IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 July 2023, accepted 13 September 2023, date of publication 28 September 2023, date of current version 4 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3320179

RESEARCH ARTICLE

HSGA: A Hybrid LSTM-CNN Self-Guided Attention to Predict the Future Diagnosis From Discharge Narratives

GASPARD HARERIMANA[1], (Member, IEEE), GUN IL KIM[2], (Member, IEEE), JONG WOOK KIM[3], (Member, IEEE), AND BEAKCHEOL JANG[2], (Member, IEEE)

[1]Department of Information Technology, Carnegie Mellon Universityy, Kigali, Rwanda
[2]Graduate School of Information, Yonsei University, Seoul 03722, South Korea
[3]Department of Computer Science, Sangmyung University, Seoul 03016, South Korea

Corresponding authors: Beakcheol Jang (bjang@yonsei.ac.kr) and Jong Wook Kim (jkim@smu.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government under Grant RS-2023-00273751 and Grant NRF-2023R1A2C1004919.

ABSTRACT The prognosis of a patient's re-admission and the forecast of future diagnoses is a critical task in the process of inferring clinical outcomes. The discharge summaries recorded in the Electronic Health Records (EHR) are stinking rich, but they are also heterogeneous, sparse, noisy, and biased, and hinder the learning algorithms that aim to extract actionable insights from them. The existing approaches use the current admission's International Classification of Diseases (ICD) codes as the input features, but they do not fully describe the progression of the patient. Other systems apply the attention mechanisms directly to these notes without the guidance of domain knowledge, resulting in distorted predictions. In this work, we propose a hybrid LSTM-CNN self-guided attention model that aims to predict the ICD diagnosis that is likely to cause the next readmission within 90 days since the current discharge using the discharge narratives. Since the notes contain unnecessary tokens, the model leverages the recent advances in deep learning to predict the patient's future diagnosis by reducing the number of tokens from the notes to be considered for prediction. We use a 1D CNN (1-Dimensional Convolutional Neural Network) to capture all features from the note and concurrently an LSTM (Long Short-Term Memory) is used to extract the features of clinically meaningful Concept Unique Identifiers (CUI) that are fetched from the note itself to build a knowledge base. The textual knowledge base guides the learning module about which n-grams from the note to focus on for prediction. We consider 3 prediction scenarios: diagnosis category prediction, the probability of the occurrence of one of the top 20 disease conditions, and ICD9 codes prediction. For the diagnosis category prediction, our proposed model achieves a macro-average ROC of 0.82 and a micro-average ROC of 0.79, an AUROC of 0.87 for the top 20 most appearing diseases prediction, and a macro-average Recall of 0.8 and a micro-average Recall of 0.84 for ICD9 codes prediction respectively. The predictive accuracy of the model is assessed through the prediction of heart failure onset and for all these prediction scenarios, the results show that the hybrid approach outperforms the existing baselines.

INDEX TERMS Electronic health records, 1D convolutional neural networks, concept unique identifier, guided attention, long short term memory.

I. INTRODUCTION

The patients in the EHR can be represented as sequences of visits with each visit containing critical records like diagnoses, procedures, medications and vital signs

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

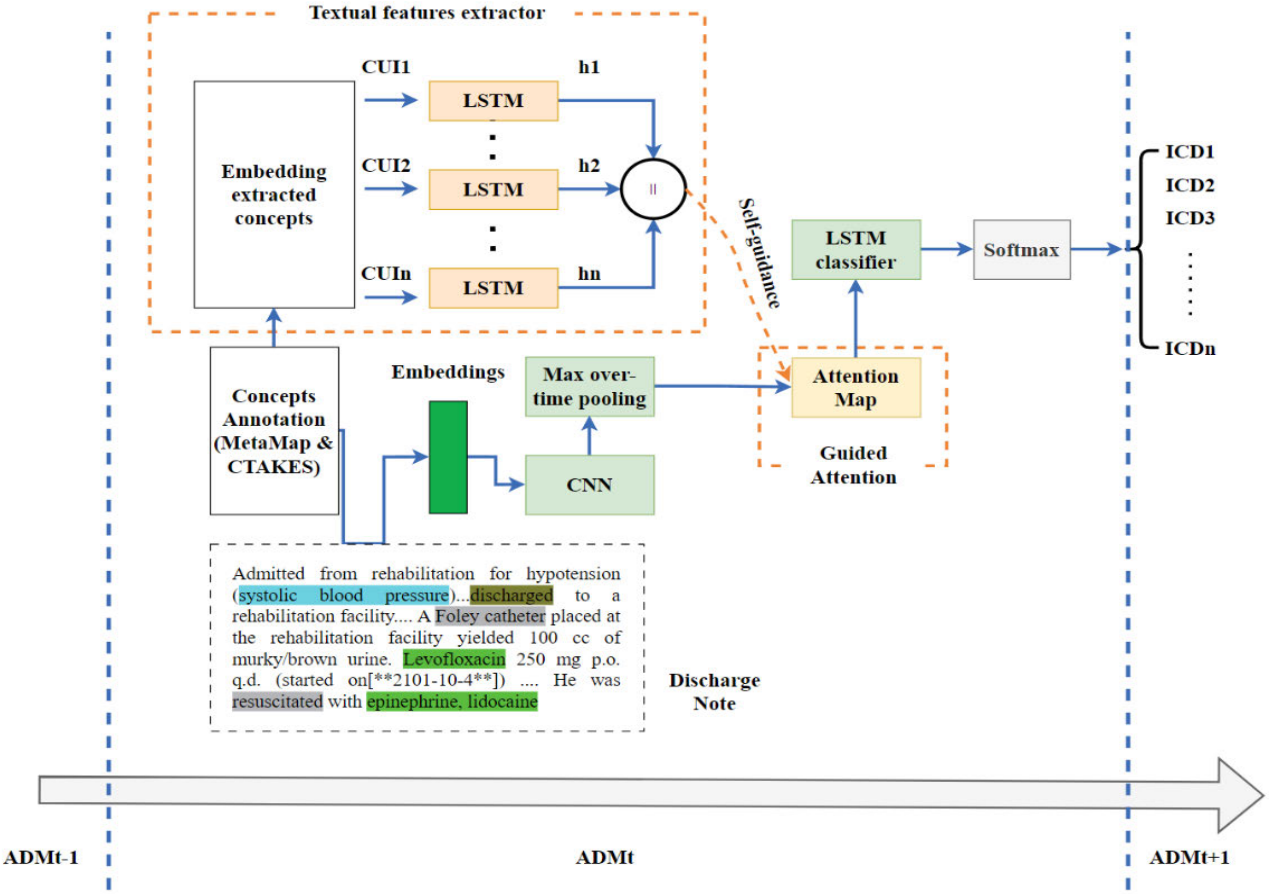

FIGURE 1. The overall structure of HSGA. The clinically meaningful tokens are first extracted, and their features obtained using an LSTM. These features are used to guide the CNN based attention map which expresses the key n-grams of the same clinical note that are most influential in the overall prediction.

charted events. The discharge notes are textual narratives recorded during admission and are commonly supplemented by formal records in the form of metadata codes like ICD (International Classification of Diseases) codes for diagnoses and procedures, as well as LOINC (Logical Observation Identifiers Names and Codes) for lab records. The most critical task of any EHR based analytic solution is to predict the adverse future events including the risks of future readmission and the disease (ICD diagnosis) that will likely be the culprit. A classic prospective cohort study [1] found out that 27% of discharged patients had preventable adverse events with ADE (Adverse Drugs Events) being the most occurring type of event (66%), followed by procedure-related injuries (17%). The ability to forecast the likely diagnosis codes in subsequent patient's visit on current discharge can improve the diagnostic process and help to deter the preventable adverse post discharge events. Clinical notes recorded in the EHR have been recently used to predict various risk factors and have proved to be the best candidates for such predictions [2], [3], [4], [5], [6]. The attention mechanism [7] is a recent improvement in the area of deep learning that improves the prediction capabilities of deep learning models and has been recently applied in various fields such as machine comprehension [8], [9], visual question answering (VQA) [10], [11] and text classification [12], [13] tasks. The attention mechanism and its variants have recently been applied and customized for clinical based predictions like mortality [14], phenotyping [15], decompensation [16], ICU readmission [17], and ultimately the ICD diagnosis codes that are likely to appear in the readmission [18], [19]. The existing studies use the discharge notes directly by applying attention mechanism with flat CNN, flat LSTM, the combination of both or simply using attention without convolution or recurrence [20]. In the quest for future diagnosis prediction using deep learning, the following issues need to be considered and addressed:

- Using discharge notes without having the domain knowledge to guide the learning algorithm on which regions of the notes to focus on will results in distorted predictions as they contain many abbreviations and unnecessary tokens.
- Building a domain knowledge and attention mechanism by combining these notes and ICD diagnosis codes for automatic ICD coding as in [21] can be

erroneous because the ICD codes do not contain patient's progression. As an example, ICD diagnosis code '486' means that the patient was diagnosed with pneumonia, but the ICD diagnosis list does not reveal his progress as well as related hidden comorbidities.

To bridge this inferential gap, we propose HSGA, a hybrid LSTM-CNN self-guided attention model that builds the domain knowledge from the clinical note itself by extracting its key UMLS CUIs (Unified Medical Language System Concept Unique Identifier) in a multi-modal pattern. The obtained knowledge base is used on the same clinical note to guide the attention process about the tokens that are correlated to the extracted CUIs inferring that they are the most influential for the final prediction of future diagnosis. This guidance facilitates the attention process by reducing the number of parameters to be learned during the training process. To properly capture the future conditions that may be related to the current discharge, we consider the future admission that happens within 90 days from the current discharge. Such problem is well studied and addressed through 3 prediction scenarios. The first scenario is top 20 which is a binary classification to predict if the patient is likely to be re-admitted due to one of the top 20 occurring diseases. The second scenario is a multi-label classification to predict the occurrence probability of one of the 285 diagnosis grouped under the CCS (Clinical Classification Software) categories. The last scenario is to predict the ICD9 codes. As depicted in Figure. 1, it shows the LSTM cells that are used for textual features extraction from the extracted concepts and these features are used to guide a CNN based attention map that helps to predict future diagnosis. To demonstrate the superiority of our approach, we use a large EHR dataset to train and evaluate the model with k-fold cross validation [22] and for each prediction scenario, the performance comparison with existing baselines is recorded. We achieve a macro-average ROC (Receiver Operating Curve) of 82% for diagnosis category prediction, an AUROC (Area Under the ROC) of 87% for the occurrence probability for the Top 20 diseases conditions, and a micro-recall of 84% for exact ICD codes predictions. Our findings show that the proposed approach can achieve better diagnosis prediction than state-of-the-art existing baselines. The improvement in macro- and micro-average ROC values obtained by our approach compared to the related baseline without the hybrid guided attention component (CNN + attention) is around 10%. To further validate our model, we investigate the prediction of heart failure onset using discharge summaries. Briefly, our major contributions are as follows:

- We propose a method that only uses discharge summaries into a multi-modal design with clinical concepts as the textual features guiding the attention mechanism, to predict the likely diagnosis for a readmission that occurs within 90 days from discharge.
- The extraction of key UMLS CUIs combines the robustness of the most popular clinical annotators, MetaMap and cTAKES.
- We embed the resulting CUIs using the pre-trained embedding vectors from a proven embedding approach.
- The problem is well studied and addressed through 3 prediction scenarios. For each scenario, the performance comparison with existing baselines is recorded.
- For the two networks (attention model and features extraction), we use the same pre-trained embedding vectors to capture the tokens from the note that are related to the extracted concepts.
- The capabilities of the model are assessed by applying it for the prediction of heart failure onset.

In this study, we use the MIMIC-III EHR [23], a publicly available critical care database that integrates the de-identified, comprehensive clinical data of patients with a total of 53,423 admissions at the Intensive Care Unit (ICU) of the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts during the period from 2001 to 2012. The remainder of this paper is arranged as follows: in section II, we present a breadth of key related works, in section III we briefly describe the key preliminaries and, in section IV we briefly discuss the data pre-processing procedure. In section V, we describe our proposed hybrid model, in section VI we perform the experiments, in section VII we present our results as well as performance evaluation, and in section. 8 we conclude with the conclusion.

II. RELATED WORKS AND BACKGROUND KNOWLEDGE

A. RELATED WORKS

The prediction of a patient's future, as well as clinical notes-based predictions have recently received a big deal of research attention. Most of the existing approaches utilize the structured EHR's ICD codes including the past records to predict future diagnoses. However, many of these models suffer from the general problem of high dimensionality due to the sheer size of the discharge notes. Lipton et al. [24] proposed an LSTM based model that predicts diagnoses from irregularly sampled multivariate records like heart rate, systolic blood pressure, and glucose levels. The study did not yield a robust interpretable predictive schema but highlighted a promising source of future diagnosis prediction. DIPOLE [25] is a category level diagnosis prediction approach that utilizes a bi-directional RNN and successive attention mechanisms to model the relationships between the patient's visits. However, dipole suffers greatly when the training dataset is small, and it does not perform well for the rare diseases whose presence in the EHR is minimal. Similarly, Beakcheol et al. [26], [27] investigated the correlation between the actual outbreak of major infectious diseases using Internet-based datasets such as Twitter and official Korean Center of Disease Control (KCDC) using the sequence-to-sequence model. RETAIN [28] utilizes a double attention method for EHR based predictions by leveraging the previous visits information. The two-level attention mechanism captures the most influential visits and most influential codes within each of these visits. Ma et al. [29]

proposed KAME, a knowledge-level attention mechanism to predict future visit data for patients with clinical oncology that uses medical knowledge throughout the prediction process. As is usually the case with other approaches that use medical ontologies for prediction, they lack the most crucial information about the patient's progressive history, hence many of the hidden comorbidities are not properly captured. Authors in [30] have developed a model that classifies clinical notes using BERT (Bidirectional Encoder Representations from Transformers) [31], state-of-the-art transfer learning-based language model for NLP tasks that utilize the transformer model, which is a type of attention mechanism that can learn contextual dependencies between words in a text. Kim et al. [32] also probed in predicting medical specialty from patient-side medical question text by pretraining with the BERT model. However, the model does not try to predict the future of the patient, but rather creates a human interpretable clinical text classifier for a more eased analysis. Peng et al. [33] built a self-attention model that automatically extracts lesion features in a region of interest from CT (Computed Tomography) reports. Qiao et al. [21] developed MMN (Multimodal Attentional Neural Networks) a multimodal attention network that combines diagnoses and clinical text to predict future diagnoses. The model first builds multimodal features and a deep feature mixer before prediction. However, the study does not elaborate on the approach used for predicting ICD-9 codes as the prediction will require the classifier to output the ICD codes sequentially or in a multi-class pattern. Mullenbach et al. [34] have proposed CAML (Convolutional Attention for Multi-Label classification), a convolutional attention model that predicts medical codes from clinical text using a per-label approach. Similarly, with the trend of Explainable Artificial Intelligence (XAI), Trigueros et al. [35] experimented in a multi-label classification task on obtaining explainable predictions of the diseases and procedures contained in EHRs based on the ICD reports using convolutional attention mechanism. Harerimana et al. [36] used the hierarchical attention network (HAN) model on forecasting the length-of-stay (LoS) and in-hospital mortality on admissions by using ICD codes and demographic data. However, clinical notes are noisy documents that contain abbreviations and other tokens that are not relevant to the medical domain, hence using direct features without the guidance of medical domain knowledge will lead to performance deterioration. Viscosi et al. [37] analyzed the histopathology reports of patients to optimize the working time spent by the Catania-Messina-Enna Integrated Cancer Registry (RTI) operators in guiding with useful information to patients using machine learning techniques. Lv et al. [38] researched on the risk factors for stroke recurrence and developed an interpretable machine learning model to predict 30-day readmissions for stroke. Liu et al. [39] has proposed a domain knowledge empowered model that uses the ECG (electrocardiogram) and Echo notes to predict the in-hospital mortality. However, patient mortality cannot be predicted by selective vital measurements, rather, a holistic approach that considers the patient journey is required for accurate prediction. Though the authors built a knowledge-based hybrid attention model as we do, the dual knowledge feed between the texts and the extracted concepts increases the number of parameters to be learned, hence can lead to performance deterioration. Moreover, while the study predicts the mortality in the same admission as a binary classification task, we introduce the prediction of future ICD diagnoses which is a complex and critical task of EHR modeling that spans two different admissions. Furthermore, the dataset used in the study contains only a handful of patients who died; hence the selected notes do not contain enough information for prediction. Fei et al. [40] proposed a Multi-Filter Residual Convolutional Neural Network (MultiResCNN) for ICD coding. The study concluded that the multi-filter convolution as well as the residual convolution boosted the performance with moderate computation cost. Liu et al. [41] developed a CNN model that tries to forecast the onset of heart disease by predicting the general readmission and 30-day readmission for heart failure using clinical notes. We use this approach as a

TABLE 1. Top 20 most occurring diagnosis codes that are used in the top 20 prediction case scenario.

S/No	Clinical Name	ICD9	%
1	Essential Hypertension	401.9	3.5
2	Congestive heart failure	428.0	2.6
3	Atrial Fibrillation	427.31	2.5
4	Coronary atherosclerosis	414.01	2.5
5	Kidney Failure, Acute	584.9	1.7
6	Diabetes mellitus	250.00	1.6
7	Hyperlipidemia	272.4	1.5
8	Respiratory Failure	518.81	1.3
9	Urinary tract infection	599.0	1.2
10	Gastroesophageal reflux disease	530.81	1.1
11	Hyperlipidemia, group A	272.0	1.0
12	immunization for viral hepatitis	V05.3	0.9
13	suspected infectious condition	V29.0	0.9
14	Anemia	285.9	0.8
15	Hypothyroidism	244.9	0.8
16	Pneumonia	486	0.8
17	Posthaemorrhagic anaemia	285.1	0.7
18	Acidosis	276.2	0.7
19	Chronic airway obstruction	496	0.7
20	Severe Sepsis	995.92	0.6
Total (%)			**29.5%**

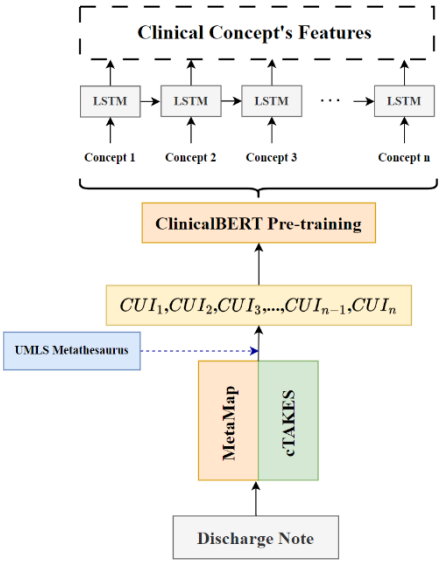

FIGURE 2. Extracting discharge summaries textual features using LSTM. The note's clinical concepts are extracted by combining Metamap and cTAKES to improve the annotation.

benchmark to assess how our model performs in the forecasting of the heart failure onset. Mallya et al. [42] used LSTMs to predict the congestive heart failure within 15 months in advance through a 12-month observation window using a large cohort of 216,394 patients and the model achieved the state-of-the-art AUROC results. Liang et al. [43] proposed a hybrid model that combines the LSTM and autoencoder to predict kidney disease in hypertension patients. The model uses a diagnosis text for prediction resulting in accurate prediction. Rajkomar et al. [44] built a general purpose and scalable approach that used the entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format to predict the mortality, risk of readmission, and the current admission's diagnosis. Though the approach is scalable and can be used to predict any adverse event, it uses many data inputs in a wholistic approach, hence it can be hard to train. Moreover, the approach predicts the risks that are only likely in the current admission. Hence, the effective use of the discharge notes which contain a critical summary of the patient's conditions provides a unique opportunity to predict the patient's future readmission and the associated diagnosis.

B. BACKGROUND KNOWLEDGE

While the breadth of the concepts used in this study will be subsequently discussed, in this section we give a background knowledge and preliminaries to the key building blocks.

1) LSTM

The standard LSTM is an improved variant of the basic recurrent neural network (RNN) that seeks to address the vulnerabilities of long-term information preservation as well as the short-term input skipping in latent variable models that the RNN suffered from. In sequence data modeling, the superiority of the LSTMs is achieved through a gating mechanism that helps in deciding on key past information to keep. A forget gate f_t monitors the amount of information from previous state c_{t-1} to be preserved. An input gate i_t determines how much the memory cell is altered by the new input and the output gate o_t determines how much memory data is supplied as a hidden state to the output. The LSTM processes are summarized in the following equations:

$$i_t = \sigma\left(W_{xi}CUI_t + W_{hi}h_{t-1} + b_i\right) \quad (1)$$

$$f_t = \sigma\left(W_{xf}CUI_t + W_{hf}h_{t-1} + b_f\right) \quad (2)$$

$$o_t = \sigma\left(W_{xo}CUI_t + W_{ho}h_{t-1} + b_o\right) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t \tanh\left(W_{xc}CUI_t + W_{hc}h_{t-1} + b_c\right) \quad (4)$$

$$h_t = o_t \tanh(c_t) \quad (5)$$

2) 1D CNN

CNNs are originally made for image data processing. However, Yoon et al. [45] proposed that textCNN can be used for text based applications. 1D CNNs do not require labor intensive feature engineering by domain experts, hence they have recently been applied in recent studies and in some applications including clinical narratives analysis they have demonstrated better performances than vanilla RNN sequences modeling [46], [47], [48]. With little specific tuning, text can be processed as a one-dimensional image, hence using a one-dimensional CNN we can capture the relationship between two adjacent words. As it is the practice in the 2D convolution for image, an input word is horizontally scanned by a kernel which is convolved to the input from the start of the word up to the end. The resulting output is a cross-correlation between the input word with the kernel and depending on the width of the kernel size we can extract the uni-gram (width of 1), bi-gram (width of 2) and tri-gram (width of 3) features. To produce a feature map per each kernel after several convolutions, we perform a 1-d max-over-time pooling operation [49] along the feature map by selecting the maximum value as the prominent feature. After the max-over-time pooling for each output channel we concatenate the pooling outputs of the n channels into a n-dimensional vector. At this stage, either we can extract the text features for other purposes, or we can add a fully connected layer for prediction. The specific structure of our 1D CNN implementation for the clinical note's features extraction is depicted in Figure. 3.

3) ATTENTION MECHANISM

The attention mechanism [50] is a process that mimics the human vision process where depending on the task at hand, the attention helps to focus on the most important regions of the viewed items and change the focal point appropriately instead of scanning the entire scene. This works also for sequence data where to understand a piece of reading, more weights are given to the tokens that are more influential to the overall meaning (classification) of the text. In machine learning, the attention process is implemented by estimating

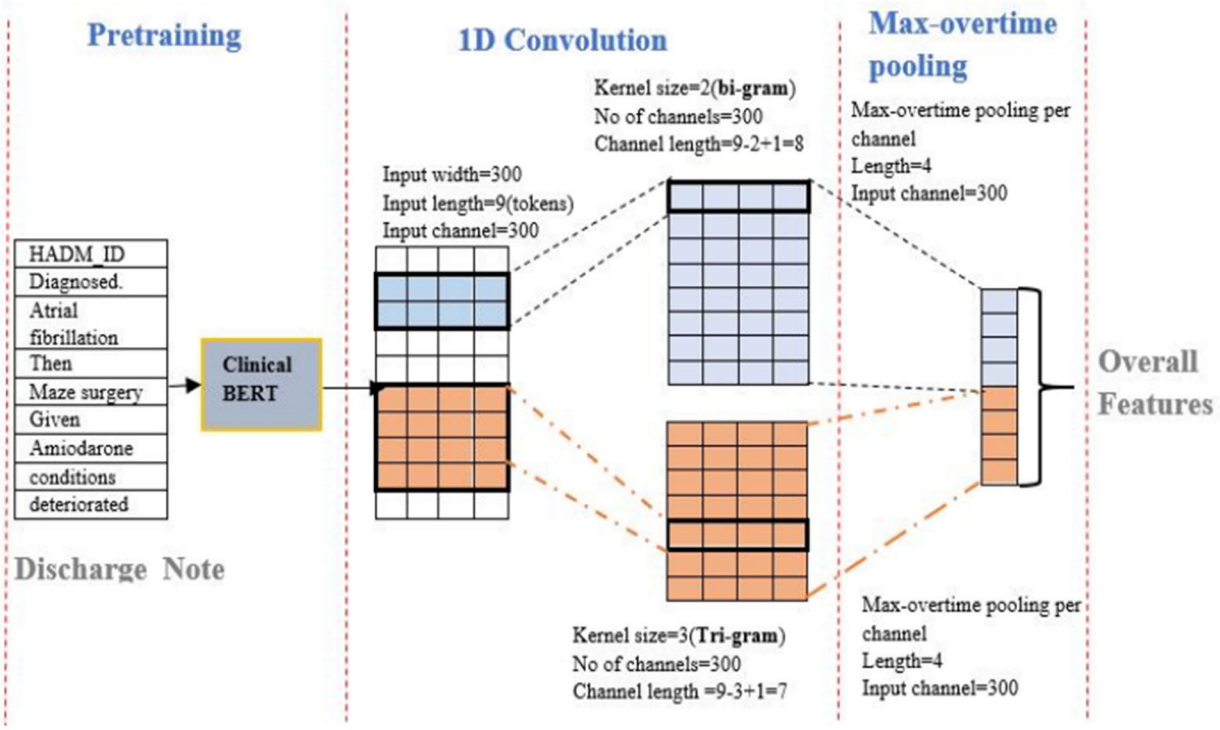

FIGURE 3. Extracting the note's features through convolutions followed by the max over-time pooling.

the weights of the input words (or pixels) by how strongly the word is correlated with the final output. A final text representation (context vector) is obtained by a weighted sum of all word features multiplied with their respective weights. The weights are computed by a correlation function like the dot product or by training a shallow neural network that can generate each word's weight.

4) UMLS AND CUI

Unified Medical Language System (UMLS) [51], is a centralized metatarsus that is comprised of medical standard ontologies. It contains the web, desktop and API based interfaces that serve to translate from one form to the other. One key feature of the UMLS that is of interest to our study is the clinical document annotation, a process that involves the discovery of meaningful clinical items called the Concept Unique Identifier (CUI) from clinical narratives. This process uses various annotators and the most popular are Metamap [52] and cTAKES [53]. The effective use of these annotators provides a more fine-grained and understandable representation of medical concepts which provides a higher abstraction of clinical texts than raw texts. Hence the content of the note's CUIs extracted and normal text filtered out can act as a knowledge base for many benchmark applications.

III. DATA DESCRIPTION AND PREPROCESSING

A. DATA SUMMARY

The MIMIC-III database that contains de-identified clinical records for 53,000 hospital admissions by patients that were admitted at the Intensive Care Unit of the Beth Israel Deaconess Medical Center from 2001 to 2012. It contains various clinical narratives like ECG, Echo, and Discharge summaries. We retained only the discharge summaries because they contain patient's outcome explicitly and contain a summary of what happened to the patient from the admission and future projections. MIMIC-III recorded a total of 59,652 discharge summaries. The discharge summary of the current admission serves as input to our model while the diagnoses of the next admission's ICD records will serve as labels.

B. DATA PREPROCESSING

The clinical notes are sparse documents with many unnecessary tokens and the average size of the discharge summary is 2600 words. Another aspect of clinical notes is that they cover the whole hospitalization of a patient hence the relationship between tokens is not as pronounced as in the normal documents. Hence, we only retained patients who are frequent fliers with at least two recorded admissions because our prediction spans two consecutive admissions. For each patient, we extracted information of the next admission and only a single note for each admission. We removed the newborns as their next admissions may be unpredictable. We also removed admissions which led to in-hospital mortality. We retained only the patients aged between 10 and 60 years old; hence, excluding the young and the elderly, because their re-admission can be due to age and other demographic conditions than on physiologic conditions. To keep

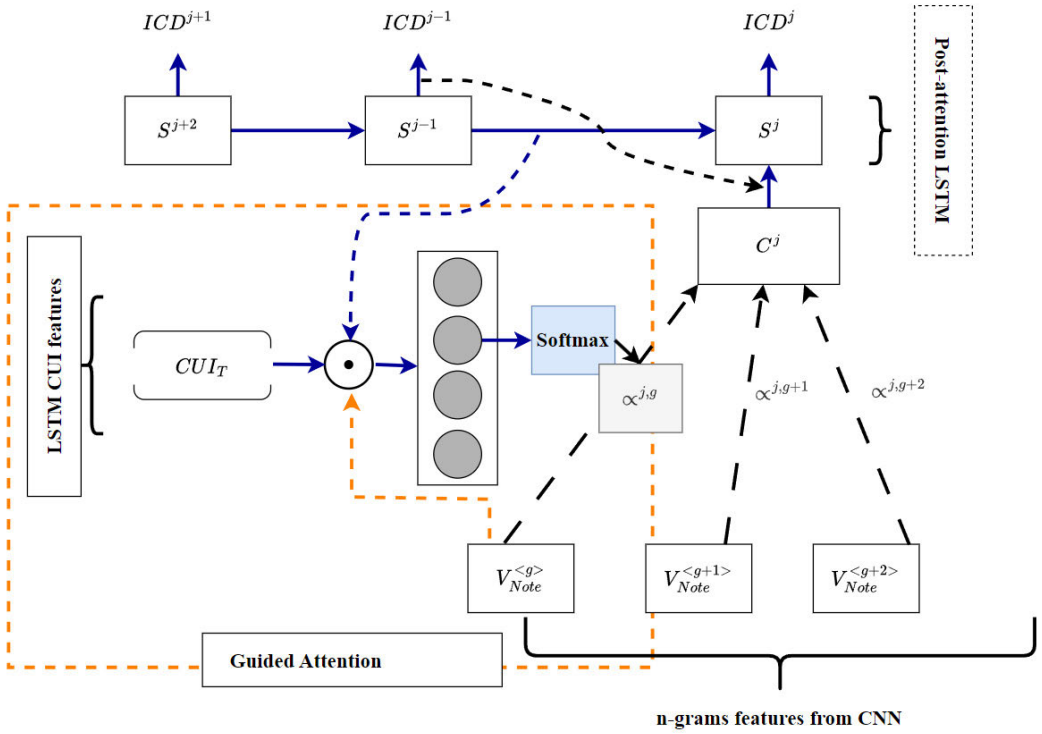

FIGURE 4. The attention given to a group of tokens(n-grams) for the prediction depends on previous n-grams, the current input and guided by the CUI features extracted by the LSTM. The regions(n-grams) of the notes that are highly correlated with the extracted tokens are considered for the prediction hence decreasing the number of parameters to be learned.

the physiological relationship between contiguous admissions, we only retained the discharge notes for admissions that had a readmission within 90 days from the current discharge. The resulting cohort consisted of 14,507 patients with one discharge note per patient. To reduce out-of-vocabulary (OOV) in the embedding stage all words in the note's words were put in lower case, special characters were stripped, words were lemmatized, known contractions and shortenings were removed, and the notes were tokenized.

C. CUIs EXTRACTION

For clinical concept extraction we use MetaMap and cTAKES, the tools that combine natural language processing and medical literature to extract and map biomedical text to concepts in the UMLS SNOMED-CT Metathausurus ontology. For improved annotation, authors in [54] have proposed an ensemble method that combines them. The authors computed the precision, recall and F1-scores and the combination proved more effective than when the two components are used alone. After getting the CUIs, we apply the UMLS Metathesaurus API to transform the CUI in structured medical terms.

D. PRE-TRAINING

The pre-training stage involves getting the distributed vector representation of words and concepts and we perform two embedding processes. The first involves the extracted CUIs which use ClinicalBERT pre-trained vectors [55]. ClinicalBERT is a variant of the BERT model [56] that uses the transformer architecture [50] to learn the embeddings for textual data. The second pre-training is the embedding of all clinical note's words also using ClinicalBERT. The advantages of using this pretrained embedding is its capability of capturing medical terms that are made of more than one token and keeping the relationships between comorbidities. This embedding scheme can capture complex clinical terms such as congestive heart failure as one vector than expressing it as three separate vectors like it is in the case for other popular word embeddings. It also captures the latent relationship between these complex clinical terms. For instance, the embedding vectors are such that the cosine similarity between the vectors of myocardial infarction and congestive heart failure is high.

E. LABELS PROCESSING AND PREDICTION SCENARIOS

In this study the prediction is performed for the following scenarios:

- **Top 20**: Table. 1 summarizes the top 20 diagnosis that cause the ICU admissions where they account for 30% of all cases. Hence in this binary classification scenario the output layer of the prediction model outputs the occurrence probability of any of these diseases within the next admission. This prediction is effective for the most occurring ICD codes but ineffective for rare diseases.

- **Single Level CCS (Diagnosis category)**: The ICD codes are grouped into 285 single level CCS categories, most of which are clinically homogeneous. For instance, all codes related to hypertension are grouped in one category (CCS code 99) rather than individual ICD codes. Hence, the output of the classification model will predict the occurrence of each of these grouped conditions in a multi-label classification approach.
- **ICD9 codes**: This is the most accurate diagnosis prediction. We predict each individual code with the guidance of extracted concepts. The benefit of the current model is that only the regions of the note that are highly correlated with the code are considered by the attention mechanism.

IV. METHOD

In this section, we describe the main HSGA process. The model has four main components; the LSTM based concept's features extraction, the whole note's features extraction using 1D CNN, the guided attention process, and the loss modeling with the final LSTM based prediction.

A. CONCEPT FEATURES EXTRACTION WITH LSTM

The CUI concept features extraction process is described in figure. 2. Each embedding vector representing a clinical concept extracted from the note is fed to an LSTM layer to get the representative features. Let $C = \{C_1, C_2, C_{i-1}, C_i, C_{i+1} \ldots C_n\}$ represents a list of meaningful concepts extracted from a patient's discharge summary with C_i a unique CUI represented by a dense vector obtained from the embedding process. At each step the LSTM cell described earlier will accept one concept vector C_i, performs an update of the neural memory cell C_i, then returns a hidden state h_t. At each time t the note is expressed by CUI_t and the textual features of the note are obtained, by extracting the last hidden layer of the LSTM.

B. OVERALL CLINICAL TEXT FEATURES FROM A CNN MODEL

The second step involves the extraction of the note's features using a 1D CNN as depicted in figure. 3. Let the clinical note be represented by an n dimension embedding and be composed of k embedding vectors. The note can be represented as a concatenation of the tokens embedding vectors as follow:

$$X_{1:k} = [x_1||x_2||x_3 \ldots ||x_k] \tag{6}$$

To extract the note's overall features from a convolution operation we apply a filter m with dimensions $m \in R^{mxn}$ where m represents a group of tokens and $0 \leq w < k$. Our implementation uses 1,2,3 filter sizes to represent unigram, bigram, and trigram respectively. We apply convolution on embedding vectors solely for feature extraction. For each filter of size mxn the output of the i-th convolution operation applied to window starting at i and ends at $i + m - 1$ is given by:

$$C_{m,i} = \tanh\left(W_m x_{i:i+m-1} + b_m\right) \tag{7}$$

W_m is the convolution weight and b_c is the bias term. For each layer l, we perform the convolution over the note's regions with the height and width of the resulting $C^{l}_{m,i}$ given by:

$$\begin{aligned} S^{[l]}_{m,i} &= \frac{S^{l-1}_h + 2p^{[l]} - m^{[l]} + 1}{d^{[l]}} \\ S^{[l]}_{m,i} &= \frac{S^{l-1}_w + 2p^{[l]} - m^{[l]} + 1}{d^{[l]}} \end{aligned} \tag{8}$$

where p and d are the padding and stride at layer l respectively. To extract quality features, we use a stride of 1 and 0 padding. Hence for filter with convolution size m the convolution operation produces the intermediate features map given by:

$$C_m = [C_{m,1}; C_{m,2}; \ldots; C_{m,k-m+1}] \tag{9}$$

The convolution operation is followed by a sub-sampling operation using max-over-time pooling. This involves the extraction of the maximum value for the feature map generated by each filter.

$$C^{final}_m = max\left(C_m\right) \tag{10}$$

The 1D max overtime pooling is applied because unlike the image inputs that have the same dimensions, the text in a clinical note contains varying dimensions. Hence this special type of pooling operation tries to capture the most important features for each feature map. Finally, the overall feature map is obtained by grouping similar feature maps as per the convolution size (i.e., uni-gram, bi-gram, and tri-gram) and is given by:

$$C = [C^{final}_1, C^{final}_2, C^{final}_3] \tag{11}$$

C. SELF-GUIDED ATTENTION MODEL

The main component of the current model is the guided attention model depicted in figure. 4. The objective of the attention model is to use the features of extracted CUI obtained from the LSTM component to guide the network in deciding which $n-$ grams from the note are the most influential in the overall prediction. In this step we use the spatial attention which depends on interaction between the CUI features and the note's CNN features. Hence the process attends to the note's features V_{Note} based on CUI features V_{CUI}. For every ICD-9 code ICD_j that we aim to predict we want to define a hidden state S_j obtained from an intermediate context vector C^j which is given by:

$$C^{<j>} = \sum_{g=1}^{n} \alpha^{<j,g>} V^{<g>}_{Note} \tag{12}$$

where n is the total n-grams contained in the note, $\alpha^{<j,g>}$ the amount of attention weight given to the g^{th} n-gram to generate the context $C^{<j>}$ and $V^{<g>}_{Note}$ represent the features of the g^{th} n-gram obtained from the CNN features extractor. Owing to the sum of weights given to individual n-grams should always

be 1, $\alpha^{<j,g>}$ is given by a SoftMax probability function as follows:

$$\alpha^{<j,g>} = \frac{exp b^{<jg>}}{\sum_{g}^{T} exp b^{<jg>}} \quad (13)$$

As seen in Figure. 4, to obtain $b^{<jg>}$ we train a single-layer neural network that learns its weights by combining the following three components:

- the previous hidden state S j−1 obtained from the previous n-gram,
- $V_{Note}^{<g>}$ Note the g^{th} n-gram's features from CNN,
- CUI_T with CUI $\in R^d$ d the attention guidance vector from CUI vectors from the LSTM, with d the total number of CUI vectors.

hence the neural network uses the SoftMax function and gradient descent to learn the attention that should be given to input $V_{Note}^{<g>}$ as follows:

$$b^{<jg>} = \tanh\left(W_C CUI_T \odot W_N V_{Note}^{<g>}\right) \odot W_s S_{j-1} \quad (14)$$

The parameters W_C, W_N and W_S which are learned in the combination process cause the final value of $\alpha^{<j,g>}$ of a given input to depend on its correlation with tokens from extracted concepts.

D. LOSS MODELING

The final LSTM layer depends on one of the predictive tasks described above. We use binary classification for Top 20 and the ICD9 codes prediction scenarios with binary cross-entropy (BCE) loss function as follow:

$$LOSS_{BCE} = -\left(y.\log\left(\tilde{y}\right) + (1-y).\log\left(1-\tilde{y}\right)\right) \quad (15)$$

where y is a set of true labels and $\tilde{y}$ are predicted labels. In single level CCS, the attention model scans through the note and outputs the occurrence probability of each one of the 285 single level CCS categories. This means that as far as one admission is concerned these classes are not mutually exclusive. Hence, in this multi-label prediction we get the sum of binary cross entropies of each of the 285 labels.

$$LOSS_{Multi-label} = \frac{1}{N}\sum_{n=1}^{N} -(y_n.\log\left(\tilde{y}\right) + (1-y_n).\log(1-\widetilde{y_n})) \quad (16)$$

where N represents the total number of labels (285 CCS categories), y is a set of true labels and $\tilde{y}$ is the predicted labels.

V. EXPERIMENTS

Each note maximum size is fixed at 5000 words, the pre-trained word-embedding vectors were fixed at 300 features per token and the number of unique words to use (i.e., number of rows in embedding vector) from the pre-trained embedding was fixed at 120,000. During the training the binary classification loss for top 20 is modeled using BCEWithLogitsLoss, a Binary Cross Entropy (BCE) loss [56] that includes a sigmoid activation. The models are optimized using Adam optimizer [57] with a learning rate of 0.0001 with dropout. The training is performed for 10 epochs with batch size of 512 and uses used the k-fold cross validation with 5-fold splits. The k-fold cross validation [58] performs better than the traditional train/test splits. All training processes were implemented in PyTorch [59] empowered with Cuda GPU [60] and the optimal results were obtained after 5 hours of training.

A. BASELINES

We compare our approach with the available methods (Though various prediction, implementations and dataset may differ from the original work due to several limitations). For each prediction scenario described earlier we implement the following:

- **Notes + CNN**: This a two layers 1D CNN. In this baseline the CNN is applied directly by treating the discharge note as a normal text and used for prediction.
- **Notes + LSTM**: This model was implemented to assess the superiority of 1D CNN over LSTMs in clinical documents analysis.
- **Notes + CNN + attention**: This approach was used in CAML [34]. A single convolutional layer and attention layer was used to generate the label-aware features for multi-label classification.
- **Notes + ICD codes + Multi-modal Attention Neural Networks (MNN)**: This baseline follows the approach used in [21] which tries to predict diagnosis codes by combining the multi-modal data from clinical notes and ICD codes.
- **RNN + attention**: This method is used in DIPOLE [25] and tries to model the patient's records using a Bidirectional RNN and attention mechanism.

The objective is not to re-produce these works but to investigate the effects of the main building blocks and effects of the hybrid self-guided attention. For instance, in CAML they perform a per ICD label prediction, hence to compare with HSGA we implement CAML but the implementation details including the dataset, data processing, the embedding process and the prediction layer follow the current implementations hence the reported results might diverge with the referenced works.

VI. RESULTS AND DISCUSSIONS

A. QUANTITATIVE EVALUATION METRICS

The performance of the model against all baselines for all prediction scenarios is reported in table. 2. The following are the metrics used to assess the superiority of our method for each prediction scenario:

For the Top 20 scenario, we implement each of the baseline's methods for this prediction and use the Area under Receiver Operator Curve (AUROC) as well as Area Under Precision-Recall Curve (AUPRC). Though the ROC curves are typically utilized for binary classification, we extend it for

TABLE 2. Performance comparison of HSGA against baselines for all prediction scenarios.

Pred.Scenario	Metric	CNN	LSTM	CNN+att. [34]	MNN att. [21]	RNN+att. [25]	Our Scheme
Top 20	AUROC	0.69	0.68	0.76	0.79	0.74	**0.87**
	AUPRC	0.59	0.51	0.73	**0.77**	0.68	0.75
SINGLE LVL CCS	Macro Avg. ROC	0.68	0.64	0.75	0.77	0.72	**0.82**
	Micro Avg. ROC	0.62	0.58	0.74	0.73	0.70	**0.79**
ICD9 CODES	Macro Avg. Recall	0.58	0.52	0.72	0.78	0.54	**0.80**
	Micro Avg. Recall	0.56	0.52	0.74	0.76	0.53	**0.84**

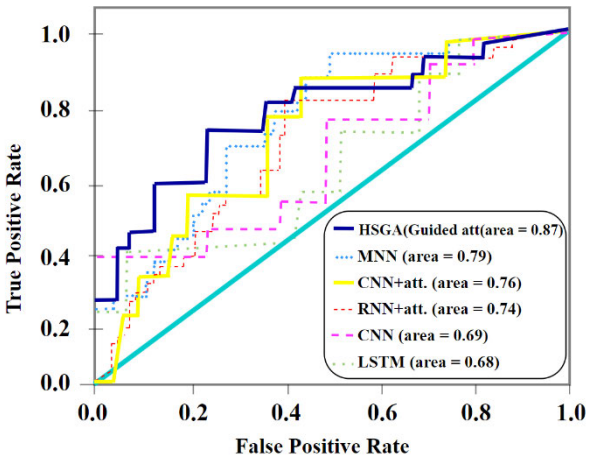

FIGURE 5. The AUROC for all baselines with our scheme for the Top 20 prediction scenario.

the single level CCS which is a multi-label scenario. Hence, we use the macro-averaged Area Under the ROC Curve (AUROC), which finds the average of per-label AUROC as well as the micro-averaged AUROC, which calculates the single AUROC score for all classes together. For ICD9 codes, we evaluate the performance of our model by also using micro-average and macro-average values obtained by averaging recall values which are calculated by considering each note and an individual output code as follows:

$$Macro - Recall = \frac{1}{n}\sum_{g=1}^{n}\frac{TP_g}{TP_g + TN_g} \tag{17}$$

$$Macro - Recall = \frac{\sum_{g=1}^{n} TP_g}{\sum_{g=1}^{n} TP_g + TN_g} \tag{18}$$

B. RESULTS

Figure. 5 shows the performance of each baseline and our method for the Top 20 prediction. A general intuition is that the embeddings used in HSGA model accounted for only 5% of out-of-order words. In MNN, the use of the ICD9 knowledge base boosted performance considerably. CAML exhibits an average performance compared to HSGA and MNN because the baseline uses a flat and fixed-length convolutional architecture hence making it difficult to learn decent document representations. From figure. 6, with the single level CCS scenario, we observe that using the notes with flat CNN results in poor AUROC scores. The reason is that various tokens from the notes are meaningless hence the CNN model cannot establish refined text representation. In table. 2, we perform a full comparison of all baselines for all the prediction scenarios. From the table, we observe that HSGA model outperforms other models in most prediction scenarios. For top 20 scenario, the approach used in MNN [21] performs better than all other baselines approach with a higher AUROC and AUPRC. The overall performance gets boosted by the adoption of hybrid approaches that combine CNN with LSTM as well as the attention mechanism. For the single level CCS prediction scenario, the improvement in micro and macro average ROC values obtained by of our approach compared to the best performing baseline (MNN) is above 5% proving that a hybrid approach that uses domain knowledge improves the performance considerably. The only difference between our approach and CAML (CNN Attention) is the guidance process. However, HSGA performance is superior at about 10% in micro and macro recall values. The reason is that rather than scanning the regions of the note to focus on key n-grams that are influential to the prediction of each ICD code, our approach considers only n-grams that are related with CUIs extracted from the notes.

C. OBSERVING THE EFFECT OF THE HYBRID ATTENTION

Observing closely in figure. 4 we can assess the effect of the guided attention over a CNN-only based attention approach. From such figure, the CNN + Attention approach used in [34] achieves an AUROC of 0.76 while the current model achieves 0.87 for the HSGA in Top 20 binary prediction scenario. Also, for the multi-labels' prediction scenarios CNN + Attention achieves a micro-average ROC of 0.79 while the proposed model achieves 0.82 for the single level CCS prediction. A similar pattern can be observed for the ICD9 codes prediction scenario with a micro-recall of 0.74 against 0.84. The only design difference between the two models and the reason for this evident improvement is the LSTM based knowledge base that guides the attention process.

D. CASE STUDY: THE PREDICTION OF HEART DISEASE ONSET

The prediction of heart failure onset has recently been used as a case study in several healthcare predictive modeling [18], [21]. The reason is that this disease leaves many

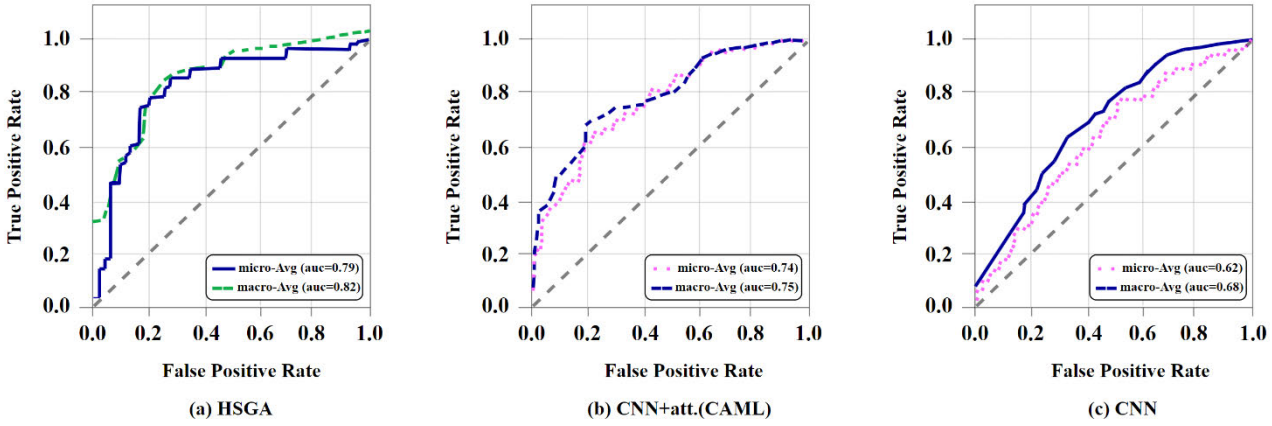

FIGURE 6. Macro/Micro AUROC for the current hybrid approach with the CNN based baselines for the Single Level CCS (diagnosis category) prediction scenario.

TABLE 3. Prediction of heart failure. the attention mechanism guided by the extracted concepts attend only on 3-grams that are near the related token. The highlighted ICD codes are correctly predicted by the model.

Concepts extracted by cTAKES & Metamap	UMLS description	key 3-grams with higher attention weights	HSGA ICD Codes Prediction
C2707305	Shortness of breath	... the 45 yr old male was admitted **shortness of breath** dry cough and low blood pressure. **insulin** was increased your nph does to 22u in the morning and 12u at night...we added the following new medications : **lisinopril** , for protection against **heart failure** , and **carvedilol** and **spironolactone** to protect his heart. we gave him **gabapentin** and morphine on an as needed basis for your pain.followup instructions : tuesday at 10 : 30am . you have an appointment for **Pulmonary function tests.**	***428.0**(heart failure) 996.72 * **V58.67**(Encounter for long-term (current) use of insulin) 447.1(Stricture of artery) ***414.01**(Coronary ath erosclerosis of native c oronary artery
C0375729	Other postoperative infection		
C0260334	Pneumonitis due to food and vomit		
C0030591	Paroxysmal ventricular tachycardia		
C0002871	Anemia		
C0021641	insulin		
C0065374	klisinopril		
C0054836	carvedilol		
C2021523	PFT airway resistance		

footprints in the EHR before its manifestation hence it can serve as a very informative case study. To access the predictive capability of our model, we proceed like in [41] and selected the admissions whose ICD codes include 398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.0, 428.1, 428.20, 428.21, 428.22, 428.23, 428.30, 428.31, 428.32, 428.33, 428.40, 428.41, 428.42, 428.43, and 428.9. These admissions were qualified as heart failure related admissions. We then extracted the discharge notes of the previous admission and fed the notes to observe the prediction performance as well as the quality of the guided attention mechanism as well as the tokens that scored higher weights to the heart disease prediction. Table. 3 depicts a real case of a patient who died from a heart failure during the final hospital admission. The discharge summary of the penultimate admission is used to predict the onset of heart failure accurately. The table shows CUI that are extracted and among 2576 tokens on the note, few n-grams that are related to the extracted CUIs are able to score the highest attention weights hence considered for the last prediction. The attention mechanism is able to focus also on other tokens which are not medically meaningful but hold key information about the conditions of the patient. From the table among 5 codes present in the admission 3 out 5 codes were properly predicted using the ICD Codes prediction scenario. The table shows some of the concepts extracted from the note (in total 35 were extracted) and textual representation. These UMLS representations (ex: shortness of breath) are used to guide the attention model about the important tri-grams to consider for prediction.

VII. CONCLUSION

We propose and evaluate HSGA model, a hybrid CNN-LSTM self-guided attention model, to predict the diagnosis that is likely to occur in the next patient's admission using the current admission's discharge narratives. These notes are highly sparse and using them without the guidance of a domain knowledge results in bad results. We consider only the next admission that will happen in the window of 90 days from the current discharge. We use the approach for 3 prediction scenarios and experimental results on a real-world dataset proves that the proposed models outperform the state-of-the-art approaches, and it corroborates the benefit of the self-guided attention mechanism. From such experiment, the CNN + Attention achieves an AUROC of 0.76 while our proposed model, HSGA, achieves 0.87 for the HSGA in Top 20 binary prediction scenario. Also, for the multi-labels' prediction scenarios, CNN + Attention achieves

a micro-average ROC of 0.79, while HSGA achieves 0.82 for the single level CCS prediction. A similar pattern can be observed for the ICD9 codes prediction scenario with a micro-recall of 0.74 against 0.84. Our attention mechanism generates more clear attention weights due to the inclusion of self-guidance. The embedding steps uses an ensemble approach making our model able to capture clinical words and non-clinical tokens that are necessary to describe the patient's progression.

SUMMARY TABLE

What was already known on the topic:

1. Deep learning can be used to predict the patient's readmission risks using information available on discharge.

2. Predicting the diagnosis to be expected in an eventual re-admission can be performed using the current ICD diagnosis and other structured ontologies.

What this study added to our knowledge:

1. Even with the attention mechanism, using the clinical notes like discharge summaries to predict the ICD codes without first building a domain knowledge can results in worse predictions as the notes are noisy and sparse

2. The prediction is boosted by the combination of deep learning models with one model helping to build a knowledge base and another model performing the prediction. This is because the attention mechanism uses the knowledge base to focus on parts of the notes that are the most influential for prediction and avoid the much noisy unnecessary tokens and abbreviations recorded in clinical narratives.

REFERENCES

[1] A. J. Forster, H. J. Murff, J. F. Peterson, T. K. Gandhi, and D. W. Bate, "The incidence and severity of adverse events affecting patients after discharge from the hospital," *Ann. Internal Med.*, vol. 138, pp. 161–167, Feb. 2003.

[2] Z. T. Korach, J. Yang, S. C. Rossetti, K. D. Cato, M.-J. Kang, C. Knaplund, K. O. Schnock, J. P. Garcia, H. Jia, J. M. Schwartz, and L. Zhou, "Mining clinical phrases from nursing notes to discover risk factors of patient deterioration," *Int. J. Med. Informat.*, vol. 135, Mar. 2020, Art. no. 104053.

[3] L. Zhou, T. Siddiqui, S. L. Seliger, J. B. Blumenthal, Y. Kang, R. Doerfler, and J. C. Fink, "Text preprocessing for improving hypoglycemia detection from clinical notes—A case study of patients with diabetes," *Int. J. Med. Informat.*, vol. 129, pp. 374–380, Sep. 2019.

[4] Y. Yu, M. Li, L. Liu, Z. Fei, F.-X. Wu, and J. Wang, "Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN," *J. Biomed. Informat.*, vol. 91, Mar. 2019, Art. no. 103114.

[5] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes," *Comput. Methods Programs Biomed.*, vol. 177, pp. 141–153, Aug. 2019.

[6] T. Gangavarapu, A. Jayasimha, and G. S. Krishnan, "Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes," *Knowl.-Based Syst.*, vol. 190, Feb. 2020, Art. no. 105321.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[8] W. Yin, S. Ebert, and H. Schütze, "Attention-based convolutional neural network for machine comprehension," 2016, *arXiv:1602.04341*.

[9] J. Guo, G. Liu, and C. Xiong, "Multiple attention networks with temporal convolution for machine reading comprehension," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 546–549.

[10] P. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6632–6641.

[11] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 30–38.

[12] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019.

[13] J. Li, Y. Xu, and H. Shi, "Bidirectional LSTM with hierarchical attention for text classification," in *Proc. IEEE 4th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Dec. 2019, pp. 456–459.

[14] R. Yu, Y. Zheng, R. Zhang, Y. Jiang, and C. C. Y. Poon, "Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 486–492, Feb. 2020.

[15] N. Xu, Y. Shen, and Y. Zhu, "Attention-based hierarchical recurrent neural network for phenotype classification," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2019, pp. 465–476.

[16] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. V. Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Sci. Data*, vol. 6, no. 1, pp. 1–12, Jun. 2019.

[17] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, and L. Jorm, "Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk," *Sci. Rep.*, vol. 10, no. 1, pp. 1–20, Jan. 2020.

[18] Y. Zhang, X. Yang, J. Ivy, and M. Chi, "ATTAIN: Attention-based time-aware LSTM networks for disease progression modeling," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1–15.

[19] J. Gao, X. Wang, Y. Wang, Z. Yang, J. Gao, J. Wang, W. Tang, and X. Xie, "CAMP: Co-attention memory networks for diagnosis prediction in healthcare," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 1036–1041.

[20] H. Song, D. Rajan, J. Thiagarajan, and A. Spania, "Attend and diagnose: Clinical time series analysis using attention models," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1–8.

[21] Z. Qiao, X. Wu, S. Ge, and W. Fan, "MNN: Multimodal attentional neural networks for diagnosis prediction," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5937–5943.

[22] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella, "The 'K' in K-fold cross validation," in *Proc. Eur. Symp. Artif. Neural Netw.*, 2012, pp. 441-446.

[23] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, May 2016.

[24] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," 2015, *arXiv:1511.03677*.

[25] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1903–1911.

[26] B. Jang, Y. Kim, G. I. Kim, and J. W. Kim, "Deep similarity analysis and forecasting of actual outbreak of major infectious diseases using internet-sourced data," *J. Biomed. Informat.*, vol. 133, Sep. 2022, Art. no. 104148.

[27] B. Jang, I. Kim, and J. W. Kim, "Long-term influenza outbreak forecast using time-precedence correlation of web data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 5, pp. 2400–2412, May 2023.

[28] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. NIPS*, 2016, pp. 1–13.

[29] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 743–752.

[30] M. Tang, P. Gandhi, M. A. Kabir, C. Zou, J. Blakey, and X. Luo, "Progress notes classification and keyword extraction using attention-based deep learning models with BERT," 2019, *arXiv:1910.05786*.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[32] Y. Kim, J.-H. Kim, Y.-M. Kim, S. Song, and H. J. Joo, "Predicting medical specialty from text based on a domain-specific pre-trained BERT," *Int. J. Med. Informat.*, vol. 170, Feb. 2023, Art. no. 104956.
[33] Y. Peng, K. Yan, V. Sandfort, R. M. Summers, and Z. Lu, "A self-attention based deep learning method for lesion attribute detection from CT reports," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2019, pp. 1–5.
[34] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2018, pp. 1–11.
[35] O. Trigueros, A. Blanco, N. Lebeña, A. Casillas, and A. Pérez, "Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention," *Int. J. Med. Informat.*, vol. 157, Jan. 2022, Art. no. 104615.
[36] G. Harerimana, J. W. Kim, and B. Jang, "A deep attention model to forecast the length of stay and the in-hospital mortality right on admission from ICD codes and demographic data," *J. Biomed. Informat.*, vol. 118, Jun. 2021, Art. no. 103778.
[37] C. Viscosi, P. Fidelbo, A. Benedetto, M. Varvarà, and M. Ferrante, "Selection of diagnosis with oncologic relevance information from histopathology free text reports: A machine learning approach," *Int. J. Med. Informat.*, vol. 160, Apr. 2022, Art. no. 104714.
[38] J. Lv, M. Zhang, Y. Fu, M. Chen, B. Chen, Z. Xu, X. Yan, S. Hu, and N. Zhao, "An interpretable machine learning approach for predicting 30-day readmission after stroke," *Int. J. Med. Informat.*, vol. 174, Jun. 2023, Art. no. 105050.
[39] N. Liu, P. Lu, W. Zhang, and J. Wang, "Knowledge-aware deep dual networks for text-based mortality prediction," in *Proc. IEEE 35th Int. Conf. Data Eng. (ICDE)*, Apr. 2019, pp. 1406–1417.
[40] F. Li and H. Yu, "ICD coding from clinical text using multi-filter residual convolutional neural network," in *Proc. AAAI Conf. Artif. Intell., AAAI Conf. Artif. Intell.*, 2019, pp. 8180–8187.
[41] X. Liu, Y. Chen, J. Bae, H. Li, J. Johnston, and T. Sanger, "Predicting heart failure readmission from clinical notes using deep learning," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 2642–2648.
[42] S. Mallya, M. Overhage, N. Srivastava, T. Arai, and C. Erdman, "Effectiveness of LSTMs in predicting congestive heart failure onset," 2019, *arXiv:1902.02443*.
[43] Y. Ren, H. Fei, X. Liang, D. Ji, and M. Cheng, "A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records," *BMC Med. Informat. Decis. Making*, vol. 19, no. S2, pp. 131–138, Apr. 2019.
[44] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, p. 18, 2018.
[45] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," 2015, *arXiv:1510.03820*.
[46] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," 2017, *arXiv:1702.01923*.
[47] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote, E. T. Moseley, D. W. Grant, P. D. Tyler, and L. A. Celi, "Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0192360.
[48] K. Lee, M. L. Famiglietti, A. Mcmahon, C.-H. Wei, J. A. L. MacArthur, S. Poux, L. Breuza, A. Bridge, F. Cunningham, I. Xenarios, and Z. Lu, "Scaling up data curation using deep learning: An application to literature triage in genomic variation resources," *PLOS Comput. Biol.*, vol. 14, no. 8, Aug. 2018, Art. no. e1006390.
[49] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," 2011, *arXiv:1103.0398*.
[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 1–15.
[51] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1–9.
[52] A. R. Aronson, "MetaMap: Mapping text to the UMLS metathesaurus," in *Proc. AMIA Symp.*, 2001, pp. 17–21.
[53] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 5, pp. 507–513, Sep. 2010.
[54] T.-T. Kuo, P. Rao, C. Maehara, S. Doan, J. D. Chaparro, M. E. Day, C. Farcas, L. Ohno-Machado, and C.-N. Hsu, "Ensembles of NLP tools for data element extraction from clinical notes," in *Proc. Annu. Symp.*, 2016, pp. 1880–1889.
[55] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," 2019, *arXiv:1904.05342*.
[56] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[58] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013, p. 18.
[59] N. Ketkar, "Introduction to PyTorch," in *Deep Learning With Python*. New York, NY, USA: Apress, 2017, pp. 195–208.
[60] P. Harish and P. J. Narayanan, "Accelerating large graph algorithms on the GPU using CUDA," in *Proc. Int. Conf. High Perform. Comput.*, 2007, pp. 197–208.

GASPARD HARERIMANA (Member, IEEE) received the B.S. degree in computer engineering from Ethiopian Defense University, in 2008, the M.S. degree in information technology from Carnegie Mellon University, in 2015, and the Ph.D. degree in computer science from Sangmyung University, Seoul, South Korea, in 2020. He was a Visiting Lecturer with the Adventist University of Central Africa (AUCA), Kigali, Rwanda. His research interests include big data and machine learning with emphasis on deep learning.

GUN IL KIM (Member, IEEE) received the B.S. degree in international studies and in business and economics and the M.S. degree in information systems from Yonsei University, Seoul, South Korea, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree in information systems. His research interests include NLP and deep learning.

JONG WOOK KIM (Member, IEEE) received the Ph.D. degree from the Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group, Teradata, from 2010 to 2013. He is currently an Assistant Professor in computer science with Sangmyung University. His primary research interests include data privacy, distributed databases, and query optimization. He is a member of the ACM.

BEAKCHEOL JANG (Member, IEEE) received the B.S. degree in computer science from Yonsei University, in 2001, the M.S. degree in computer science from the Korea Advanced Institute of Science and Technology, in 2002, and the Ph.D. degree in computer science from North Carolina State University, in 2009. He is currently an Associate Professor with the Graduate School of Information, Yonsei University. His primary research interests include artificial intelligence, big data analytics, and natural language processing. He is a member of the ACM.