1904.03244*.

[18] M. Feucht, Z. Wu, S. Althammer, and V. Tresp, “Description-based label attention classifier for explainable ICD-9 classification,” 2021, *arXiv:2109.12026*.

[19] Y. Liu, H. Cheng, R. Klopfer, M. R. Gormley, and T. Schaaf, “Effective convolutional attention network for multi-label clinical document classification,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 5941–5953.

[20] G. Kurata, B. Xiang, and B. Zhou, “Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2016, pp. 521–526.

[21] S. Banerjee, C. Akkaya, F. Perez-Sorrosal, and K. Tsioutsiouliklis, “Hierarchical transfer learning for multi-label text classification,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6295–6300.

[22] H. Chen, Q. Ma, Z. Lin, and J. Yan, “Hierarchy-aware label semantics matching network for hierarchical text classification,” in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 4370–4379.

[23] J. Zhou et al., “Hierarchy-aware global model for hierarchical text classification,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1106–1117.

[24] L. Xu et al., “Hierarchical multi-label text classification with horizontal and vertical category correlations,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2459–2468.

[25] J. Shen, W. Qiu, Y. Meng, J. Shang, X. Ren, and J. Han, “Taxoclass: Hierarchical multi-label text classification using only class names,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 4239–4249.

[26] A. Galassi, M. Lippi, and P. Torroni, “Attention in natural language processing,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4291–4308, Oct. 2021.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” 2013, *arXiv:1301.3781*.

[28] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” 2020, *arXiv:2004.05150*.

[29] H. Lu, L. Ehwerhemuepha, and C. Rakovski, “A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance,” *BMC Med. Res. Methodol.*, vol. 22, no. 1, 2022, Art. no. 181.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751. [Online]. Available: https://aclanthology.org/D14-1181

[32] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] K. Cho et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proc. Conf. Emp. Methods Natural Lang. Process.*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[34] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo, “Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences,” 2022, *arXiv:2201.11838*.

[35] Z. Niu, G. Zhong, and H. Yu, “A review on the attention mechanism of deep learning,” *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[36] G. Brauwers and F. Frasincar, “A general survey on attention mechanisms in deep learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3279–3298, Apr. 2023.

[37] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[38] WHO, “ICD-9 code descriptions,” 2014. Accessed: Nov. 9, 2022. [Online]. Available: https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/Downloads/ICD-9-CM-v32-master-descriptions.zip

[39] A. E. Johnson et al., “MIMIC-III, A freely accessible critical care database,” *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[40] S. Gao et al., “Limitations of transformers on clinical text classification,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3596–3607, Sep. 2021.

[41] F. Li and H. Yu, “Icd coding from clinical text using multi-filter residual convolutional neural network,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8180–8187.

[42] T.-S. Heo, Y. Yoo, Y. Park, B. Jo, K. Lee, and K. Kim, “Medical code prediction from discharge summary: Document to sequence bert using sequence attention,” in *Proc. IEEE 20th Int. Conf. Mach. Learn. Appl.*, 2021, pp. 1239–1244.

[43] C.-W. Huang, S.-C. Tsai, and Y.-N. Chen, “PLM-ICD: Automatic ICD coding with pretrained language models,” in *Proc. 4th Clinical Natural Lang. Process. Workshop*, T. Naumann, S. Bethard, K. Roberts, and A. Rumshisky, Eds., Seattle, WA, 2022, pp. 10–20, doi: 10.18653/v1/2022.clinicalnlp-1.2.

[44] J. Edin et al., “Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study,” in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, New York, NY, USA, 2023, pp. 2572–2582, doi: 10.1145/3539618.3591918.

[45] O. P. Richter and R. Wattenhofer, “Normalized attention without probability cage,” 2022. [Online]. Available: https://openreview.net/forum?id=PeG-8G5ua3W

[46] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: Review, opportunities and challenges,” *Brief. Bioinf.*, vol. 19, no. 6, pp. 1236–1246, 2018.

[47] Z. Ahmad, S. Rahim, M. Zubair, and J. Abdul-Ghafar, “Artificial intelligence (AI) in medicine, current applications and future role with special emphasis on its potential and promise in pathology: Present and future impact, obstacles including costs and acceptance among pathologists, practical and philosophical considerations. a comprehensive review,” *Diagn. Pathol.*, vol. 16, pp. 1–16, 2021.