Multi-channel, Convolutional Attention based Neural Model for Automated Diagnostic Coding of Unstructured Patient Discharge Summaries[⋆]

Veena Mayya[a,*], Sowmya Kamath S[a], Gokul S Krishnan[a], Tushaar Gangavarapu[b,1]

[a] *Healthcare Analytics and Language Engineering (HALE) Lab, Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India*
[b] *Automated Quality Assistance (AQuA) Machine Learning Research, Kindle Content Quality, Books Org., Amazon.com, Inc.*

Abstract

Effective coding of patient records in hospitals is an essential requirement for epidemiology, billing, and managing insurance claims. The prevalent practice of manual coding, carried out by trained medical coders, is error-prone and time-consuming. Mitigating this labor-intensive process by developing diagnostic coding systems built on patients' Electronic Medical Records (EMRs) is vital. However, developing nations with low digitization rates have limited availability of structured EMRs, thereby demanding the need for systems built on unstructured data sources. Despite the rich clinical information available in such unstructured data, modeling them is complex, owing to the variety and sparseness of diagnostic codes, complex structural and temporal nature of summaries, and prolific use of medical jargon. This work proposes a context-attentive network to facilitate automatic diagnostic code assignment as a multi-label classification problem. The proposed model facilitates information aggregation across a patient's discharge summary via multi-channel, variable-sized convolutional filters to extract multi-granular snippets. The attention mechanism enables selecting vital segments in those snippets that map to the clinical codes. The model's superior performance underscores its effectiveness compared to the state-of-the-art on the MIMIC-III database. Additionally, experimental validation using the CodiEsp dataset exhibited the model's interpretability and explainability.

Keywords: Disease prediction, explainability, healthcare informatics,

 The final authenticated version is available online at: https://doi.org/10.1016/j.future.2021.01.013.

[*] Corresponding author.

Email addresses: mayya.veena@gmail.com (Veena Mayya), sowmyakamath@nitk.edu.in (Sowmya Kamath S), gsk1692@gmail.com (Gokul S Krishnan), tusgan@amazon.com (Tushaar Gangavarapu)

[1] Work done at the National Institute of Technology Karnataka.

interpretability, predictive analytics, unstructured text modeling.

1. Introduction

In hospitals, the International Statistical Classification of Diseases and Related Health Problems (ICD-9[2] [1] and ICD-10[3] [2]) medical coding taxonomy is widely employed to describe patients' clinical conditions and associated diagnoses. These classification systems are maintained by the World Health Organization, and several publicly available large healthcare datasets record instances of patient data mapped to ICD-9 and ICD-10 clinical procedure and diagnostic codes. ICD is essentially a hierarchical classification that defines unique codes for patient conditions, diseases, infections, symptoms, causes of injury, and others. These unique diagnostic codes are assigned to patient records to facilitate clinical and financial decisions made by the hospital management for various tasks, including billing, insurance claims, and reimbursements [3, 4]. Based on clinicians' free-text notes and other patient records such as discharge summaries, doctors' notes, nursing notes, and other relevant sources, trained professional medical coders employed by the Medical Records Department in hospitals transcribe patient records into a set of appropriate medical diagnostic codes (from a potentially large number of over $15,000$ codes). These medical coders utilize their medical domain expertise along with a plethora of coding rules and terminologies to facilitate patient-record-to-diagnostic-codes (one-to-many) mapping.

Given the enormous volume of patient records generated every day in urban and rural hospitals alike, such manual coding processes are highly cost-intensive and often inexact, time-consuming, and error-prone [5, 6]. Interestingly, the additional costs incurred due to inaccurate coding and the financial investment towards improving diagnostic coding efficacy is estimated to be more than $25 billion per year (in the United States alone) [7, 8]. Furthermore, automated systems reliant on Structured Electronic Medical Records (S-EMRs) find limited applicability in developing nations with relatively low digitization rates. It is crucial to develop intelligent computational systems that accommodate these needs by facilitating automated diagnostic coding of *unstructured* patient records. Such a code assignment can be regarded as a multi-label classification problem involving binary classification of multiple diagnostic labels, with each code label pertaining to a specific diagnostic condition (recorded as a binary indicator). Over the years, there has been a significant interest in developing and utilizing machine learning models to facilitate automated ICD coding as a

[2] `https://www.cdc.gov/nchs/icd/icd9cm.htm`.
[3] `https://icd.who.int/browse10/2019/en`.

multi-label classification task. Strategies and models utilizing Support Vector Machines (SVMs) [9, 10, 11], naïve Bayes [12, 13], nearest neighbors [14, 15], unsupervised topic modeling [16, 17], and several others have been employed for the clinical prediction task. Recent surveys on applications of deep learning approaches for the analysis of S-EMRs [18, 19, 20] highlight the need for interpretability of predictions made and explainability of automated prediction systems. By understanding the input features that contribute to the output decisions, trust can be built in the predictions and recommendations enabled by such learned models, which is crucial in healthcare applications. In this study, we attempt to dissect the black-box decisions facilitated by the proposed deep neural model by visualizing the associated clinical terms that contributed to the prediction of the respective disease code. We argue that such analyses and interpretation of the obtained predictions enhance the explainability of the proposed automated system.

More recently, research on automated code assignment has been attempted by modeling the unstructured clinical text [21, 4, 22, 23, 24, 25, 26, 6, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39], thus, exploring the richness of patient-specific information in such free-text. While supervised learning approaches are applicable in cases of accessible large-scale annotated datasets, it is not uncommon for researchers to explore modeling approaches that are beneficial in targeted studies with minimal data resources. In this regard, deep neural models and modeling strategies, including the *DeepLabeler* [4], Convolutional Networks (ConvNets) [27, 28, 29, 30], Long Short-Term Memory (LSTM) models [24], and transfer learning [6, 26], have been quite successful. However, the availability of healthcare clinical datasets is relatively abundant (e.g., PCORnet[4], Open NHS[5], eICU-Philips[6], MIMIC[7], VistA[8], ACS-NSQIP[9], and others [40]), owing to the volume of medical patient data generated day-to-day, thus promoting active healthcare research in modeling such data. Despite the data abundance, only a limited number of these data sources include unstructured text-based patient diagnosis data, such as discharge summaries and nursing notes. Most state-of-the-art studies have utilized the standard, openly-available MIMIC-III (Medical Information Mart for Intensive Care) database [41], comprising over $40,000$ patients' data. Several researchers [22, 24, 4, 26, 6, 27] have attempted to utilize the predictive power of machine and deep learning based models to enhance the diagnostic coding performance on the patient data available in

[4]https://pcornet.org/data-driven-common-model/.
[5]https://digital.nhs.uk/data-and-information/data-collections-and-data-sets.
[6]https://eicu-crd.mit.edu/.
[7]https://mimic.physionet.org/.
[8]https://www.data.va.gov/widgets/4d7k-fkpu.
[9]https://www.facs.org/Quality-Programs/ACS-NSQIP/joinnow/data.

the MIMIC-III database, making the database one of the most widely employed sources for performance benchmarking. An alternate dataset, CodiEsp, released as a part of the CLEF eHealth challenge [42, 43], contains explainability-specific annotated patient data by clinical experts. The CodiEsp dataset facilitates the exploration of the extent to which the proposed automated coding solution is interpretable and explainable, thus enabling the dissection of the black-box decisions output by the underlying neural system.

Existing studies [35, 32, 36, 27] facilitating automated ICD-based clinical coding corroborate the critical nature of the task at hand. Moreover, the applicability, deployability, and adaptability of the proposed intelligent systems in real-world scenarios demand high performance (exceeding that by the manual clinical coders), both in code prediction and system explainability. However, the nature of the underlying data poses several modeling challenges, including the variety and sparseness of diagnostic codes, complex structural and temporal nature of unstructured data, and prolific use of medical jargon, limiting the reported performance in the existing works. Thus, the problem of accurate ICD code assignment remains a long-standing open research challenge in the field of healthcare informatics and machine learning. To cope with the modeling complexities, specifically the vast imbalance in the code distribution across patient data, prior studies discarded medical records corresponding to less frequent diagnosis codes, thus reporting the performance on modeling the top$-k$ diagnostic procedures. Furthermore, several researchers and recent surveys on the use of deep neural approaches for patients' risk stratification [18, 19, 20] highlight the urgent need for the interpretability and explainability of the proposed automated prediction systems. In this study, we emphasize the significance of interpretable intelligent healthcare solutions in ensuring the trustworthiness of the underlying computational clinical decision support systems. In designing an automated explainable intelligent ICD coding system, this work proposes the Enhanced Convolutional Attention network for Multi-Label classification (*EnCAML*). The *EnCAML* model employs multi-channel, variable-sized convolution filters and multiple attention layers that reveal the associations of medical text with the predicted diagnostic code as a result of the interactions between the neurons.

To enable extensive performance benchmarking with state-of-the-art works detailing the automated ICD-based code assignment as a multi-label problem, we employed the MIMIC-III (v1.4) database. Additionally, to demonstrate the explainability and interpretability of the proposed neural model, we utilized the CodiEsp dataset. In line with the existing works, this study benchmarks the performance using (a) *top$-k$ diagnostic codes*, covering over 76.93% ($k = 10$) and 93.60% ($k = 50$) of the database, (b) *top$-k$ diagnostic code categories*, covering over 84.24% ($k = 10$) and 96.79% ($k = 50$) of the database, and (c) *all* 6, 918 *disease diagnosis codes*, corresponding to the discharge summaries of the patient cohort. Our extensive benchmarking across several variations in the cohort data selection (presented as (a), (b), and (c) above) facilitates detailed analysis of the obtained prediction performance, thus enabling recommendations on the targeted use of the proposed automated system. While benchmarking our performance on the CodiEsp dataset, we utilized the top−10 and top−50 most frequently utilized diagnostic codes to account for the limited corpus size. The results from our exhaustive experimentation revealed the superiority of the proposed approach over several state-of-the-art diagnostic code prediction models. Moreover, our analysis indicated the minimal impact of the initial embedding layer on the overall ICD code prediction performance, thereby corroborating the robustness and flexibility of the proposed *EnCAML* model. The key contributions of this work in advancing the efforts of the state-of-the-art can be summarized as follows:

- Design of *EnCAML*, a multi-channel, variable-sized convolution attention neural model that facilitates the reliable assignment of diagnostic codes using unstructured text-based patient discharge summaries, focusing on the interpretability and explainability of the neural system.
- Enable detailed analysis on the impact of the initial embedding layer on the overall performance of the proposed *EnCAML* model, using several state-of-the-art embedding approaches on voluminous discharge summaries. Our results reveal that the effect of the initial embedding layer on the overall performance is minimal, thus indicating the robustness and flexibility of the proposed *EnCAML* model.
- Present extensive benchmarking results that underscore the superior performance of the proposed *EnCAML* model compared to the current works on ICD-9 code prediction using MIMIC-III unstructured discharge summaries. Furthermore, we expand on the interpretability and explainability of the proposed system using our analysis on the CodiEsp dataset.

The rest of the paper is organized as follows: Section 2 presents a detailed overview of the related works discussing the prediction of clinical events and outcomes. Section 3 documents the methods employed for data extraction and preprocessing, while Section 4 details the specifics on the proposed *EnCAML* deep neural model employed in automated ICD code assignment. The proposed *EnCAML* model's overall benchmarking performance and a detailed discussion on the model's explainability are presented in Section 5. Finally, Section 6 summarizes this work with highlights on future research possibilities.

2. Related Work

Automated diagnostic coding of patient records is a field of active and extensive research interest, dating back to as early as the 1990s. Owing to the time-hallowed nature of the ninth version of the ICD coding system among the existing clinical datasets and hospitals alike, most of the existing works [44, 25, 22, 45, 23, 24, 26, 27, 28, 29, 30, 31, 36, 32, 35] reported their performance on ICD-9 code assignment. However, with the recent shift towards ICD-10 coding, certain works [5, 46, 38, 39] employed the much convoluted ICD-10 coding taxonomy. To enable extensive performance benchmarking and ensure rapid deployability of the proposed automated system, we utilize the more-established ICD-9 code taxonomy. The seminal work on automated ICD-9 code assignment by de Lima et al. [47] employed a cosine similarity between the term weighting vectors of text-based clinical notes and ICD-9 codes to facilitate the clinical task. Several significant studies on solving the task of automated ICD code assignment have emerged ever since. These works can be broadly classified as (a) rule-based systems [48, 49, 8], (b) primitive learning-based systems involving Bayesian classifiers, nearest neighbors, and relevance feedback [50, 51, 52], (c) advanced neural-learning-based systems [34, 35, 32, 33, 36], and (d) explainable intelligent systems [20, 27]. In this section, we present an overview of the existing works built on large healthcare datasets.

In a broader sense, rule-based systems mimic the approaches employed by trained clinical coders by using a set of handcrafted expert directives. These systems are heavily reliant on the knowledge of the medical professionals for the construction of rules and procedures, thus making it impractical to scale, given the wide variety of diseases and ever-increasing diagnostic codes. Despite their impracticality, rule-based systems draw up the decision trees, thus enabling extensive explainability of the predictions output by the automated system. Conversely, learning-based automated coding systems built to spontaneously learn patterns (virtual rules) from the underlying data ensure constant rule updation, thus accounting for such diagnostic coding systems' scalability. Such systems can be further categorized into feature-engineering-based learning systems [44, 53] and end-to-end, data-driven learning systems [4, 33, 35, 34, 32, 36]. Approaches reliant heavily on the input representation and the extraction of relevant features fall under the former category. However, over the years, research has shifted in favor of end-to-end, data-driven intelligent predictive systems built on deep neural models, owing to their time-aware predictive capabilities. Deep neural models have been shown to achieve promising results in modeling EMRs to facilitate a multitude of clinical prediction tasks, including mortality prediction [45, 54, 55, 56], chronic disease prediction [57, 58], length-of-stay estimation [45, 54, 59], hospital readmission prediction [60, 61, 62], disease phenotyping [45, 54, 63], precision medicine modeling [64], ICD-9 code group prediction [33, 35, 34, 32, 36], and automated ICD-9 coding [65, 23, 22, 27]. Furthermore, since neural models perform some sense of implicit feature selection, the need for external extensive feature engineering is minimized.

With the latest advancements and success in deep neural modeling, ConvNets have been utilized widely to facilitate the classification of various free-text documents [66, 58], including voluminous unstructured healthcare records [67, 68, 69]. Researchers have recently studied the significance of ConvNet-based methods for automated diagnostic code assignment based on free-text critical care discharge summaries [23, 4, 22, 27, 28, 29, 30, 31, 36, 35, 34]. In critical healthcare applications such as clinical decision support systems, trust is rooted in more than just their performance; such systems also need to justify and explain their actions based on the principles that present the dynamics of the concerned domain. In an attempt to develop explainable intelligent systems, current research aims to combine neural models such as ConvNets and recurrent networks with an attention mechanism [22, 27, 28, 29, 30, 31]. Baumel et al. [22] proposed a hierarchical neural attention model to discern relevant portions of a given free-text document that corresponded to a specific ICD-9 code label, based on which a deep neural Gated Recurrent Unit (GRU) was trained to enable the clinical task of automated coding. The *DeepLabeler*, designed by Li et al. [4], facilitates the assignment of ICD-9 codes to discharge summaries—the authors utilized ConvNets on Doc2Vec embeddings of the discharge summaries. Mullenbach et al. [27] proposed a convolutional attention network to facilitate multi-label classification of ICD-9 codes, advancing the field of explainable predictive systems. The authors benchmarked their prediction performance using $8,921$ unique ICD-9 codes, including $6,918$ diagnostic codes and $2,003$ procedural codes. To encode the hierarchy of ICD-9 codes and facilitate diagnostic coding, Xie and Xing [24] utilized LSTM networks with attention on the *diagnosis description* portion of the discharge summaries. Huang et al. [23] evaluated and benchmarked the performance of several existing deep neural models, including feed-forward neural networks, ConvNets, LSTMs, and GRUs, on patient discharge summaries, for the clinical prediction task of ICD-9 coding. Additionally, the authors also benchmarked their performance using traditional machine learning classifiers, including logistic regression and random forest.

Exploiting the nature of the problem at hand, Zeng et al. [6] employed transfer learning from indexing the medical subject headings to automated diagnostic coding. The authors utilized a ConvNet for the ICD-9 code prediction task and compared their performance against machine classifiers, including SVMs and flat-SVM models. Extending their work, Rios and Kavuluru [26] modified the initial transfer learning model to improve the predictive accuracy of top-10 ICD-9 codes. The investigation of modeling performance variations by initializing the embedding layer using pre-learned weights derived from various pre-trained word embedding models such as Word2Vec [70], fastText [71], and Bidirectional Encoder Representations from Transformers (BERT) [72] is vital to analyze the impact of healthcare document representations on the overall diagnostic code predictability. Guo et al. [37] and Mullenbach et al. [27] employed Word2Vec embeddings for the ICD-9 coding task. On the other hand, Huang et al. [23] experimented with both the Word2Vec Continuous Bag-of-Words (CBoW) model and fine-tuned PubMed pre-trained word embeddings. Additionally, certain works, including Baumel et al. [22], modeled the patient information using a ConvNet and hierarchical-attention-based GRU, without using any pre-learned embeddings, while others such as Zeng et al. [6] utilized the word embeddings learned during the transfer learning phase.

It is essential to learn the reasoning behind the black-box predictions made by a deep neural model to facilitate evidence-based diagnosis, thus building trust and confidence among medical personnel on the model's capabilities and limitations. Recent studies have focused on analyzing the ConvNet output maps and predictions to decipher the learnings formulated by the neural system. In vision-specific tasks, researchers have utilized coarse localization maps to highlight the essential regions of the image that contribute towards the final output prediction in natural images [73, 74, 75, 76]. These maps and visualization mechanisms have been adapted to medical images as well [77, 78]. On the natural language front (text-based unstructured documents), attempts to generate human-readable explanations through topic coherence and attention-based mechanisms are in progress [73, 79, 80, 81, 82, 83]. Gangavarapu et al. [35] employed coherence models to analyze the topic clusters extracted from clinical nursing notes. Baumel et al. [22] utilized the attention scores obtained from their proposed hierarchical-attention-based GRU model to understand the contributions of summary sentences and their constituent tokens towards each predicted diagnostic code. Mullenbach et al. [27] extracted the most important $n-$gram ($n = 4$) in the discharge summary along with a window of five tokens on either side (for context) to enable interpretability of the neural model in predicting ICD-9 codes. Owing to the ease of analysis and visualization using attention weights, they have been employed in most existing studies for designing interpretable models [84, 85, 86, 87]. Alternate techniques such as Class Activation Maps (CAM) [75, 76], occlusion studies [73], and saliency maps [74] also facilitate effective visualization but remain inadequately explored.

From a modeling standpoint, extensions to the convolutional attention network proposed by Mullenbach et al. [27] were facilitated by using residual convolution blocks [28], multiple convolution layers [29, 30], and bidirectional LSTM networks [31]. However, most of these prior studies employ rudimentary preprocessing techniques and benchmark their results on ICD-9 code prediction using clinical notes transcribed in English. Additionally, most research is heav-

Table 1: Statistics of the Spanish clinical records in CodiEsp corpus, presenting textual evidence and reasoning between the record and its mapped diagnostic codes.

Parameter	**Total**	**Average**
Patient clinical cases	$1,000$	–
Spanish cases cohort	750	–
Unique ICD-10 codes (chosen cohort)	$2,194$	11.13
Unique words in the clinical cases	$34,108$	214.35
Words in the discharge summaries	$2,62,583$	350.11
Words in the longest discharge summary	$1,172$	–
Words in the shortest discharge summary	69	–

ily skewed towards model performance rather than its interpretability, resulting in complex models with increased layers and a fusion of additional external inputs. These limitations hinder the adaptability of the exiting state-of-the-art works in more pragmatic settings, especially in developing nations. This study aims at extending the efforts of the state-of-the-art approaches in utilizing patient-specific information to enhance evidence-based clinical decision support, with minimal risk of clinical deterioration and improved triaging accuracy. We propose a multi-channel, variable-sized convolution attention neural model (*EnCAML*) for the multi-label classification task of ICD-9 code prediction. To minimize classification errors, we determined the optimal threshold on the probability of a discharge summary corresponding to a specific code using the Fisher-Jenks clustering approach. Additionally, we study the impact of initial word embeddings on the overall performance of the proposed neural model, and report on the flexibility and robustness of the *EnCAML* model in the context of the choice of embedding. We present an exhaustive benchmarking of the proposed model for the top-10, top-50 most-frequent codes and code categories, and all $6,918$ codes (corresponding to discharge summaries in the chosen MIMIC-III cohort) against several state-of-the-art models. To establish the language-agnostic nature and adaptability of our proposed model, we validated our performance on the CodiEsp corpus comprising clinical notes in Spanish, annotated with ICD-10 codes. Finally, we demonstrate the explainability and interpretability of the proposed *EnCAML* model in enabling intelligent automated diagnostic coding for enhanced clinical decision-making.

3. Materials and Methods

The proposed multi-channel, variable-sized convolutional attention neural model for diagnostic code prediction, is benchmarked on the patient records available in the MIMIC-III and CodiEsp databases. The CodiEsp corpus is

Table 2: Statistics of the discharge summaries corpus extracted from the MIMIC-III database for the clinical task of diagnostic code (and code category) prediction.

Parameter	**Total**	**Average**
Unstructured discharge summaries	$52,726$	–
Patients in the chosen cohort	$46,520$	–
Unique ICD-9 codes (chosen cohort)	$6,918^{a}$	11.73
Unique words in the discharge summaries	$150,854$	606.465
Words in the discharge summaries	$79,731,657$	$1,513.51$
Words in the longest discharge summary	$10,500$	–
Words in the shortest discharge summary	51	–

[a]A total of $6,984$ diagnostic codes were present in the extracted MIMIC-III discharge summaries corpus. However, post cohort selection and preprocessing 66 of these codes were removed.

relatively small, containing a total of $1,000$ clinical cases (in Spanish) that were manually annotated by medical experts into $2,194$ ICD-10 codes. Furthermore, the Spanish records in CodiEsp dataset (minimal) textual evidence corroborating their mapping to respective diagnostic codes, qualifying the dataset to be best-suited to test the proposed model's interpretability. Owing to the modest size of the CodiEsp corpus, we subject the corresponding clinical text records to minimal preprocessing of character case folding and specific punctuation removal. Table 1 tabulates the statistics of Spanish records in the CodiEsp dataset.

The MIMIC-III database is a comprehensive collection of diverse, clinical and physiological healthcare data of critical care patients admitted to the Beth Israel Deaconess Medical Center, Boston, between June 2001 to October 2012. For our work, the discharge summaries corresponding to $46,520$ intensive unit patients were considered. It is vital to note that the occurrence of ICD-9 diagnostic codes associated with the extracted discharge summaries was highly imbalanced, indicating that the amount of data available to learn more infrequent codes is highly selective. Therefore, it is essential to understand the relevant portions of the clinical free-text that contribute towards the assignment of a particular diagnostic code. In contrast to the CodiEsp corpus, the MIMIC-III database is relatively large and requires exclusive preprocessing to enable accurate data representations. The subsequent sections describe the steps involved in extracting and preprocessing the unstructured text from the discharge summaries to facilitate this clinical task of ICD-9 code (and code category) prediction formulated as a classification problem.

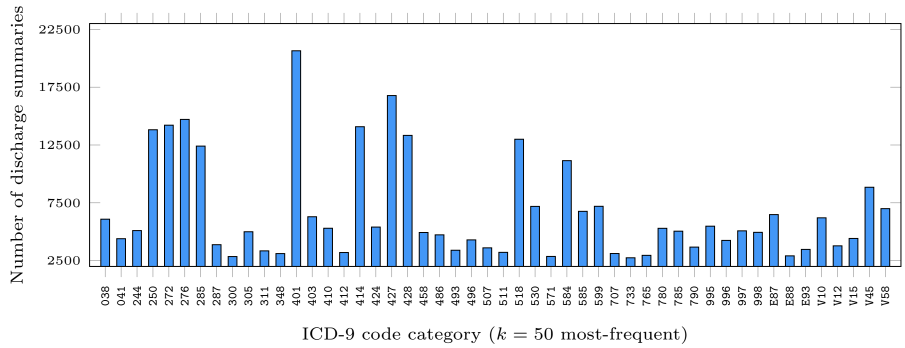
3.1. Patient Records Extraction, Cohort Selection, and Data Cleaning

The MIMIC-III database comprises 26 relational tables, and the required cohort data utilized in this study is extracted from two specific tables. A total of $52,726$ discharge summaries corresponding to various hospital admissions were extracted from the ***noteevents*** table, and the ICD-9 codes corresponding to these summaries were extracted from the ***diagnoses_icd*** table. Specific structural and linguistic details concerning the extracted discharge summary corpus are tabulated in Table 2. We employed a cohort selection criteria in line with that adopted by several state-of-the-art works [27, 23, 35, 32, 36, 54] to enable comparative evaluation of the obtained performance. Accordingly, we only considered discharge summaries that corresponded to the first hospital admission of a patient and discarded data from the subsequent admissions. As argued by Gangavarapu et al. [35, 32, 36], such conditions ensure risk assessment using the earliest detected symptoms.

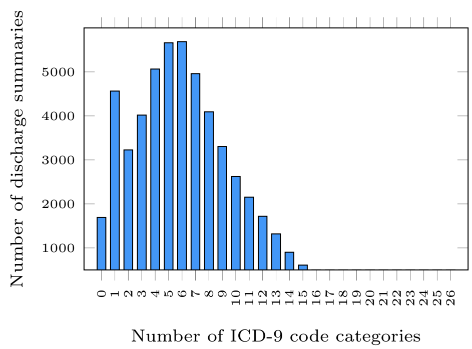
The discharge summaries obtained from the MIMIC-III database included duplicate entries, which were identified and deduplicated. The resulting data corresponded to $6,918$ unique ICD-9 codes in total. Additionally, stemming from the manifold nature of the disease symptoms (e.g., *nephrolithiasis* (formation of kidney stones) caused due to *hyponatremia* (low natrium presence in the blood)), the dataset included multiple records per patient, mapped to different ICD-9 codes. To account for this, we aggregated the content and diagnostic codes across multiple records of a patient, thus enabling multi-label classification. Our work employs binary predictions as the target scores, with a pairwise comparison of actual and predicted values.

3.2. Diagnostic Code Category Assignment

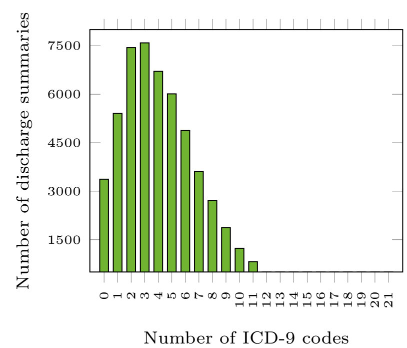
To enable the prediction of diagnostic ICD-9 code categories, we grouped the corresponding diagnostic codes into categories based on the hierarchical nature of the ICD-9 coding taxonomy, resulting in 942 code categories. The multi-label classification of discharge summaries is facilitated through pairwise comparison of the binary predictions with true code categories. The distributions of 50 most-frequent ICD-9 code categories and codes are depicted using Figures 1a and 1d, respectively. The distributions indicating the number of ICD-9 code categories and codes among the discharge summaries are also shown in Figures 1b and 1c, respectively, demanding the need for multi-label prediction. It is interesting to note that with just $k = 10$ and $k = 50$ most-frequent ICD-9 code categories and codes, we can cover majority of the dataset ($k = 10$ codes and categories: 76.93% and 84.24%; $k = 50$ codes and categories: 93.60% and 96.79%). Since we aim to evaluate the proposed modeling strategies on various constructed datasets, they will hereby be referred to as: (a) *top-10-code*, for top-10 ICD-9 codes, (b) *top-10-cat*, for top-10 ICD-9 code categories, (c) *top-50-code*, for



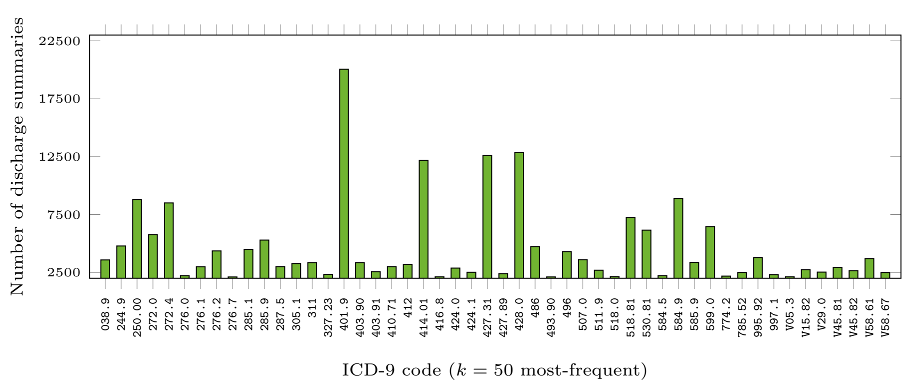
(a) Distribution of ICD-9 code categories across discharge summaries.



(b) Distribution indicating the manifold nature (in terms of categories) of discharge summaries.



(c) Distribution indicating the manifold nature (in terms of diagnostic codes) of discharge summaries.



(d) Distribution of ICD-9 codes across discharge summaries.

Figure 1: Data statistics (distributions) of the patient discharge summaries extracted from the publicly available MIMIC-III database.

Algorithm 1 Procedure employed for automated removal of medical jargon

```
1:  Find tags ending with "<string-1> <string-2>:" using regular expressions
2:  Filter-out all the tags ending with medications
3:  Retain tags containing to-be-excluded keywords (e.g., discharge).
4:  Store the extracted medication tags in a tags-specific database.
5:  Repeat steps 1 through 4 to extract all the patient history tags.
6:  for each text ∈ discharge summary do
7:      if text contains a medication or history tag then
8:          Extract the subsequent tag within the text
9:          Remove content in the text between tags using regular expressions
10:     end if
11: end for
```

top−50 ICD-9 codes, (d) *top-50-cat*, for top−50 ICD-9 code categories, and (e) *all-codes*, for all 6,918 ICD-9 codes. Additionally, we also benchmark our approach using $k = 50$ most-frequent diagnostic (6,918) and procedural (2,003) codes, referred to as *top-50-dp-code*.

3.3. Data Preprocessing

Given the rich information present in unstructured discharge summaries, there is a need to transform the raw clinical text into a canonical form to account for the complex linguistic structure, medical jargon, and voluminosity of the clinical corpus. The discharge summaries obtained from the MIMIC-III database are drawn from a sizeable vocabulary of 150,854 words ($= |\mathbb{V}|$), and each summary consists of a variable length of tokens (see Table 2). In addition to the extensive vocabulary of the selected cohort, multiple discharge summaries maintained per patient adds to the computational complexity and cost of training the underlying neural language models. Hence, it is vital to transform the corpus into a machine-processable format with a manageable vocabulary size. To enhance the manageability of the data, we removed certain medications (e.g., discharge and transfer medications) and patient history sections (e.g., family and social history) from the data. The procedure followed to facilitate such removal is described using Algorithm 1. Next, we eliminated punctuation marks, numeric tokens, and enabled character case folding. Additionally, we tagged all those tokens occurring in less than three summaries as *out-of-vocabulary* words.

To further normalize the content in the summaries, we enabled typographical error correction for those tokens that were not present in the biomedical word embedding vocabulary employed by McDonald et al. [88]. The biomedical word embeddings were trained with approximately 28,000,000 articles compris-

Table 3: A few examples of misspelled tokens from the MIMIC-III discharge summary corpus, corrected using the biomedical word embedding vocabulary from [88].

Observed token	**Corrected token**	**Observed token**	**Corrected token**
ab**cs**ess	ab**sc**ess	abdominal**l**	abdominal
an**ix**ety	an**xi**ety	arrhythm**n**ia	arrhythmia
calcifi**c**ed	calcified	calci**c**um	calcium
calcifed	calcif**i**ed	cardio**gol**ist	cardio**log**ist
cardiolo**l**gy	cardiology	coron**ora**y	coron**ar**y

ing titles and abstracts obtained from the PubMed baseline 2018 collection[10], which accounts for a medical vocabulary of over $2,540,000$ terms. Utilizing the large PubMed vocabulary, we corrected the typographical errors of those tokens (ηs) whose Levenshtein distance ($Lev_{\eta,\rho}$) [89] with the terms in the PubMed vocabulary (ρs) was less than three ($\approx 25,000$ tokens). The Levenshtein distance is computed as:

$$Lev_{\eta,\rho}(n,p) = \begin{cases} \max(n,p), & \text{if } \min(n,p) = 0, \\ \min \begin{cases} Lev_{\eta,\rho}(n-1,p)+1 \\ Lev_{\eta,\rho}(n,p-1)+1 \\ Lev_{\eta,\rho}(n-1,p-1)+\mathbf{1}\{\eta_n \neq \rho_p\} \end{cases} & \text{otherwise} \end{cases} \tag{1}$$

where $Lev_{\eta,\rho}(n,p)$ indicates the distance between the first n characters of η and first p characters of ρ (n and p are $1-$based indices), and $\mathbf{1}\{\bullet\}$ denotes an indicator function. A few examples illustrating the use of Levenshtein distance for correcting the misspelled tokens in the extracted discharge summary corpus are shown in Table 3.

The modeling and representation of the sizeable discharge summary corpus into a $d-$dimensional space ($d \ll |\mathbb{V}|$) was performed by employing a Continuous Bag-of-Words (CBoW) Word2vec model [70], trained on the underlying corpus. Table 4 lists the parameters utilized in generating the word embeddings. We fixed the learning rate to a default value of 0.025 (same as that of the base Word2Vec model presented by Mikolov et al. [70]) and the number of iterations to 10. We experimented with varying embedding sizes of 50, 100, and 200 to empirically determine the optimal embedding size for the underlying clinical task. The implementations of the Word2Vec model available in the Python Gensim library [90] were utilized in generating the embeddings. Ad-

[10] https://www.nlm.nih.gov/databases/download/pubmed_medline.html.

Table 4: Parameters of the Word2Vec models employed to effectively represent the extracted and cleaned discharge summaries.

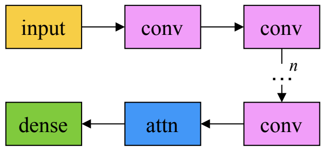
Parameter	Value(s)
Number of iterations	10
Vocabulary size of the summaries without medical jargon removal or typographical error correction[b]	$51,917$
Vocabulary size of the summaries post processing using Algorithm 1	$45,268$
Vocabulary size of the summaries post processing using Algorithm 1, followed by typographical error correction	$42,170$
Employed word embedding sizes	$\{50; 100; 200\}$
CBoW context window size	5
Learning rate of the neural model	0.025

[b]At this stage, all the numeric tokens are removed, infrequent tokens marked as out-of-vocabulary words, and summaries are truncated to a maximum of $2,500$ tokens (as done in [27]).

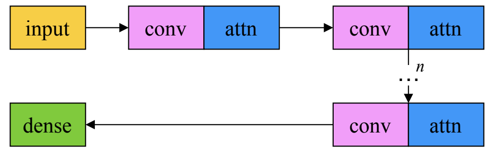
ditional details, including the rationale behind choosing the CBoW Word2Vec model over other recent neural word embedding approaches, including BERT, are presented with experimental validation in Section 5.2.

4. Diagnostic Code Prediction

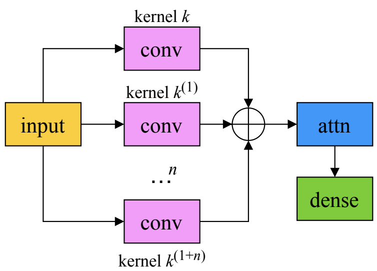
The proposed *EnCAML* convolutional attention network was designed to enhance the predictability of diagnostic codes corresponding to a given discharge summary while enhancing the ease of model interpretability and performance explainability. A linear combination of the features (rather, feature weights) weighted by the convolutional filter convolves the input representation into a more informative feature. Smaller kernel sizes are often the more popular choice over larger sizes, as they capture the desired amount of context without over or undershooting. However, choosing larger kernel sizes could be beneficial when handling highly context-dependent data, as is the case in most healthcare applications. The proposed *EnCAML* neural model utilizes variable-sized multi-channel (parallel) convolution filters to ensure the choice of appropriate kernel size. Furthermore, we employ attention weighting over the convolution filters to highlight the text snippets within the discharge summaries, responsible for mapping the respective summary to a diagnostic code, thus mimicking the actual



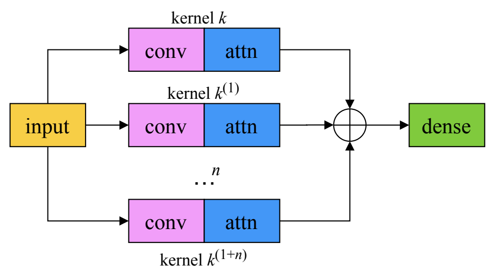
(a) Neural model with $(n+1)$ sequential convolution units.



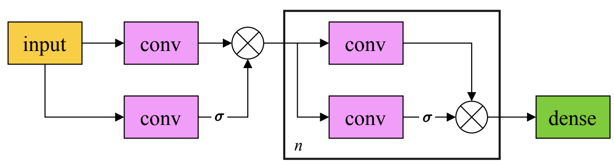
(b) Neural model with $(n+1)$ sequential convolutional attention units.



(c) Neural model with $(n+1)$ parallel convolution units with same or different kernels ($\oplus$ denotes concatenation).



(d) Neural model with $(n+1)$ parallel convolutional attention units with same or different kernels ($\oplus$ denotes concatenation).



(e) Neural model with $(n+1)$ gated convolution units ($\otimes$ denotes element-wise multiplication).

Figure 2: Convolutional attention neural model variants for the task of diagnostic code prediction as multi-label classification. Note that the architecture in (d) with different kernels across parallel convolutional attention units forms the basis for the proposed *EnCAML* architecture.

diagnosis procedure followed at hospitals. Based on our observations, we argue that the proposed model facilitates enhanced predictability and interpretability over alternate variants depicted in Figure 2. The overall architecture of the proposed *EnCAML* neural model is presented in Figure 3.

Let $\mathcal{D}^{(d)} = \{t_1^{(d)}; t_2^{(d)}; \ldots; t_L^{(d)}\}$ be the d-th ($d \in \{1; 2; \ldots; D\}$) discharge summary of length $L = |\mathcal{D}^{(d)}|$ ($\leq 2,500$) comprising tokens $t_i^{(d)}$s, each represented as an $e-$dimensional embedding. The token embeddings adjacent to the token of interest (i.e., the context) are combined using the convolution operation with a filter $F_k \in \mathbb{R}^{f \times e \times k}$, where f is the number of feature maps ($\mathcal{F}_j$s) and $k \in \{3; 5; 7; 9\}$ is the kernel size. Each feature map $\mathcal{F}_j \in \mathbb{R}^L$ and the entire convolution operation over the discharge summary $\mathcal{D}^{(d)}$ results in (four) matrices H_ks of dimension $\mathbb{R}^{f \times L}$ for each kernel size k. Note that we do not perform pooling across the length of the summary to ensure no loss in information, i.e.,

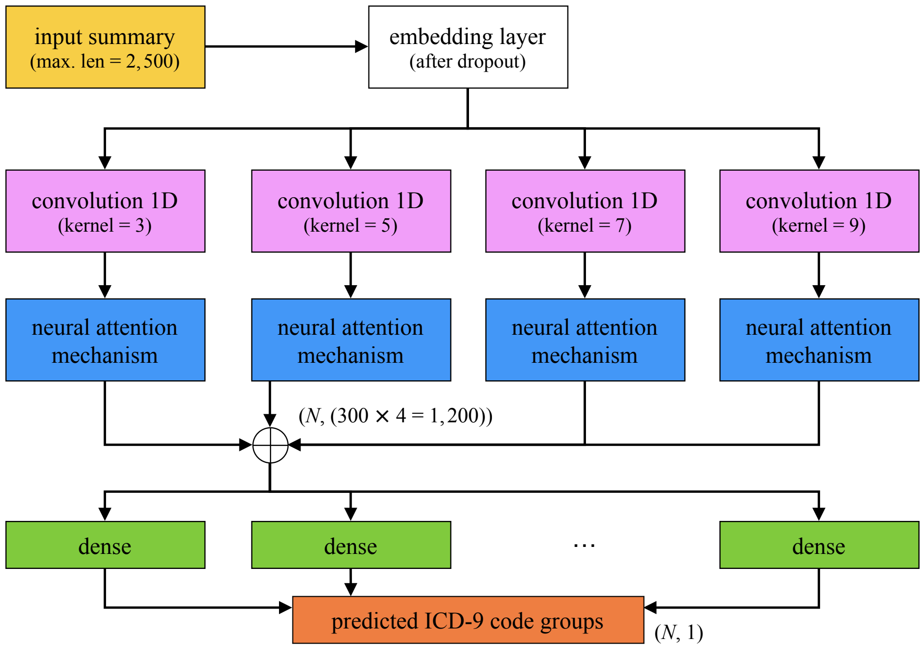


Figure 3: The overall flow employed in the proposed multi-channel, variable-sized convolutional attention neural architecture ($\oplus$ denotes concatenation).

different portions of the summary could be relevant to different diagnostic codes. Next, we mimic the process of diagnosis at hospitals (and manually annotating the patient records) by narrowing down the entire discharge summary to a specific textual portion that most contributes towards the respective diagnostic code. In this regard, we employ the attention mechanism applied per code to highlight the text snippets in the convolution output matrices. The attention weights a_c for a code c are computed using the trainable vector parameter $u_c \in \mathbb{R}^f$ as $a_c = \text{softmax}(H_k^\intercal \cdot u_c)$. The attention weights a_c can help visualize which tokens contribute to code c. The final output representations obtained using the attention vector result in (four) matrices $A_k \in \mathbb{R}^{f \times N}$, one per kernel size, where N is the number of output codes (here, $N \in \{10; 50; 6,918\}$).

To facilitate the classification task of diagnostic code prediction, we built individual classifiers atop the $\oplus\{A_k\}$ $\forall k$ vector representations ($\oplus$ denotes concatenation). We present that modeling diagnostic codes independently instead of employing a single prediction layer is beneficial as the model parameters are fine-tuned independently at the penultimate layer, thus enhancing the predictability of the proposed automated system. This way, the neural model can effectively learn and generalize what features best contribute to a particular diagnostic code. Therefore, a fully-connected layer with a sig-

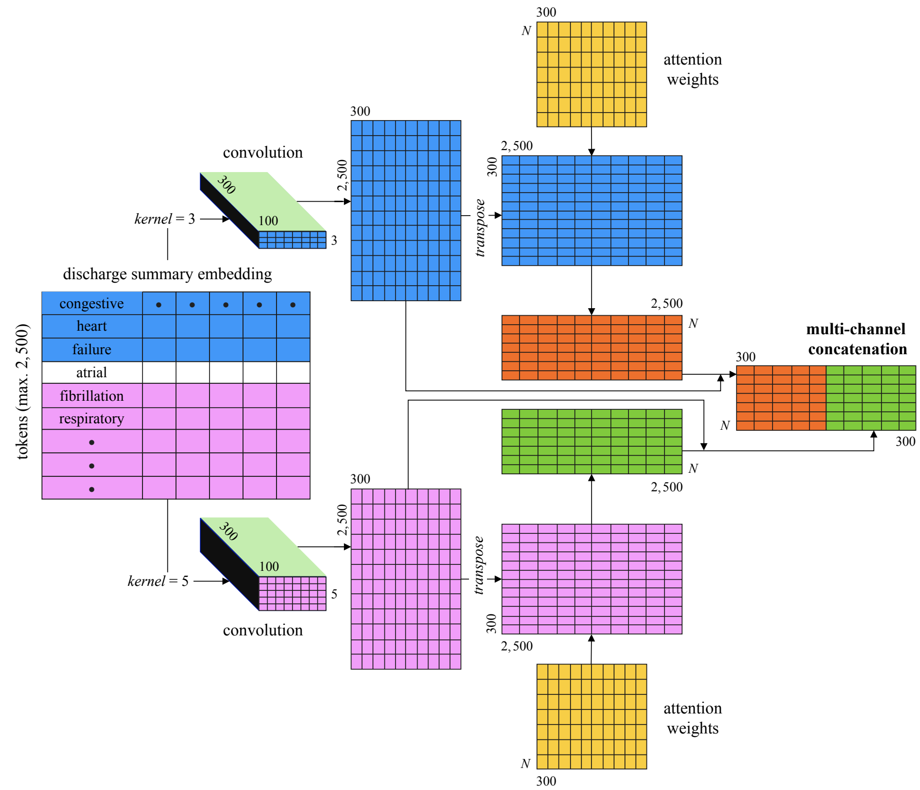


Figure 4: An illustration of the convolutional attention neural architecture (with parallel kernels of 3 and 5) employed by the proposed *EnCAML* model in extracting the discharge summary vector representations for diagnostic code classification.

moid activation function is employed to facilitate binary code prediction, i.e., $\hat{y}_c = \text{sigm}(W^{\intercal}(H \cdot a_c) + b)$, where W and b are the corresponding weight matrix and bias vector, respectively. We trained the neural model to minimize binary cross-entropy loss using Adam optimizer [91]. Additionally, we employed early stopping criterion to mitigate any overfitting of the model. When modeling for the prediction of diagnostic codes among all $6,918$ codes, we employed a single linear layer as opposed to individual code-specific classifiers to lower the computational overhead incurred in training a large number of independent classifiers. An illustration depicting the overall convolutional attention architecture employed in generating discharge summary vector representations for classification is depicted in Figure 4.

The choice of the threshold (θ) on the sigmoid activation layer regulates the predictive performance of the proposed automated diagnostic coding system. Most of the existing studies [23, 27, 36, 35, 34] round-up the obtained output

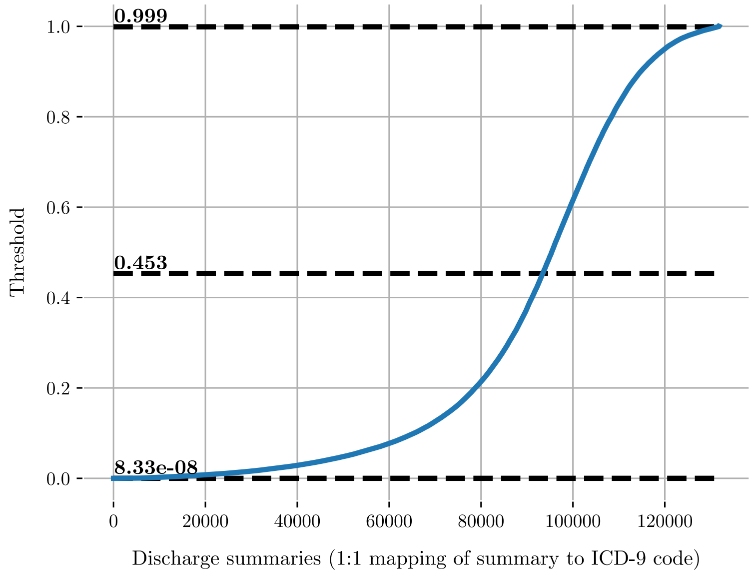


Figure 5: Data-level threshold values obtained using the Fisher-Jenks Natural Breaks algorithm for the *top-10-code* data category.

values to the closer of 0.0 and 1.0 (i.e., an implicit threshold of 0.5), while others, including Li et al. [4], empirically determined the optimal threshold through experimentation with $\theta \in [0.1, 0.95]$. In this study, we employed the Fisher-Jenks Natural Breaks algorithm [92] to find an optimal threshold that maximizes the predictability of $\hat{y}$. The algorithm aims at determining the most-suitable arrangement of values into different classes, i.e., the natural breaks in the data, by minimizing the intra-class variance while maximizing the inter-class variance. These natural breaks can be precomputed from the training data to be employed while testing. In this study, we compute both *code-level* and *data-level* threshold values. For instance, computing the code-level threshold for the diagnostic code `414.01` (*coronary atherosclerosis of native coronary artery*) would involve detecting the natural breaks in $\hat{y}_{\texttt{414.01}}$, i.e., the most optimal threshold that can cluster the training data into + and − classes. Alternately, computing the data-level threshold involves the use of $\hat{y}_c \; \forall c \in y$ to best group the input data according to the distribution of the output classes. We employed the implementations of the algorithm available in the Python Jenkspy library to find the optimal classification threshold values. The generated data-level thresholds for the *top-10-code* prediction task is depicted in Figure 5.

Table 5: The hyperparameter ranges and the experimentally-determined optimal values for the proposed *EnCAML* neural model ($\|$ denotes parallel operation).

Hyperparameter	Experimental value(s)	Optimal value(s)
Embedding sizes (e)	$\{50; 100; 200\}$	100
Kernel sizes (k)	$\{1 \parallel 3 \parallel 5 \parallel 10; 3 \parallel 5 \parallel 7 \parallel 9\}$	$3 \parallel 5 \parallel 7 \parallel 9$
Number feature maps (f)	$\{100; 200; 300; 400\}$	300
Dropout probabilities	$\{0.2; 0.3; 0.5; 0.8\}$	0.2
Learning rates	$\{1\text{e}-4; 3\text{e}-4; 1\text{e}-3; 3\text{e}-3\}$	$1\text{e}-4$
Exponential decay rates	$\beta_1 = 0.9$; $\beta_2 = 0.999$	$\beta_1 = 0.9$; $\beta_2 = 0.999$

5. Experimental Results and Discussion

This section presents the observations from our extensive performance evaluation, both in terms of predictability and interpretability, on CodiEsp and extracted MIMIC-III datasets. The proposed *EnCAML* deep neural model was implemented the functionalities available in the Python PyTorch library [93]. All the experiments, training, and validation were performed using a server running Ubuntu OS with 56 cores of Intel Xeon processors, 128 GiB RAM, 3 TB hard drive, and two NVIDIA Tesla M40 GPUs, running CUDA v10.1. The proposed *EnCAML* model is trained using the curated discharge summaries from the MIMIC-III database (capped at $2,500$ tokens, see Table 4 for more details) and their corresponding ICD-9 code mappings. We tuned the model hyperparameters using relevant experimental values obtained from prior studies and retrieved the optimal values for those parameters through experimental validation. The results of our hyperparameter tuning are summarized in Table 5. Using the *EnCAML* model trained with the chosen optimal hyperparameters, we establish the predictive and interpretive superiority of our proposed approach over several state-of-the-art benchmarks.

In an attempt to enable accurate benchmarking of the obtained performance, we grouped the datasets into train, validation, and test sets exactly as reported by the respective state-of-the-art studies. For datasets with diagnostic codes and rolled-up categories (*top-10-code*, *top-50-code*, *top-10-cat*, and *top-50-cat*), we employed the 50-25-25 split facilitated by the hospital admission identifiers in the train-validation-test-HADM_IDs set utilized by Huang et al. [23]. While modeling the *top-50-dp-code* dataset, we employed the hospital admission identifiers from the train_50-HADM_IDs set reported in [27]. For the code prediction task employing all $6,918$ codes (*all-codes* dataset), we used a 90-to-10 train-to-test split, enabling maximum training instances to ensure model generalizability on a large number of target classes. As stated earlier, we incorporated the early stopping criterion (tolerance of five epochs) while training to overcome possible overfitting of the deep neural model, thus enabling the most optimally-trained neural model to enhance the code predictability.

The CodiEsp dataset presents an inherent division of its $1,000$ clinical records into training (500 instances), validation (250 instances), and test (250 instances) sets. However, since the clinical texts in the test set remain unannotated, we present our results with the validation texts as the test set. In modeling the CodiEsp clinical corpus annotated with ICD-10 codes, we present our performance on top-10 and top-50 most-frequent codes (referred to as *top-10-ce-code* and *top-50-ce-code*; "ce" indicates the CodiEsp corpus) owing to the limited number of available training instances.

5.1. Evaluation Metrics

To informatively report the performance of our proposed model, we employ the extensively utilized micro-averaged and macro-averaged F_1 scores [94]. The F_1 (more generally, $F_{\beta=1}$) aims to seek a balance between precision and recall and is interpreted as a weighted harmonic mean of the two [95]. Therefore, models with relatively higher F_1 scores are expected to enhance the predictability of the system. Since the F_1 measure accounts for the true and false positives (TP and FP) as well as true and false negatives (TN and FN), it is often regarded to be more indicative than the standard accuracy score. The F_1 score is computed as follows:

$$F_{\beta=1} = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{2}$$

where the values of precision and recall are micro-averaged or macro-averaged over the target output classes (here, codes and code categories). For a neural system trained to predict over N ICD-9 codes or code categories, the micro-averaged precision and recall are computed as:

$$\text{precision}_{\text{micro}} = \frac{\sum_{c=1}^{N} \text{TP}_c}{\sum_{c=1}^{N} (\text{TP}_c + \text{FP}_c)};\ \text{recall}_{\text{micro}} = \frac{\sum_{c=1}^{N} \text{TP}_c}{\sum_{c=1}^{N} (\text{TP}_c + \text{FN}_c)} \tag{3}$$

On the other hand, the macro-averaged precision and recall scores computed as the average observed precision and recall over N ICD-9 codes or code categories are obtained using:

$$\text{precision}_{\text{macro}} = \frac{1}{N} \sum_{c=1}^{N} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c};\ \text{recall}_{\text{macro}} = \frac{1}{N} \sum_{c=1}^{N} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{4}$$

In addition to the F_1 scores, we also report our performance using the Jaccard similarity score [96] and Hamming loss multi-label classification metrics. The Jaccard loss computed using (5) captures the amount of dissimilarity between the actual (y) and predicted ($\hat{y}$) code or code category sets, averaged over the entire validation or test set. Alternately, the Hamming loss as computed in (6) estimates the ratio of misassigned codes or code categories from the given sequences of actual and predicted label sets, and is also averaged over the entire validation or test set. The models with lower Jaccard and Hamming loss values are regarded to be high performing.

$$\text{Jaccard}(y, \hat{y}) = \frac{1}{D} \sum_{d=1}^{D} \frac{|y^{(d)} \cap \hat{y}^{(d)}|}{|y^{(d)} \cup \hat{y}^{(d)}|} \tag{5}$$

$$\mathcal{L}_{\text{Hamming}}(y, \hat{y}) = \frac{1}{D}\frac{1}{N} \sum_{d=1}^{D}\sum_{c=1}^{N} \mathbf{1}\{y_c^{(d)} \neq \hat{y}_c^{(d)}\} \tag{6}$$

Finally, we also evaluate the performance of the proposed neural model using micro-averaged and macro-averaged Area Under the Receiver Operating Characteristic curve (AUROC). Since the ROC curve is a probability curve (plotted as sensitivity against the fall-out), the area under the curve represents the measure of class separability, i.e., a quantitative measure of the capability of the model in distinguishing between target codes or code categories [32]. By analogy, the higher the value of AUROC value, the better the model at distinguishing between patients with and without corresponding diseases.

5.2. Word Embeddings and Predictability of EnCAML

The employed word embedding neural network determines the representation of the underlying clinical text, thereby effectively capturing the document's semantics. By extension, it seems intuitive to establish that vector representations capturing a higher level of document semantics (e.g., intra-word associations mined using self-attention) would outperform more simplistic approaches. However, it must be noted that a flexible and robust classification model must be capable of generalizing over minimalistic representations, as well as learning adequately from highly semantics-specific representations without overrepresenting the underlying patterns. To analyze the impact of the choice of initial word embedding on the proposed *EnCAML* neural classification model, we experimented with several state-of-the-art word embedding approaches, including Word2Vec (skip-gram and CBoW variants), fastText (skip-gram and CBoW variants), and BERT (pre-trained on clinical corpora and fine-tuned). By reporting any variations in the classification performance of the proposed *EnCAML* model, we aim to establish the robustness of the proposed approach in modeling unstructured clinical text.

The Word2Vec (or a close variant) neural model for generating word embeddings [70] has been widely employed in modeling clinical text across several state-of-the-art studies [4, 33, 32, 35, 36, 34], owing to its ability to capture the text semantics in a simple yet efficient manner. However, models reliant on

Table 6: Results depicting the effect of initial word embedding choice on the overall predictive performance of the proposed *EnCAML* model, recorded using discharge summaries of the *top-10-cat* data category.

Embedding model	F_1 **micro**
Skip-gram Word2Vec	0.7784
CBoW Word2Vec	**0.7811**
Skip-gram fastText	0.7811
CBoW fastText	**0.7821**
Fine-tuned BERT (clinical texts + PubMed abstracts)	0.7760
Pre-trained BERT (Alsentzer et al. [97])	0.7729
Xavier uniform initialization	0.7668

Word2Vec approaches often cluster all the out-of-vocabulary words into a single vector representation, defaulted for all unknown tokens. In this regard, the more flexible fastText neural model [71] aims at representing the unknown tokens as some combination of known sub-tokens, thus overcoming the limitations of the Word2Vec model. Finally, a more advanced self-attention-based BERT model [72] captures the context of the given token from both left-to-right and right-to-left, aiming to extract the exact intended semantics of the underlying text, which would otherwise go unnoticed. For comparison, we obtained the BERT embeddings pre-trained on the entire MIMIC-III discharge summaries corpus from [97]. Additionally, we also generated fine-tuned BERT embeddings by re-training the embedding model with our clinical vocabulary and discharge summaries corpus (see Table 2). For brevity, we utilized the pre-trained checkpoints obtained while training on clinical texts, released by Alsentzer et al. [97], and those obtained while training on PubMed abstracts, released by Peng et al. [98], to initialize the BERT model. Word2Vec and fastText models were employed through the implementations available in the Python Gensim library [90], while BERT embeddings were generated using the openly available BERT-as-service framework[11]. Since the BERT (base model) outputs a vector embedding of 768 dimensions, the same embedding size was employed while modeling Word2Vec and fastText models for comparison. Furthermore, the Word2Vec and fastText embeddings were deployed with a window size of five, trained for 30 iterations over the corpus.

We report the obtained performance (measured as micro F_1 score) of our proposed *EnCAML* classification model for various neural embeddings in Table 6.

[11] https://github.com/hanxiao/bert-as-service.

Additionally, we also present the micro F_1 score obtained using a random Xavier uniform initialization of $768-$dimensional vector per token as the baseline. It can be observed that the CBoW variants of Word2Vec and fastText models always outperform their skip-gram counterparts. One possible interpretation for this behavior could be that predicting a target word, given the neighboring noisy context, is far simpler than predicting the exact noisy context for a given target token. Despite the fastText CBoW variant achieving the highest performance, the speedup obtained for Word2Vec models was nearly ten-fold ($10\times$) at a similar (i.e., insignificantly lower) performance. We attribute this speedup in that the fastText model aimed at representing several out-of-vocabulary medical jargon tokens that were rather uncontributing to the final prediction output. Considering these findings, we chose to model the input discharge summaries as vector representations output by the CBoW Word2Vec embedding network.

As can be observed from Table 6, the presented baseline, i.e., Xavier uniform initialization at random, also provides comparable performance with respect to other more sophisticated models. This corroborates that the values of initial embedding vector components play little to no role in enhancing the predictability of the *EnCAML* model. Since the proposed *EnCAML* model employs multiple attention layers, thus enabling the learning of per-code attention weights over training samples, the initialization of input vectors with pre-trained embedding weights is quite redundant and cost-intensive (requiring additional storage space of up to 1.5 GiB). The robustness of the proposed *EnCAML* model over other state-of-the-art models lies in its ability to learn from and generalize over the input discharge summaries in a rather end-to-end fashion. Hence, it is arguable that such a system could enable rapid prototyping and deployability in real-world scenarios, especially in modeling noisy clinical data obtained from the hospitals of developing nations, which are far less ideal than the standard datasets utilized in academic research.

5.3. Performance Benchmarking

We enable effective performance benchmarking of our proposed *EnCAML* model against several state-of-the-art studies. As stated earlier, we curated six data categories from the obtained MIMIC-III corpus to facilitate exhaustive comparison. For the *top-k-code* ($k = 10, 50$) data categories, the discharge summaries mapped to the top$-k$ ICD-9 diagnostic codes were employed in benchmarking. On the other hand, for the *top-k-cat* ($k = 10, 50$) data categories, we rolled-up the ICD-9 diagnostic codes up to three digits (e.g., `225.2` (*benign neoplasm of cerebral meninges*) and other codes within the `225.x` class were rolled-up into the `225` category (*benign neoplasm of brain and other parts of nervous system*)) and extracted the discharge summaries corresponding to top$-k$ categories. Since most of the existing works presented their performance on the combined set of most-frequent diagnostic and procedural ICD-9 codes,

Table 7: Results from our performance benchmarking of the proposed *EnCAML* neural model against several prior state-of-the-art works. The highest achieved performance in a given code category among various models (including *EnCAML*) is indicated in **bold**, while the best-performing model (macro or micro F_1 score) from prior studies is marked using (*).

Data category	Study (model)	F_1 score	
		macro	micro
top-10-code	This work (multi-channel CAML)	**0.7624**	**0.7772**
	Huang et al. [23] (GRU)	0.6957*	$-^c$
	Samonte et al. [25] (hierarchical attention + topic modeling)	0.6870	$-^c$
	Rios and Kavuluru [26] (transfer learning)	0.6200	$-^c$
top-10-cat	This work (multi-channel CAML)	**0.7782**	**0.7840**
	Huang et al. [23] (GRU)	0.7233*	$-^c$
top-50-code	This work (multi-channel CAML)	**0.6028**	**0.6733**
	Huang et al. [23] (GRU)	0.3263	$-^c$
	Guo et al. [37] (bidirectional LSTM)	$-^c$	0.5720*
top-50-cat	This work (multi-channel CAML)	**0.6363**	**0.6908**
	Huang et al. [23] (LSTM)	0.3367*	$-^c$
top-50-dp-code	This work (multi-channel CAML)	**0.6109**	**0.6764**
	Mullenbach et al. [27] (description-regularized CAML)	0.5760	0.6330
	Mullenbach et al. [27] (single-channel CAML)	0.532	0.6140
	Li and Yu [28] (multi-filter residual ConvNets)	0.6060	0.6700*
all-codes	This work (multi-channel CAML)	**0.0859**	**0.5258**
	Zeng et al. [6] (transfer learning)	$-^c$	0.4200*
	Li et al. [4] (Doc2Vec + ConvNet + $\theta = 0.2$)	$-^c$	0.4080
	Baumel et al. [22] (ConvNet)	$-^c$	0.4070

[c]The score for the corresponding metric was not reported by the underlying study.

we also benchmark our performance on the combined set, represented as *top-50-dp-code* data category. Finally, we evaluate our proposed model trained on all the 6,918 ICD-9 diagnostic codes observed in the obtained MIMIC-III cohort, under the *all-codes* data category.

The obtained results from our performance benchmarking of the proposed *EnCAML* neural model against the state-of-the-art models are presented in Table 7. As presented earlier, we employed the test-validation-train sets similar to that reported by the prior studies, thus mitigating the necessity to reimplement their proposed models for comparison. Additionally, in cases where the under-

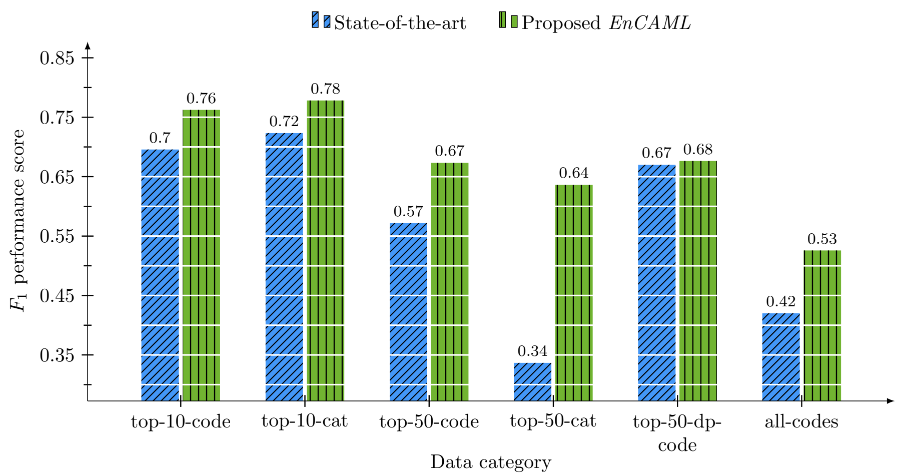


Figure 6: Performance comparison (measured as macro or micro F_1 score) of the proposed *EnCAML* approach against best-performing state-of-the-art models for the corresponding curated data category (entries marked using (*) in Table 7).

lying studies did not report certain metrics, we attempted to reimplement their works; however, this was quite challenging due to the lack of exact modeling specifics, including precise data splits, external data curation and annotation strategies, and others. We benchmark our results to the best possible extent, depending on the replicability of the prior works for metrics not captured by them. From Table 7 and consequently Figure 6, it can be observed that the proposed *EnCAML* model outperforms the state-of-the-art works by a significant margin on all the data categories, owing to the enhanced predictability attributed by the multi-channel, variable-sized convolutional attention layers. We seek to draw attention towards the single-channel convolutional attention network [27] whose architecture is quite close to the one presented in this study. Observe the improved performance by shifting from a single-channel convolution to a multi-channel variable-sized convolution. Such significant improvement results from the proposed model being able to expand its reach to three, five, seven, and nine (employed kernel sizes) context spaces, thereby mitigating the need to establish the most optimal context size for capturing the essence of the underlying discharge summary manually. Furthermore, the incorporation of per-code classification and Fisher-Jenks thresholds also has favorable effects on the overall predictability of the model, as quantified by the significant performance improvement of up to 89% (for *top-50-cat* data category) achieved by the *EnCAML* model over prior works.

We tabulate the obtained performance of the proposed *EnCAML* model per data category using additional multi-label evaluation metrics (see Section 5.1 for details) in Table 8. As expected, it can be observed that the Jaccard similarity

Table 8: Extensive performance benchmarking of the proposed *EnCAML* deep neural model per data category, using additional standard multi-label classification metrics, including Jaccard similarity score, Hamming loss, and AUROC.

Data category	**Jaccard score**	**Hamming loss**	**AUROC**	
			macro	**micro**
top-10-code	0.6887	0.0912	0.9377	0.9447
top-10-cat	0.6770	0.1187	0.9230	0.9331
top-50-code	0.5231	0.0573	0.9223	0.9439
top-50-cat	0.5460	0.0773	0.9136	0.9345
top-50-dp-code	0.5178	0.0751	0.9056	0.9309
all-codes	0.3701	0.0015	0.9861	0.8985

score decreases with an increase in the number of target codes or code categories (compare *top-10-x* with *top-50-x* in Table 8). This indicates the difficulties in obtaining exactly-matched predicted and actual output sets for large sets of target labels. Any inferences drawn from the Jaccard score beyond this are meaningless since the score is computed as a fraction of the union of predicted and actual codes. Since Hamming loss accounts for a normalized score of the number of mismatches between the predicted and actual code sets, it serves to be more informative than the Jaccard score. It can be noted that the Hamming loss decreases with an increase in the number of target codes or code categories, which could be the result of the normalization factor $(1/N)$ at play. However, in general, it can be seen that the Hamming loss is relatively low across all data categories, indicating a smaller number of misclassifications (both FN and FP). Finally, the AUROC scores vibrate in the range of 0.90 to 0.99 (close to 1.0), indicating the efficacy of the proposed model in differentiating the discharge summaries associated with a particular code from those not associated with that code.

As reported earlier, owing to the limited size of the CodiEsp Spanish clinical notes corpus, we benchmark the proposed *EnCAML* model on top-10 and top-50 most-frequent ICD-10 diagnostic codes (*top-10-ce-code* and *top-50-ce-code*), respectively. Since the test set code labels were not made publicly available at the time of this study, we present our performance on the CodiEsp corpus as a way to demonstrate the flexibility and adaptability of our proposed model. The obtained performance, measured as (micro and macro) F_1 and AUROC scores, is presented in Table 9. Observe the recorded high performance despite minimal preprocessing employed while handling CodiEsp clinical notes. A decreased yet competitive performance while modeling with *top-50-ce-code* data category could be explained by the availability of a limited number of training

Table 9: Results from our benchmarking experiments on the CodiEsp Spanish clinical notes corpus for the clinical task of ICD-10 code prediction, using the proposed *EnCAML* deep neural model.

Data category	F_1 **score**		**AUROC**	
	macro	**micro**	**macro**	**micro**
top-10-ce-code	0.7684	0.8188	0.9521	0.9631
top-50-ce-code	0.6195	0.7008	0.9079	0.9321

instances (500) mapping to a relatively larger number of diagnostic codes. All in all, the proposed *EnCAML* model is shown to generalize over non-English texts with a more convoluted ICD-10 coding system, thus establishing the superiority of the proposed approach over prior works.

5.4. Performance Analysis and Discussion

In the previous subsection, we presented our extensive benchmarking experiments and demonstrated the superiority of the proposed *EnCAML* model compared to several state-of-the-art works. This subsection presents our major findings from comparing the *EnCAML* model with the prior studies beyond the obtained performance and attempts to draw contrasts between them. Most of the existing studies presented little to no stress on the utilized preprocessing steps. Even the works that did perform substantial preprocessing, their approaches were quite rudimentary, mostly limited to tokenization, non-alphanumeric and stopword removal, stemming, and case folding. In this aspect, our preprocessing pipeline was far more extensive, involving typographical error correction (using an external voluminous biomedical corpus), automated removal of medical jargon and irrelevant content pruning (through handcrafted keyword searches), and capping the number of tokens per discharge summary (see Section 3.3 for details). As a result, the medical vocabulary size reduced from over 1.5 million entries to a mere $42,170$ ($3.6\times$ smaller). Moreover, since each token in the vocabulary is translated into an $e-$dimensional vector ($e = 100$), such a significant reduction in the vocabulary size resulted in a substantial optimization. Furthermore, despite the proposed *EnCAML* model employing four parallel convolutional attention layers, our model still has $560,000$ less trainable parameters than the single-channel convolution attention network presented by Mullenbach et al. [27], built on a vocabulary of $51,917$ tokens. When applied to the single-channel model, our preprocessing pipeline facilitated a significant speedup in the training process while improving the predictability of the neural model (see Table 15 for results from our ablation study).

From the neural model training perspective, our *EnCAML* model starts to converge (performance saturation) between 34 to 36 epochs, while the single-channel convolutional attention model proposed by Mullenbach et al. [27] takes twice as long (i.e., 63 to 65 epochs). We attribute this fast convergence to the use of multi-channel convolutions and per-label classifiers as opposed to a single linear layer. Furthermore, the task of multi-label classification of N diagnostic codes or code categories, facilitated through N binary classifiers, enables the neural model to generalize over relevant features that correspond to the underlying code more effectively. On the aspect of extracting features from the given data, the multi-channel variable-sized convolution filters extract crucial information from the underlying discharge summary at varying contexts, which are then searched attentively (through neural attention) for vital portions that are responsible for the corresponding output diagnostic code. The use of multi-channel convolution instead of a fixed-length filter enhances the model's flexibility in choosing the context of representation and relies entirely on the attention layer to segregate between the convolved outputs. Employing a pooled convolution output (as opposed to an attention-based aggregation) often results in a loss of information (relevant features corresponding to specific code labels), especially when classifying data with a large number of sparse and diverse target labels (e.g., *all-codes* data category), as observed with the use of traditional ConvNet models in [4] and [22]. Additionally, the *EnCAML* model facilitates an unrestricted use of variable-sized filters resulting in variable-sized contexts that are weighed by attention, enhances the interpretability of the obtained neural predictions to a large extent.

It is reasonable to argue that modeling text-based discharge summaries could be facilitated by recurrent neural models such as LSTM or GRU that effectively capture the dependencies within the text. However, since most of the discharge summaries range between 500 to $2,500$ tokens in length (after truncating), sequence models could experience severe vanishing gradient problems. However, our proposed model with multiple convolutional layers is able to adequately cope with such issues, as is evident from the reported high performance of *EnCAML* compared to GRU [23], LSTM [23], and bidirectional LSTM [37] models. Additionally, employing more sophisticated neural models such as BERT to handle the limitations with recurrent networks is also challenging, especially due to the high computational cost of training, exacerbated by its fixed input sequence length of 512 tokens (lower end of the discharge summaries length range), warranting additional runs to accommodate longer texts.

We analyzed the discharge summaries of the MIMIC-III database corresponding to the misclassifications from our proposed model in an attempt to explain the predictions output by the model. For the more severe false-negative scenarios (existing disease goes unidentified), it was observed that several discharge summaries under this category included minimal disease-specific reference text and several links to alternate sources of patient-specific information

Table 10: Sample discharge summaries from the MIMIC-III corpus with vague and unusable information with respect to the mapped ICD-9 diagnostic codes, illustrating the intrinsic complexities in modeling unstructured clinical data.

ICD-9 code(s)	**Discharge summary**
`584.9`: Acute renal failure, unspecified	... see outside medical records for history of present illness, physical examination, pertinent laboratories, x-ray electrocardiogram, and other tests ...
`428.0`: Congestive heart failure, unspecified	... her discharge was delayed one day due to bed unavailability at rehab ...
`427.31`: Atrial fibrillation	... please see discharge summary record from outside medical record notes ...
`998.32`: Disruption of external operation (surgical) wound	... this addendum will serve to confirm that in addition to the previous discharge summary the admission diagnosis should be included ...
`401.9`: Unspecified essential hypertension	... this is an addendum to the initial discharge summary which was dictated when the patient remained in the hospital awaiting appropriate rehabilitation facility ...
`V45.81`: Aortocoronary bypass status	... please refer to the discharge summary dictated by myself with discharge date for content ...

such as nursing notes or outside medical records. With little to no diagnostic-code-specific text in the underlying summary, our *EnCAML* model was unable to provide conclusive predictions. Several such sample discharge summaries and their associated ICD-9 diagnostic codes are documented in Table 10. In the more tolerant false-positive cases (nonexistent disease gets marked-up), the discharge summaries included prolonged patient histories that signaled the *EnCAML* model to mark-up the content within the history as evidence to predict the corresponding nonexistent ICD-9 diagnostic code as existent. Specific examples of discharge summaries falling into the false-positive category are highlighted in Table 12.

5.5. Evaluation of Interpretability

We now present details on the interpretability of the diagnostic code predictions facilitated by the proposed *EnCAML* model, specifically through the attention layers of the neural model trained at the individual diagnostic code level. Table 11 presents sample patient discharge summaries extracted from the MIMIC-III database whose content is highlighted using the learned attention weights (a_cs) corresponding to the respective diagnostic code c. These

Table 11: Examples of patient discharge summaries extracted from the MIMIC-III database establishing the interpretability and explainability of the proposed *EnCAML* model. The text snippets indicating the possibility of the respective ICD-9 diagnostic code in the discharge summary are highlighted in blue.

Parameter	**Value**
Extracted $n-$grams using attention weights	... mass he *received units of packed red blood* cells ... discharge diagnosis *upper gastrointestinal bleed* discharge ...
Extracted $n-$grams using Grad-CAM	... presented with **hematocrit drop** and had guaiac ... mass he **received units of packed red blood** cells ...
Top-3 tokens	bleed, drop, and hematocrit
Associated ICD-9 code	`285.1`: Acute posthemorrhagic anemia
Extracted $n-$grams using attention weights	... a history of *hypothyroidism morbid obesity* polycystic ovarian ... in the *evening levothyroxine* mcg oral ...
Extracted $n-$grams using Grad-CAM	... on **exertion paroxysmal nocturnal dyspnea orthopnea** ankle ... in the **evening levothyroxine** mcg oral ...
Top-3 tokens	levothyroxine, hypothyroidism, and levoxyl
Associated ICD-9 code	`244.9`: Unspecified hypothyroidism

highlighted tokens were considered most contributing towards the corresponding ICD-9 code by the *EnCAML* model, and Table 11 also presents the top-3 tokens that were highly weighted by the neural system. The visualization of the text snippets demonstrates the effectiveness of the proposed model in learning the most relevant and vital keywords adequately to facilitate enhanced predictability of the corresponding ICD-9 codes. As reported earlier, the attention mechanism extracts patterns that signal the presence of a corresponding code based on the entire discharge summary (without any pooling over the convolved outputs). Therefore, in cases of summaries containing extended patient histories with minimal disease-specific indicators, the attention mechanism seems to classify the patient history as if it were the current illness. Examples of such discharge summaries extracted from the MIMIC-III database, resulting in false-positive predictions, are tabulated in Table 12.

To benchmark the interpretability and explainability of the proposed *EnCAML* approach, we compare the resultant attention output for a discharge summary to that obtained using the Gradient-weighted CAM (Grad-CAM) [76] approach. Grad-CAM employs the gradients of a target class, flowing into the

Table 12: The predictability and interpretability of the proposed *EnCAML* model on sample patient discharge summaries extracted from the MIMIC-III database. Observe that the predicted false-positive ICD-9 codes (indicated in ~~strikethrough~~ text) are evidently signaled from the text snippets (marked in red; in the first summary (top), `401.9` corresponds to the term *hypertension*; in the second summary (bottom), `414.01` corresponds to the terms *coronary artery disease* and *cardiac catheterization*.

Parameter	**Value**
Extracted $n-$grams using attention weights	... complaint *giant paraesophageal hernia* major ... past medical history pulmonary *hypertension* *depression lyme disease osteopenia* ...
Extracted $n-$grams using Grad-CAM	... diagnosis giant paraesophageal **hernia gerd** **hypertension** **osteopenia depression** ...
Predicted ICD-9 code(s)	`311`: Depressive disorder, not elsewhere classified `530.81`: Esophageal reflux `401.9`: ~~Unspecified essential hypertension~~
Actual ICD-9 code(s)	`518.81`: Acute respiratory failure
Extracted $n-$grams using attention weights	... and family history of *coronary artery disease* ... who presents for *cardiac catheterization* to evaluate ...
Extracted $n-$grams using Grad-CAM	... past medical history prostate brachytherapy years ago ... and underwent **aortic valve** replacement ...
Predicted ICD-9 code(s)	`39.61`: Extracorporeal circulation auxiliary to open heart surgery `401.9`: Unspecified essential hypertension `414.01`: ~~Coronary atherosclerosis of native coronary artery~~
Actual ICD-9 code(s)	`39.61`: Extracorporeal circulation auxiliary to open heart surgery `401.9`: Unspecified essential hypertension

final convolution layer (before the attention layers in *EnCAML*), to produce a localization map highlighting the important candidate $n-$grams in the underlying summary for predicting the corresponding code. Since Grad-CAM allows for the visualization of all possible contributing $n-$grams, it spans a much broader aspect than the attention outputs of the *EnCAML* model. However, on the flip side, because attention outputs are quite narrowed down, they are more precise and depict accurate understandings of what the underlying deep neural model looks at. Upon experimentation, we observed that the Grad-CAM and attention outputs are quite similar for most of the discharge summaries (see Tables 11 and

Table 13: Several examples of attention visualization, comparing [27] to the proposed *EnCAML* for discharge summaries in the MIMIC-III database. Text color corresponds to softmax($H_k^\top \cdot u_c$) (obtained attention weight for the actual code c), where blue (■) indicates low code-correspondence and red (■) indicates high code-correspondence. The false-positive predictions are marked using ~~strikethrough~~ text.

Actual code(s)	**Mullenbach et al. [27]**		**This work**	
	Predicted code(s)	**Attention visualization**	**Predicted code(s)**	**Attention visualization**
39.61: Extracorporeal circulation auxiliary to open heart surgery	39.61: Extracorporeal circulation auxiliary to open heart surgery 995.92: ~~Severe sepsis~~	he recently underwent a urologic procedure developed urosepsis with mrsa bactermia and an echocardiogram was performed to check systolic murmur	39.61: Extracorporeal circulation auxiliary to open heart surgery	name if chief complaint aortic stenosis do major surgical invasive procedure aortic valve replacement jude epic porcine history of present illness year
530.81: Esophageal reflux	530.81: Esophageal reflux 401.9: ~~Unspecified essential hypertension~~	found to have edh was sent to hospital medical history pmhx gerd hospital course 25m admitted for close clinical observation of mental status epidural	530.81: Esophageal reflux	hospital past medical history pmhx gerd hospital course admitted for observation ⋯ hospital discharge diagnosis epidural hematoma r temporal bone fx discharge condition
285.1: Acute posthemorrhagic anemia	285.1: Acute posthemorrhagic anemia 401.9: ~~Unspecified essential hypertension~~ 38.93: ~~Venous catheterization, not elsewhere classified~~	prevent this side effect medication refills cannot be written after noon on fridays anticoagulation take lovenox for dvt prophylaxis for weeks post therapy	285.1: Acute posthemorrhagic anemia	left calcaneus fracture right above elbow amputation post operative blood loss anemia discharge condition mental status coherent level of consciousness alert
305.1: Tobacco use disorder 311: Depressive disorder, not elsewhere classified	– (no predictions)	last name namepattern md telephone fax building sc hospital ward name clinical ctr location un campus east best parking hospital ward name and number completed	305.1: Tobacco use disorder 311: Depressive disorder, not elsewhere classified 276.2: ~~Acidosis~~	tachycardic with blood sugar in 600s and found to have anion gap metabolic acidosis and ketonuria consistent with dka was treated with insulin drip and ivf and

Table 14: Case study on clinical notes from the CodiEsp corpus demonstrating the predictability and interpretability of the proposed *EnCAML* model. For the second note (bottom), our *EnCAML* model also predicted `r52` (false-positive, indicated in ~~strikethrough~~ text), signaled from the use of the term *hinchazón mandibular* (translates to *jaw swelling*).

Parameter	Value
Interpretation using attention weights	... único antecedente **de hipertensión arterial** presentaba ... cefaleas y *vómitos* no asociados ...
Actual expert-provided text evidence	**hipertensión arterial** *vómitos*
Predicted ICD-10 code(s)	`i10`: **Essential (primary) hypertension** `r11.10`: *Vomiting, unspecified*
Actual ICD-10 code(s)	`i10`: **Essential (primary) hypertension** `r11.10`: *Vomiting, unspecified*
Interpretation using attention weights	... presentar dolor e *hinchazón mandibular* ... progresión de **la enfermedad y** deterioro ...
Actual expert-provided text evidence	**enfermedad**
Predicted ICD-10 code(s)	`r69`: **Illness, unspecified** `r52`: ~~*Pain,unspecified*~~
Actual ICD-10 code(s)	`r69`: **Illness, unspecified**

12 to compare attention and Grad-CAM outputs). Additionally, we also compared the model interpretability between *EnCAML* and the single-channel convolutional attention network proposed by Mullenbach et al. [27], employing a kernel size $k = 10$. More recent studies [29, 31] facilitated enhanced learning from external data sources such as Wikipedia knowledge, in addition to training on the discharge summaries, and showed some improvements in the predictability of the system. However, we argue that such external-data-based boosting approaches often trade-off model interpretability for higher accuracy of predictions. The mappings between the underlying clinical text and the corresponding diagnostic codes are often blurred in such models. For clinical decision support systems to be adaptable in real-world scenarios, providing an explainable decision (even when incorrect) is far more acceptable than just producing a highly accurate black-box decision. Table 13 presents several examples of attention visualization, comparing the single-channel convolutional attention model [27] to the proposed *EnCAML* model.

The CodiEsp Spanish clinical notes corpus presents compact text $n-$grams extracted from the notes' content as evidence for the ICD-10 code(s) assigned to the respective notes. This provides an unprecedented opportunity to benchmark the interpretability of the proposed *EnCAML* model using manually-annotated data. For a given clinical note $\mathcal{D}^{(d)} = \{t_1^{(d)}, t_2^{(d)}, \ldots, t_L^{(d)}\}$ comprising L tokens $t_i^{(d)}$s, let $P_{10}^{(d,c)} \subseteq \mathcal{D}^{(d)}$ be the set of top-10 clinical text tokens that contribute most to the predicted ICD-10 code c, obtained using the attention weights a_cs. Now, let $E^{(d,c)}$ be the set of tokens obtained from the expert-provided evidence for a clinical note $\mathcal{D}^{(d)}$ mapping to the actual ICD-10 diagnostic code c. From the inspection of the CodiEsp corpus, we have $|E^{(d,c)}| \leq |P_{10}^{(d,c)}|$. We compute the overall interpretability score ($\mathcal{I} \in [0,1]$) for the proposed *EnCAML* model as follows:

$$\mathcal{I} = \sum_{d=1}^{D} \sum_{c=1}^{N} \frac{\mathbf{1}\{(E^{(d,c)} \neq \phi) \wedge (E^{(d,c)} \subseteq P_{10}^{(d,c)})\}}{\mathbf{1}\{E^{(d,c)} \neq \phi\}} \tag{7}$$

Notice that the interpretability score penalizes false-negative scenarios, i.e., cases where the attention-based evidence fails to capture all the contributing tokens specified by the manually-annotated evidence. Table 14 presents few sample clinical notes from the CodiEsp corpus, demonstrating the predictability and interpretability of our proposed neural model. Using (7), we obtained $\mathcal{I}$ scores of 0.9550 and 0.9130 for *top-10-ce-code* and *top-50-ce-code* CodiEsp data categories, respectively. These recorded high values of $\mathcal{I}$ scores corroborate an extensive overlap between the expert-annotated textual evidence and the attention-output-based evidence obtained from the proposed *EnCAML* model, thus establishing the enhanced interpretability of the proposed system.

5.6. Ablation Study

In this subsection, we report the findings from our ablation study aimed at establishing the contributions of various modules in the proposed diagnostic code prediction system. The study was performed using the discharge summaries in the extensively-benchmarked *top-50-dp-code* data category, and the results are summarized in Table 15. The results indicate that replacing multi-channel variable-sized convolutional attention layers with a single-channel model degrades the prediction performance of the neural system significantly. Additionally, it can be seen that each component in the proposed system, including preprocessing, multi-channel variable-size kernels, and the optimal threshold setting contributed towards improving the overall predictability of ICD diagnostic codes. Moreover, Table 15 also presents the total number of trainable parameters obtained per model. It can be seen that our preprocessing pipeline reduces the number of trainable parameters in the order of 10e6. Furthermore, our proposed *EnCAML* model with multi-channel variable-sized convolutional attention layers employs considerably less trainable parameters than a more straightforward single-channel model. As detailed in the previous subsection,

Table 15: The results from the ablation study of major components in the proposed system for the prediction of codes in the widely-benchmarked *top-50-dp-code* data category ($\|$ denotes parallel convolutions with varying kernel sizes).

Model	F_1 **micro**	**Total parameters**
Preprocessing (§ 3.3) + multi-channel CAML (§ 4) + Fisher-Jenks thresholds (§ 4)	**0.6764**	5.58e6
Preprocessing (§ 3.3) + multi-channel CAML (§ 4)	$0.6698^{-\theta}$	$5.58\text{e}6^{-\theta}$
Preprocessing (§ 3.3) + single-channel CAML ($e = 100,\ k = 10,\ f = 300$)	$0.6197^{-\theta, -\|}$	$4.27\text{e}6^{-\theta, -\|}$
Single-channel CAML ($e = 100,\ k = 10,\ f = 300$)	$0.6138^{-\theta, -\|, -\text{pre}}$	$6.14\text{e}6^{-\theta, -\|, -\text{pre}}$

the model interpretability achieved using the attention weights of the proposed model is on-par with that facilitated by expert medical coders. Finally, the flexibility, robustness, and enhanced interpretability of the proposed *EnCAML* model establish the extensive adaptability of our approach for rapid prototyping and deployment in developing nations with low digitization rates.

6. Conclusion

Enabling diagnostic code assignment is vital for clinical decision support, epidemiology, billing, and managing hospital resources; however, manual facilitation of such assignment is often error-prone and time-consuming. In this study, we proposed *EnCAML*, a multi-channel variable-sized convolutional attention model, to enable the clinical task of diagnostic code assignment as a multi-label classification problem. We demonstrated that the proposed model enhances the code predictability by extracting multi-granular text snippets, using which the attention mechanism enables the selection of those segments that are most contributing to the corresponding diagnostic code. Our extensive benchmarking against several state-of-the-art models, including convolution-based models, sequence models, single-channel convolutional attention models, models employing transfer learning, and others, revealed the efficacy of our proposed approach in modeling noisy, unstructured discharge summaries of the MIMIC-III corpus. In part, we attribute our reported high performance to the proposed preprocessing pipeline, which facilitated the effective pruning of irrelevant content in the free-text summaries. Furthermore, to demonstrate the robustness and adaptability of our proposed model, we established the minimal effect of the choice of initial embedding layer on the overall performance. We also presented our promising results in modeling a more convoluted ICD-10 coding taxonomy employed in the CodiEsp Spanish clinical notes corpus, thereby exhibiting the flexibility and language-agnostic nature of the proposed system. Finally, we demonstrated the enhanced interpretability of the predictions output by the *EnCAML* model using the learned per-code attention weights, thereby establishing the impact of the proposed model on instigating trust in the proposed intelligent healthcare system.

In the future, we aim at extending the model and approaches presented in this study to accommodate alternate sources of patient data, especially in cases where the underlying discharge summaries are rather uninformative. Additionally, we propose to explore the challenge of patient profiling via automated generation of summarized and well-formatted reports, sourced from multiple patient data sources, including discharge summaries, nursing notes, radiology reports, and various others. Such aggregated, rich semi-structured data can then be employed in enhancing the interpretability and predictability of the underlying clinical decision support systems.

Acknowledgments

This research was funded by the DST-SERB Early Career Research Grant ECR/2017/001056, provided by the Government of India. Any findings, opinions, and recommendations or conclusions expressed in this study are those of the authors and do not reflect the views of the funding agency.

Declaration of Competing Interests

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work.

References

[1] Vergil N. Slee. The International Classification of Diseases: Ninth Revision (ICD-9). *Annals of Internal Medicine*, 88(3):424–426, 03 1978. ISSN 0003-4819. doi: 10.7326/0003-4819-88-3-424. 2

[2] World Health Organization. *ICD-10 : International statistical lassification of iseases and related health problems / World Health Organization.* World Health Organization Geneva, 10th revision, 2nd ed. edition, 2004. ISBN 9241546492 9241546530 9241546549. 2

[3] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395, 2012. 2

[4] M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, and J. Wang. Automated icd-9 coding via a deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(4):1193–1202, 2019. 2, 3, 6, 7, 19, 22, 25, 29

[5] YunZhi Chen, HuiJuan Lu, and LanJuan Li. Automatic icd-10 coding algorithm using an improved longest common subsequence based on semantic similarity. *PloS one*, 12(3):e0173410, 2017. 2, 6

[6] Min Zeng, Min Li, Zhihui Fei, Ying Yu, Yi Pan, and Jianxin Wang. Automatic icd-9 coding via deep transfer learning. *Neurocomputing*, 324:43–50, 2019. 2, 3, 7, 8, 25

[7] Dee Lang. Consultant report-natural language processing in the health care industry. *Cincinnati Children's Hospital Medical Center, Winter*, 6, 2007. 2

[8] Richárd Farkas and György Szarvas. Automatic construction of rule-based icd-9-cm coding systems. In *BMC bioinformatics*, volume 9, page S10. BioMed Central, 2008. 2, 6

[9] Jose C Ferrao, Filipe Janela, Mónica D Oliveira, and Henrique MG Martins. Using structured ehr data and svm to support icd-9-cm coding. In *2013 IEEE International Conference on Healthcare Informatics*, pages 511–516. IEEE, 2013. 3

[10] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237, 2013. 3

[11] Sen Wang, Xue Li, Lina Yao, Quan Z Sheng, Guodong Long, et al. Learning multiple diagnosis codes for icu patients with local disease correlation mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(3):31, 2017. 3

[12] Serguei VS Pakhomov, James D Buntrock, and Christopher G Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association*, 13(5):516–525, 2006. 3

[13] Julia Medori and Cédrick Fairon. Machine learning and features selection for semi-automatic icd-9-cm encoding. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 84–89. Association for Computational Linguistics, 2010. 3

[14] Patrick Ruch, Julien Gobeill, Imad Tbahriti, and Antoine Geissbühler. From episodes of care to diagnosis codes: automatic text categorization for medico-economic encoding. In *AMIA Annual Symposium Proceedings*, volume 2008, page 636. American Medical Informatics Association, 2008. 3

[15] Madhav Erraguntla, Belita Gopal, Satheesh Ramachandran, and Richard Mayer. Inference of missing icd 9 codes using text mining and nearest neighbor techniques. In *2012 45th Hawaii International Conference on System Sciences*, pages 1060–1069. IEEE, 2012. 3

[16] Adler J Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. Hierarchically supervised latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2011. 3

[17] Mohamed Dermouche, Julien Velcin, Rémi Flicoteaux, Sylvie Chevret, and Namik Taright. Supervised topic models for diagnosis code assignment to discharge summaries. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 485–497. Springer, 2016. 3

[18] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, Sep 2018. ISSN 2168-2208. doi: 10.1109/jbhi.2017.2767063. URL http://dx.doi.org/10.1109/JBHI.2017.2767063. 3, 4

[19] Paschalis Bizopoulos and Dimitrios Koutsouris. Deep learning in cardiology. *IEEE Reviews in Biomedical Engineering*, 12:168–193, 2019. ISSN 1941-1189. doi: 10.1109/rbme.2018.2885714. URL http://dx.doi.org/10.1109/RBME.2018.2885714. 3, 4

[20] Inês Domingues, Gisèle Pereira, Pedro Martins, Hugo Duarte, João Santos, and Pedro Henriques Abreu. Using deep learning techniques in medical imaging: a systematic review of applications on ct and pet. *Artificial Intelligence Review*, Nov 2019. ISSN 1573-7462. doi: 10.1007/s10462-019-09788-3. URL https://doi.org/10.1007/s10462-019-09788-3. 3, 4, 6

[21] Mary Jane C Samonte, Bobby D Gerardo, Arnel C Fajardo, and Ruji P Medina. Icd-9 tagging of clinical notes using topical word embedding. In *Proceedings of the 2018 International Conference on Internet and e-Business*, pages 118–123. ACM, 2018. 3

[22] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3, 6, 7, 8, 25, 29

[23] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer Methods and Programs in Biomedicine*, 177:141–153, Aug 2019. ISSN 0169-2607. doi: 10.1016/j.cmpb.2019.05.024. URL http://dx.doi.org/10.1016/j.cmpb.2019.05.024. 3, 6, 7, 8, 11, 18, 20, 25, 29

[24] Pengtao Xie and Eric Xing. A neural architecture for automated icd coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076, 2018. 3, 6, 7

[25] Mary Jane C. Samonte, Bobby D. Gerardo, Arnel C. Fajardo, and Ruji P. Medina. Icd-9 tagging of clinical notes using topical word embedding. In *Proceedings of the 2018 International Conference on Internet and e-Business*, ICIEB '18, pages 118–123, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-6375-4. doi: 10.1145/3230348.3230357. URL http://doi.acm.org/10.1145/3230348.3230357. 3, 6, 25

[26] Anthony Rios and Ramakanth Kavuluru. Neural transfer learning for assigning diagnosis codes to EMRs. *Artificial Intelligence In Medicine*, 96 (December 2018):116–122, 2019. ISSN 0933-3657. doi: 10.1016/j.artmed.2019.04.002. URL https://doi.org/10.1016/j.artmed.2019.04.002. 3, 6, 7, 25

[27] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://www.aclweb.org/anthology/N18-1100. 3, 4, 6, 7, 8, 11, 15, 18, 20, 25, 26, 28, 29, 33, 34

[28] Fei Li and Hong Yu. Icd coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020. 3, 6, 7, 8, 25

[29] Fei Teng, Wei Yang, Li Chen, LuFei Huang, and Qiang Xu. Explainable prediction of medical codes with knowledge graphs. *Frontiers in Bioengineering and Biotechnology*, 8:867, 2020. ISSN 2296-4185. doi: 10.

3389/fbioe.2020.00867. URL https://www.frontiersin.org/article/10.3389/fbioe.2020.00867. 3, 6, 7, 8, 34

[30] Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 73–78, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.clinicalnlp-1.8. URL https://www.aclweb.org/anthology/2020.clinicalnlp-1.8. 3, 6, 7, 8

[31] Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for icd coding from clinical text. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/461. URL https://doi.org/10.24963/ijcai.2020/461. Main track. 3, 6, 7, 8, 34

[32] Tushaar Gangavarapu, Aditya Jayasimha, Gokul S. Krishnan, and Sowmya Kamath S. Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowledge-Based Systems*, page 105321, 2019. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2019.105321. URL http://www.sciencedirect.com/science/article/pii/S0950705119305982. 3, 4, 6, 7, 11, 22

[33] Tushaar Gangavarapu, Aditya Jayasimha, Gokul S Krishnan, and Sowmya Kamath S. TAGS: Towards Automated Classification of Unstructured Clinical Nursing Notes. In *International Conference on Applications of Natural Language to Information Systems*, pages 195–207. Springer, 2019. URL https://doi.org/10.1007/978-3-030-23281-8_16. 3, 6, 7, 22

[34] Aditya Jayasimha, Tushaar Gangavarapu, Sowmya Kamath S, and Gokul S Krishnan. Deep Neural Learning for Automated Diagnostic Code Group Prediction Using Unstructured Nursing Notes. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CoDS-COMAD '20, pages 152–160, New York, NY, USA, 2020. ACM. URL https://doi.org/10.1145/3371158.3371176. 3, 6, 7, 18, 22

[35] Tushaar Gangavarapu, Gokul S Krishnan, and Sowmya Kamath S. Coherence-based modeling of clinical concepts inferred from heterogeneous clinical notes for ICU patient risk stratification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1012–1022, Hong Kong, China, November 2019. Association for

Computational Linguistics. doi: 10.18653/v1/K19-1095. URL https://www.aclweb.org/anthology/K19-1095. 3, 4, 6, 7, 8, 11, 18, 22

[36] Tushaar Gangavarapu, Gokul S Krishnan, Sowmya S Kamath, and Jayakumar Jeganathan. Farsight: Long-term disease prediction using unstructured clinical nursing notes. *IEEE Transactions on Emerging Topics in Computing*, 2020. doi: 10.1109/TETC.2020.2975251. URL https://ieeexplore.ieee.org/document/9007352. 3, 4, 6, 7, 11, 18, 22

[37] Donglin Guo, Guihua Duan, Ying Yu, Yaohang Li, Fang-Xiang Wu, and Min Li. A disease inference method based on symptom extraction and bidirectional long short term memory networks. *Methods*, 2019. ISSN 1046-2023. doi: https://doi.org/10.1016/j.ymeth.2019.07.009. URL http://www.sciencedirect.com/science/article/pii/S1046202319301033. 3, 8, 25, 29

[38] Keyang Xu, Mike Lam, Jingzhi Pang, Xin Gao, Charlotte Band, Piyush Mathur, Frank Papay, Ashish K. Khanna, Jacek B. Cywinski, Kamal Maheshwari, Pengtao Xie, and Eric P. Xing. Multimodal machine learning for automated icd coding. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 197–215, Ann Arbor, Michigan, 09–10 Aug 2019. PMLR. URL http://proceedings.mlr.press/v106/xu19a.html. 3, 6

[39] Ssu-Ming Wang, Yu-Hsuan Chang, Lu-Cheng Kuo, Feipei Lai1 andYun Nung Chen, Fei-Yun Yu, and Chih-Wei Chen. Using deep learning for automated icd-10 classification from free text data. *EJBI*, pages 1–10, 01 2020. doi: 10.24105/ejbi.2020.16.1.1. 3, 6

[40] Jeff Marshall, Abdullah Chahin, and Barret Rush. *Review of Clinical Databases*, pages 9–16. Springer International Publishing, Cham, 2016. ISBN 978-3-319-43742-2. doi: 10.1007/978-3-319-43742-2_2. URL https://doi.org/10.1007/978-3-319-43742-2_2. 3

[41] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016. 3

[42] Antonio Miranda, Aitor Gonzalez-Agirre, and Martin Krallinger. CodiEsp corpus: Spanish clinical cases coded in ICD10 (CIE10) - eHealth CLEF2020, March 2020. URL https://doi.org/10.5281/zenodo.

3693570. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL). 4

[43] Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, and Martin Krallinger. Clef ehealth evaluation lab 2020. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, pages 587–594, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45442-5. 4

[44] Julia Medori and Cédrick Fairon. Machine learning and features selection for semi-automatic icd-9-cm encoding. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, Louhi '10, pages 84–89, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1867735.1867748. 6

[45] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112 – 134, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2018.04.007. URL http://www.sciencedirect.com/science/article/pii/S1532046418300716. 6

[46] Elyne Scheurwegs, Boris Cule, Kim Luyckx, Léon Luyten, and Walter Daelemans. Selecting relevant features from the electronic health record for clinical code prediction. *Journal of Biomedical Informatics*, 74:92 – 103, 2017. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2017.09.004. URL http://www.sciencedirect.com/science/article/pii/S1532046417302010. 6

[47] Luciano R. S. de Lima, Alberto H. F. Laender, and Berthier A. Ribeiro-Neto. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, pages 132–139, New York, NY, USA, 1998. ACM. ISBN 1-58113-061-9. doi: 10.1145/288627.288649. URL http://doi.acm.org/10.1145/288627.288649. 6

[48] Agnieszka Mykowiecka, Małgorzata Marciniak, and Anna Kupść. Rule-based information extraction from patients' clinical data. *Journal of Biomedical Informatics*, 42(5):923 – 936, 2009. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2009.07.007. URL http://www.sciencedirect.com/science/article/pii/S1532046409001002. Biomedical Natural Language Processing. 6

[49] Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and Steven Carroll. Automatic code assignment to medical text. In

Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, BioNLP '07, pages 129–136, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1572392.1572416. 6

[50] Serguei Pakhomov, James Buntrock, and Christopher Chute. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *Journal of the American Medical Informatics Association : JAMIA*, 13:516–25, 09 2006. doi: 10.1197/jamia.M2077. 6

[51] Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine*, 65, 05 2015. doi: 10.1016/j.artmed.2015.04.007. 6

[52] Leah S Larkey and W Bruce Croft. Automatic assignment of icd9 codes to discharge summaries. Technical report, Technical report, University of Massachusetts at Amherst, Amherst, MA, 1995. 6

[53] Hanna Suominen, Filip Ginter, Sampo Pyysalo, Antti Airola, Tapio Pahikkala, Sanna Salanterä, and Tapio Salakoski. Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. *In: Proceedings of the ICML/UAI/COLT Workshop on Machine Learning for Health-Care Applications*, 11 2007. 6

[54] H. Harutyunyan, H. Khachatrian, D.C. Kale, G. Ver Steeg, and A. Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), 2019. ISSN 20524463. doi: 10.1038/s41597-019-0103-9. 6, 11

[55] William Caicedo-Torres and Jairo Gutierrez. Iseeu: Visually interpretable deep learning for mortality prediction inside the icu. *Journal of Biomedical Informatics*, 98:103269, 2019. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2019.103269. URL http://www.sciencedirect.com/science/article/pii/S1532046419301881. 6

[56] Gokul S. Krishnan and Sowmya Kamath S. A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data. *Applied Soft Computing*, 80:525 – 533, 2019. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2019.04.019. URL http://www.sciencedirect.com/science/article/pii/S1568494619302108. 6

[57] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent

neural networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 301–318, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR. URL http://proceedings.mlr.press/v56/Choi16.html. 6

[58] Jingshu Liu, Zachariah Zhang, and Narges Razavian. Deep ehr: Chronic disease prediction using medical notes. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 440–464, Palo Alto, California, 17–18 Aug 2018. PMLR. URL http://proceedings.mlr.press/v85/liu18b.html. 6, 7

[59] T. Gentimis, A. J. Alnaser, A. Durante, K. Cook, and R. Steele. Predicting hospital length of stay using neural networks on mimic iii data. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 1194–1201, Nov 2017. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191. 6

[60] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh. Deepr: A convolutional net for medical records. *IEEE Journal of Biomedical and Health Informatics*, 21(1):22–30, 2017. ISSN 21682194. doi: 10.1109/JBHI.2016.2633963. 6

[61] Ahmad Hammoudeh, Ghazi Al-Naymat, Ibrahim Ghannam, and Nadim Obied. Predicting hospital readmission among diabetics using deep learning. *Procedia Computer Science*, 141:484 – 489, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.10.138. URL http://www.sciencedirect.com/science/article/pii/S1877050918317873. The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops. 6

[62] Awais Ashfaq, Anita Sant'Anna, Markus Lingman, and Sławomir Nowaczyk. Readmission prediction using deep learning on electronic health records. *Journal of Biomedical Informatics*, 97:103256, 2019. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2019.103256. URL http://www.sciencedirect.com/science/article/pii/S1532046419301753. 6

[63] Sina Rashidian, Janos Hajagos, Richard Moffitt, Fusheng Wang, Xinyu Dong, Kayley Abell-Hart, Kimberly Noel, Rajarsi Gupta, Mathew Tharakan, Veena Lingam, Joel Saltz, and Mary Saltz. Disease phenotyping using deep learning: A diabetes case study. *arXiv e-prints*, art. arXiv:1811.11818, November 2018. 6

[64] T. Pham, T. Tran, D. Phung, and S. Venkatesh. Deepcare: A deep dynamic memory model for predictive medicine. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9652 LNAI:30–41, 2016. ISSN 03029743. doi: 10.1007/978-3-319-31750-2_3. 6

[65] Finneas Catling, Georgios P. Spithourakis, and Sebastian Riedel. Towards automated clinical coding. *International Journal of Medical Informatics*, 120:50 – 61, 2018. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2018.09.021. URL http://www.sciencedirect.com/science/article/pii/S1386505618304039. 7

[66] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1181. URL https://www.aclweb.org/anthology/D14-1181. 7

[67] Anthony Rios and Ramakanth Kavuluru. Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '15, pages 258–267, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3853-0. doi: 10.1145/2808719.2808746. URL http://doi.acm.org/10.1145/2808719.2808746. 7

[68] M. A. Parwez, M. Abulaish, and Jahiruddin. Multi-label classification of microblogging texts using convolution neural network. *IEEE Access*, 7: 68678–68691, 2019. doi: 10.1109/ACCESS.2019.2919494. 7

[69] Yuqi Si and Kirk Roberts. Deep patient representation of clinical notes via multi-task learning for mortality prediction. *AMIA Summits on Translational Science Proceedings*, 2019:779, 2019. 7

[70] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. International Conference on Learning Representations, ICLR, 2013. 8, 14, 22

[71] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the*

Association for Computational Linguistics, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://www.aclweb.org/anthology/Q17-1010. 8, 23

[72] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423. 8, 23

[73] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. 8

[74] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014. 8

[75] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 8

[76] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 8, 31

[77] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017. URL http://arxiv.org/abs/1711.05225. 8

[78] Incheol Kim, Sivaramakrishnan Rajaraman, and Sameer Antani. Visual Interpretation of Convolutional Neural Network Predictions in Classifying Medical Image Modalities. *Diagnostics (Basel)*, 9(2):38, April 2019. doi: 10.3390/diagnostics9020038. 8

[79] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2015. 8

[80] H. Liu, Q. Yin, and W.Y. Wang. Towards explainable nlp: A generative explanation framework for text classification. pages 5570–5581. Association for Computational Linguistics (ACL), 2020. ISBN 9781950737482. 8

[81] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5408. URL https://www.aclweb.org/anthology/W18-5408. 8

[82] Ning Wang, MINGXUAN CHEN, and Koduvayur P. Subbalakshmi. Explainable cnn-attention networks (c-attention network) for automated detection of alzheimer's disease. *medRxiv*, 2020. doi: 10.1101/2020.06.24.20139592. URL https://www.medrxiv.org/content/early/2020/06/26/2020.06.24.20139592. 8

[83] Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–18, 2020. ISSN 2162-2388. doi: 10.1109/tnnls.2020.3019893. URL http://dx.doi.org/10.1109/TNNLS.2020.3019893. 8

[84] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://www.aclweb.org/anthology/N19-1357. 8

[85] Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1282. URL https://www.aclweb.org/anthology/P19-1282. 8

[86] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://www.aclweb.org/anthology/D19-1002. 8

[87] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. Attention Interpretability Across NLP Tasks. *arXiv e-prints*, art. arXiv:1909.11218, September 2019. 8

[88] Ryan McDonald, George Brokos, and Ion Androutsopoulos. Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1860, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1211. URL https://www.aclweb.org/anthology/D18-1211. 13, 14

[89] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, January 1974. ISSN 0004-5411. doi: 10.1145/321796.321811. URL https://doi.org/10.1145/321796.321811. 14

[90] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. 14, 23

[91] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 18

[92] G. F. JENKS. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190, 1967. URL https://ci.nii.ac.jp/naid/10021899676/en/. 19

[93] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 20

[94] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *In Data Mining and Knowledge Discovery Handbook*, pages 667–685, 2010. 21

[95] Tushaar Gangavarapu, CD Jaidhar, and Bhabesh Chanduka. Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artificial Intelligence Review*, pages 1–63, 2020. 21

[96] Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 44:223–70, 01 1908. doi: 10.5169/seals-268384. 21

[97] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://www.aclweb.org/anthology/W19-1909. 23

[98] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65, 2019. 23