SDCANet: Enhancing Symptoms-Driven Disease Prediction with CNN-Attention Networks

Thao Minh Nguyen Phan[1], Cong-Tinh Dao[1], Tai Tan Phan[2], and Hai Thanh Nguyen[2](✉)

[1] National Yang Ming Chiao Tung University, Hsinchu City, Taiwan
pnmthaoct@gmail.com
[2] Can Tho University, Can Tho City, Vietnam
ntthai.cit@ctu.edu.vn

Abstract. Deep learning algorithms have revolutionized healthcare by improving patient outcomes, enhancing diagnostic accuracy, and advancing medical knowledge. In this paper, we propose an approach for symptom-based disease prediction based on understanding the intricate connections between symptoms and diseases by accurately representing symptom sets, considering the varying importance of individual symptoms. This framework enables precise and reliable disease prediction, transforming healthcare diagnosis and improving patient care. By incorporating advanced techniques such as a one-dimensional convolutional neural network (1DCNN) and attention mechanisms, our model captures the unique characteristics of each patient, facilitating personalized and accurate predictions. Our model outperforms baseline methods through comprehensive evaluation, demonstrating its effectiveness in disease prediction.

Keywords: disease prediction · healthcare · symptom

1 Introduction

Artificial Intelligence (AI)-based tools can support physicians in the diagnostic process, increasing the chances of timely healthcare care for a broader population [13]. AI provided solutions to key challenges, including reducing diagnosis time for critical conditions, improving access to comprehensive care, and reducing healthcare expenses. In addition, AI took advantage of Computer-Aided Diagnosis (CAD) systems, primarily to address healthcare crises such as the COVID-19 pandemic. The implementation of AI-driven solutions has witnessed extensive adoption across various fields.

However, several factors contributed to the limitations in implementing AI solutions in healthcare. One major limitation was the stringent compliance regulations imposed by the Health Insurance Portability and Accountability Act (HIPAA), which restricts the public availability of healthcare data, even for

N. Thai-Nghe et al. (Eds.): ISDS 2023, CCIS 1949, pp. 15–30, 2024.
https://doi.org/10.1007/978-981-99-7649-2_2

research purposes [16]. As a result, the development of new AI-driven healthcare systems faced significant challenges. Moreover, the medical data tended to be imbalanced and skewed due to disparities in the number of participants between patient groups and healthy individuals in medical studies. As a result, the sample sizes for data collection were typically small, making rectifying the data imbalance issue complicated. Additionally, the need for more transparency in data collection posed a critical concern when developing AI-based healthcare systems. In a published report by the World Health Organization (WHO) [8], with the lack of physicians for patients in many developing countries, providing timely care and medical services to needy individuals has become an enormous challenge. The scarcity of healthcare professionals created a situation where it is nearly impossible for medical professionals to cater to the growing healthcare demands promptly. This disparity in the healthcare workforce posed significant barriers to accessing essential medical care. Despite these challenges, the benefits of an AI-led healthcare system outweighed the obstacles. Developing a robust and accurate predictive system that provides initial diagnoses and fulfills the minimum healthcare support required by individuals becomes crucial. Deep learning (DL) techniques demonstrated immense promise and potential in medicine, particularly in accurate prognosis and diagnosis [16]. DL algorithms transformed healthcare using large volumes of medical data and uncovering hidden patterns, leading to better patient outcomes, better diagnostic accuracy, and advancements in medical knowledge.

Symptoms were primary clinical phenotypes that are significant resources but often neglected. They played a pivotal role in clinical diagnosis and treatment, serving as essential indicators of an individual's health status. For instance, when considering a heart attack, key symptoms encompass diverse manifestations such as chest pain, discomfort radiating to the arms, shoulders, jaw, neck, or back, feelings of weakness, lightheadedness, fainting, and shortness of breath. This broad spectrum of symptoms vividly demonstrated the interconnectedness of homeostatic mechanisms, whose disruptions ultimately cause the emergence of a disease. Community health professionals and general practitioners' knowledge regarding symptoms of specific diseases was primarily derived from their observations within hospital environments [5]. It should be noted that symptoms, the most directly observable disease characteristics, served as the foundational elements for classifying clinical diseases.

Given the limitations of doctors in avoiding diagnostic errors due to restricted expertise and potential human negligence, a symptom-based prognosis model held significant potential in providing valuable assistance [17]. This model could assist physicians by providing a systematic approach to the prognosis based on a given set of symptoms, offering convenience and accessibility in the diagnostic process. However, developing and implementing this novel approach presented several unique challenges that must be addressed to ensure its effectiveness. The first challenge revolved around modeling the complex relationships between symptoms and diseases. The intricate interplay between symptoms and their corresponding diseases required a robust framework that accurately captured these relationships. This involved understanding the various factors that influence the manifestation of symptoms and the underlying mechanisms that link them to specific diseases. Developing a comprehensive model that effectively captures and represents these relationships could significantly improve prognostic accuracy, leading to better patient outcomes and healthcare decision-making. The second challenge focused on improving the accuracy of the representation of the symptom set. Accurately representing the collective information became crucial to a reliable prognosis when dealing with symptoms. It was essential to consider the varying importance and relevance of individual symptoms.

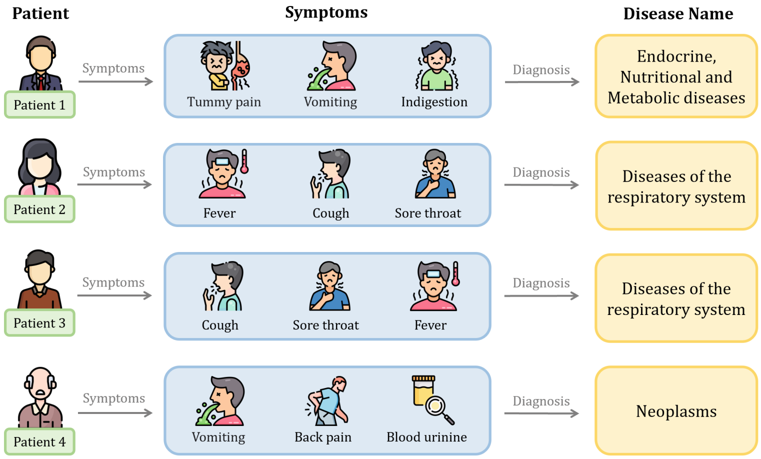

Fig. 1. An Illustrative Demonstration of Symptom-Based Disease Prediction

Addressing these challenges, DL methodologies, with their ability to uncover latent and hidden patterns within data, identified underlying connections between symptoms and diseases, thereby facilitating the development of an AI-based healthcare system [11]. The goal of this study was to predict diseases using their symptoms. The goal is to give doctors helpful resources to use while making diagnoses. We proposed a simplified approach that involves only presenting symptoms rather than relying on complex historical clinical records. Doctors may manually enter these symptoms, or computer programs may automatically extract them from medical records. The symptom-based disease prediction system ensures privacy and can be widely used to help doctors prescribe the proper medications using symptoms, which can provide information on a patient's physical condition while protecting their personal information. Figure 1 illustrates that patients 2 and 3 with identical symptoms, such as fever accompanied by cough and sore throat, are more likely to have the same disease diagnosed by medical professionals, specifically respiratory system diseases.

In this paper, we presented our proposed framework, known as **S**ymptoms-based **D**isease Prediction with **C**NN-**A**ttention **Net**works (SDCANet), with the primary goal of predicting diseases by leveraging both patient symptoms and demographics. Our framework addresses existing limitations in healthcare diagnosis, aiming to reduce diagnostic errors and enhance patient outcomes. Our model effectively analyzes and interprets the intricate relationships between symptoms and diseases by harnessing the power of 1D convolutional neural networks (1DCNN) and attention mechanisms. It overcomes the challenges posed by diverse symptom sets and considers the varying importance of individual symptoms, resulting in precise and reliable disease prediction. The incorporation of 1DCNN allows our model to capture sequential patterns and local dependencies within the symptom data, enabling a deeper understanding of disease dynamics. Simultaneously, the attention mechanisms focus on critical symptom information, enhancing the model's ability to make personalized and accurate predictions. The critical contributions of our work are as follows:

- We introduced an innovative framework called SDCANet, which effectively combines CNN and Attention Networks to significantly enhance the accuracy of disease diagnosis by leveraging both symptoms and demographics information.
- We highlighted the pivotal role of utilizing symptoms as valuable information for disease prediction, leading to early intervention and improved patient outcomes.
- We conducted a comprehensive comparative evaluation of the proposed model against existing works, demonstrating its superior accuracy, precision, F1 score, and recall metrics.
- We utilized a private dataset to ensure the authenticity and practicality of our study, allowing the model to capture the complexities of clinical settings and thereby increasing the relevance and applicability of the proposed disease prediction framework

The subsequent sections of this chapter are organized as follows. Section 2 provides a concise overview of current state-of-the-art approaches and the application of DL in disease prediction. Section 3 details the problem formulation relevant to this study. Section 4 presents an in-depth exploration of a DL-based approach for disease prediction. Subsequently, in Sect. 5, the proposed solution is evaluated and assessed. Finally, Sect. 6 concludes the chapter by summarizing the essential findings and implications of the research.

2 Related Work

2.1 Text Classification

There has been significant research on text classification and analysis methods, which are essential tasks in natural language processing (NLP). Several approaches have been proposed, including traditional machine learning (ML) methods such as support vector machines and decision trees and deep learning techniques such as CNNs and RNNs. In recent years, there has been increasing interest in using pre-trained language models, such as BERT and GPT-2, for text classification and analysis tasks. These models have achieved state-of-the-art performance in various NLP tasks, including sentiment analysis, named entity recognition, and question-answering. Significant progress has been made in developing text representation methods, including contextualized word embeddings. The text classification and analysis field rapidly evolved, with new methods and techniques being developed and tested regularly.

In medical diagnosis, a growing trend in recent years has been using NLP techniques to analyze text data. This approach has facilitated the development of novel methodologies to predict a diverse range of medical conditions based on information derived from textual data, including symptoms, demographics, and medical history. In particular, there were three broad categories of methods: rule-based, traditional ML-based, and DL-based. Rule-based approaches utilize expert knowledge to create rules to identify symptoms and infer diagnoses. Traditional ML-based approaches require labeled data sets to train statistical models, which are then used to predict diagnoses based on new text data. In contrast, DL-based approaches automatically use neural networks to learn relevant features from raw input data. Each method presented unique advantages and limitations, with the choice of which approach to use depending on the specific task requirements and available data. However, several challenges remained to be overcome, including dealing with imbalanced data sets and ensuring the reliability and interpretability of the results. Nonetheless, NLP techniques continued to promise to enhance medical diagnosis and patient outcomes, thereby reducing healthcare costs.

2.2 Symptoms-Based Disease Prediction Models

Many studies have proposed disease prediction models using patient-collected symptoms. These studies stressed the significance of utilizing the abundant symptom data to enhance diagnostic accuracy, enable early disease detection, and support proactive healthcare interventions. Researchers aimed to leverage advanced data analysis and machine learning to build robust models that could revolutionize disease prediction and enhance patient outcomes.

Kanchan et al. [2] introduced a comprehensive system for predicting diseases based on the symptoms exhibited by patients. To achieve accurate predictions, the researchers employed two machine learning algorithms, K Nearest Neighbor (KNN) and CNN. The CNN algorithm demonstrated an overall disease prediction accuracy of 84.5%, surpassing the performance of the KNN technique. However, it was observed that KNN required more time and memory resources than CNN. Besides, Keniya et al. [10] employed the KNN algorithm to predict diseases by assigning data points to the class containing most of the K closest data points. However, this method was susceptible to noise and missing data. Similarly, Taunk et al. [15] also utilized the KNN method and demonstrated its high Precision in various cases, including predicting diabetes and heart risks. However, there needed to be more data for disease classification. Moreover, Cao et al. [3] proposed a methodology that utilizes a Support Vector Machine (SVM) to classify diseases based on symptoms. The SVM model was effective in disease prediction but required more time for accurate predictions. The method had the potential for improved accuracy. However, it relied on classifying objects using a hyperplane, which was only partially effective. In the medical context, where symptoms correspond to multiple diseases, this binary classification approach was limited as it could only handle two classes, which needs to be improved for accurate diagnosis.

The work [14] proposed a method that utilized the Naïve Bayes algorithm to predict a limited number of diseases, including Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis. However, their research did not involve working with an extensive data set to predict a broader range of diseases. Another approach [4] also utilized the Naïve Bayes classifier. However, their disease prediction model yielded poor accuracy, and they did not employ a standard dataset for training purposes. The work in [9] presented an approach to automate patient classification during hospital admission, focusing on symptoms extracted from text data. They leveraged the Bag of Words (BOW) model to generate word features from the textual information on diagnosing ten common diseases based on the set of symptoms, utilizing various machine learning algorithms such as Random Forest, SVM, Decision Tree, Multinomial Naive Bayes, Logistic Regression.

The approaches mentioned earlier have explored various machine-learning techniques for disease prediction. However, these existing works did not address factors such as efficiency, accuracy, the limited size of the data set used for model training, and the consideration of a restricted set of symptoms for disease diagnosis.

3 Problem Formulation

We aimed to tackle this challenge by developing a predictive model that effectively utilizes patients' symptoms and demographic information. All algorithms were presented using a single admission scenario to simplify the process. Our ultimate objective was to leverage this predictive model to provide a specific disease as a suitable treatment option based on a comprehensive analysis of the patient's symptoms and demographic factors. Hence, we strived to enhance the accuracy and efficiency of diagnosing patients and guided them toward the most appropriate course of treatment. The input and output were defined as follows:
Input: For each admission, the input data of our model consists of a patient demographics p and set of symptoms $(s_1, s_2, ..., s_N)$

- The patient's demographic data, denoted as p, encompasses various attributes of the patient, such as age and sex, commonly documented in electronic health records (EHR). The demographic p is a generated sentence; for instance, if the patient is a man and 24 years old, the demographic p would be "He is 24 years old".
- A patient's set of symptoms s contains multiple symptoms. We construct s by discretizing each symptom.

Output: The predicted disease, denoted $\hat{y}$, represents the result of our model, which is determined based on the symptoms a patient exhibits and their demographic information.

4 Methodology

In this section, we present our proposed model **S**ymptoms-based **D**isease Prediction with **C**NN-**A**ttention **Net**works (SDCANet) for predicting diseases by utilizing the symptoms and demographics of patients.

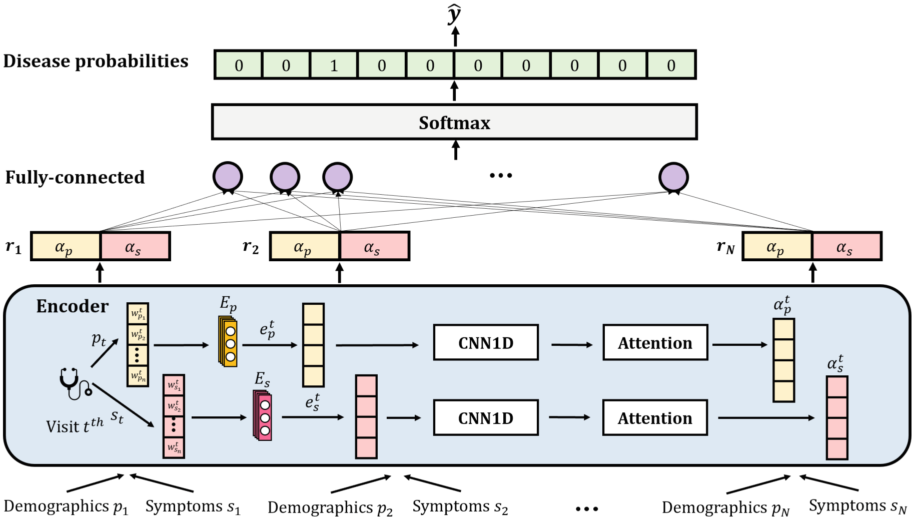

Fig. 2. The proposed model for symptom-based disease prediction

4.1 SDCANet Overview

Fig. 2 shows that the encoder block plays a vital role in our approach. SDCANet, specifically designed for predicting future diagnoses, incorporates two primary components: a patient representation encoder and a disease predictor. The patient representation encoder module teaches us intricate representations of symptoms and patient demographics. The accuracy and reliability of our predictions are enhanced by comprehensively understanding symptoms and demographic characteristics employing 1DCNN. Incorporating an attention mechanism further strengthens our ability to learn patient representations based on exhibited symptoms. This approach focuses on the relevant symptoms and their significance, capturing the unique characteristics of each patient. The attention-based approach ensures adaptability to individual patients, resulting in personalized and accurate disease predictions. Finally, disease probabilities are computed using the representations obtained from previous modules. Leveraging these learned representations and considering various factors based on attention scores, our model provides valuable insights into the likelihood of a patient having a specific disease.

4.2 Patient Representation Encoder

The objective of patient representation was to acquire a concise and informative vector that captures the patient's condition. During a clinical visit, doctors diagnose diseases by considering symptoms and demographic information. Our model also incorporated these two features in its analysis. Given that symptoms can be of varying importance, we proposed an innovative approach to represent them effectively. Taking inspiration from the attention mechanism in neural networks [1,6], we deviated from the conventional approach of merging all symptoms into a single lengthy text. Instead, we employed an attentive multi-view learning framework to learn comprehensive representations of unified symptoms. This framework enabled us to learn comprehensive and unified representations of symptoms by treating each symptom's information as a distinct viewpoint within the broader context of symptomatology.

Inspired in [17], the grouping strategy can be implemented to convert a collection of symptom embeddings into a unified representation. However, it could not effectively capture the varying importance of individual symptoms, a crucial factor that doctors must consider carefully during the diagnostic process. For instance, as illustrated in Fig. 1, cough is a more significant symptom than other symptoms (such as fever and sore throat) for Patient 2 and Patient 3, as it directly contributes to identifying diseases related to the respiratory system. Furthermore, given that symptoms occur at different frequencies, an averaging strategy alone would disregard unusual symptoms and fail to detect certain diseases. For example, Blood in the urine is a rare symptom associated with kidney stone disease, necessitating greater attention when representing a set of symptoms to accurately recommend diseases specific to the kidneys.

Symptoms and Demographics Encoders. The initial layer was word embedding, which transformed a symptom category from a sequence of words into a series of semantic vectors with lower dimensions. We denoted the word sequence of a category of demographics and symptoms as $[w^t_{p_1}, w^t_{p_2}, ..., w^t_{p_n}]$, $[w^t_{s_1}, w^t_{s_2}, ..., w^t_{s_n}]$, where p_n, s_n represent the number of words of demographics and symptoms, respectively. Through the utilization of a word embedding look-up table $W_e \in \mathbb{R}^{V \times N}$, this sequence was converted into a sequence of word vectors $[e^t_1, e^t_2, ..., e^t_N]$ where V represents the vocabulary size, while N denoted the dimension of the word embeddings.

Patient Representation. The second layer consists of 1DCNN, which considers the importance of local word contexts within demographics for acquiring their representations. Similarly, in the case of symptom sets, specific symptoms held significant contextual information for determining the corresponding disease. For example, we considered the set of symptoms such as fever, cough, and sore throat. In this case, the local contexts of symptoms such as sore throat and cough were crucial for recognizing their relevance to respiratory-related conditions. Hence, we employed CNNs to effectively learn contextual representations of words by capturing their local contexts. We denoted the contextual representation of the i-th word as c_i^t, calculated as Eq. 1.

$$c_i^t = ReLU(F_t \times e_{(i-K):(i+K)}^t) + b_t \tag{1}$$

where $e_{(i-K):(i+K)}^t$ is the concatenation of word embeddings from position $(i-K)$ to $(i+K)$. $F_t \in \mathbb{R}^{N_f \times (2K+1)D}$ and $b_t \in \mathbb{R}^{N_f}$ are the kernel and bias parameters of the CNN filters, N_f is the number of CNN filters and $2K+1$ is their window size. ReLU is the nonlinear activation function. The output of this layer is a sequence of contextual word representations, i.e., $[c_1^t, c_2^t, ..., c_N^t]$.

Then, we applied the third layer, which encompasses a word-level attention network [7], which addressed the fact that various words within the same symptom category tend to possess varying levels of informativeness when learning representations of symptoms. Certain symptoms held greater informativeness within a symptom category than others in representing the underlying disease. For example, in a set of symptoms such as fever, cough, and sore throat, we observed that the sore throat may carry more significance in indicating a specific disease than the Fever symptom. Acknowledging the significance of identifying crucial words within different symptom categories offered the potential to acquire more informative symptom representations. To accomplish this, we introduced a word-level attention network that selectively highlights important words within the context of each symptom category. The attention weight of the i-th word in a symptom category is denoted as α_i^t, and formulated as Eqs. 2 and 3:

$$a_i^t = q_t^T tanh(V_t \times c_i^t + v_t) \tag{2}$$

$$\alpha_i^t = \frac{exp(a_i^t)}{\Sigma_{j=1}^N exp(a_j^t)} \tag{3}$$

where V_t and v_t are the projection parameters, q_t denotes the attention query vector. The ultimate representation of a category of symptoms is determined by the summation of contextual representations of its words, each representation weighted by its corresponding attention weight. In other words, the final representation can be calculated as the sum of the contextual representations of the words multiplied by their respective attention weights. It is formulated as Eq. 4:

$$r_t = \Sigma_{j=1}^N \alpha_j^t c_j^t \tag{4}$$

Within our SDCANet approach, the symptoms encoder module played a crucial role in acquiring representations for both historical symptoms reported by patients and the candidate symptoms that are to be recommended. This module was responsible for learning and capturing the essential features of these symptoms, enabling accurate representation and subsequent analysis.

4.3 Disease Predictor

In the disease prediction stage, attention scores played an important role in improving the accuracy and reliability of diagnoses. Attention scores are calculated by assigning importance scores to various demographic and symptom representation features. This allowed the model to focus on the most relevant information and capture its contributions to the disease prediction task. Different attention mechanisms, such as dot-product attention or self-attention, are employed depending on the model's architecture. By incorporating attention scores, the model gained the ability to extract meaningful patterns and relationships from input data effectively. The combination and integration of attention scores are achieved by concatenating or merging them to form a unified representation. These representations were then passed through a fully connected layer, where each neuron applies a weight to the corresponding attention score. This transformation and integration process enabled the model to incorporate the significance of different features and enhanced the overall understanding of the input data. The model became more adept at capturing the critical information necessary for accurate disease prediction by giving higher weights to essential features.

The softmax layer played a pivotal role in disease prediction by converting the transformed representation from the fully connected layer into probability distributions. Each disease was assigned a probability, indicating the likelihood of a patient having that specific disease based on their symptoms and demographic information. This enabled the model to quantify the confidence of its predictions and prioritize the most probable diseases. The model determined the predicted diagnosis by selecting the disease with the highest probability or applying a threshold. The utilization of attention scores and the softmax layer enhanced the accuracy of disease prediction and provided valuable insights for medical professionals in making informed decisions and improving patient outcomes.

5 Experiments

5.1 Dataset

In this investigation, an in-depth analysis was performed on the Patient Admission dataset [9], comprising 230,479 samples such as age, gender, and patient clinical symptoms. We used data from March 2016 to March 2021 from the Medical Center of My Tho City, Tien Giang in Vietnam, from the admissions and discharge office, outpatient department, accident, emergency department, and related reports. Patient information was collected through manual or semi-automatic retrieval, primarily utilizing the QRCode embedded in the patient's health insurance card. The dataset has fields including the patient's age (captured in the AGE field), gender (represented as 1 for male and 0 for female in the SEX field), an extensive compilation of clinical symptoms (stored within the CLINICAL SYMPTOMS field), and the ID DISEASES field showed the type of diagnosed disease by the patient's ICD10 disease code. The data set encompasses ten commonly encountered disease types in Vietnamese hospitals, as listed in Table 1. Documenting the clinical symptoms falls upon the medical staff stationed at the admission and discharge office. These healthcare professionals meticulously record the observed clinical symptoms upon the patient's declaration. The clinical symptom data is comprehensive, covering various aspects such as physical fitness, abnormal vital signs, and the manifestation of symptoms before and during the patient's arrival at the hospital.

Table 1. Statistical Analysis of a Patient Admission Dataset for Disease Prediction

No.	Disease name	#samples
1	Neoplasms	16271
2	Endocrine, Nutritional and metabolic diseases	38672
3	Diseases of the eye and adnexa	18443
4	Diseases of the circulatory system	37782
5	Diseases of the respiratory system	41888
6	Diseases of the skin and subcutaneous tissue	7044
7	Diseases of the musculoskeletal system and connective tissue	35427
8	Diseases of the genitourinary system B212	17503
9	Pregnancy, childbirth, and the puerperium	3666
10	Injury, poisoning and certain other consequences of external	13783

5.2 Experimental Setup

The experimental results were obtained by testing an Ubuntu 18.04172 operating system server. The server had 20 different CPU configurations and boasted a substantial 64GB RAM capacity. The convolutional neural network in the experiments incorporated an attention mechanism constructed using the Keras library. The experiments are evaluated through 5-fold cross-validation.

5.3 Results

Table 2 demonstrates a comparison of different metrics, such as Accuracy (ACC), Precision, Recall, F1 score, and Area Under the Curve (AUC), for various optimizers, including Adam, SGD, and RMSProp, across different values of the MAX_SYMPTOM_LEN parameter (100, 150, and 200), which represents the maximum length of symptoms used. Besides these hyper-parameters, some others are a batch size of 64, a learning rate of 0.001, and the ReduceLROnPlateau scheduler, which are also utilized. Overall, the results show that the optimizers achieve comparable performance across most metrics and values of the number of symptoms. In terms of accuracy, all optimizers perform reasonably well, with

Table 2. Performance Evaluation (mean ± standard deviation) of SDCANet Model. The best results are marked in bold, and the second-highest results are underlined.

Metrics	Optimizers	Maximum number of symptoms		
		100	150	200
ACC	**Adam**	0.879 ± 0.0044	0.880 ± 0.0030	0.879 ± 0.0037
	SGD	0.879 ± 0.0052	0.879 ± 0.0048	0.878 ± 0.0045
	RMSProp	0.882 ± 0.0007	0.882 ± 0.0008	**0.883** ± **0.0024**
Precision	**Adam**	0.900 ± 0.0008	0.898 ± 0.0017	0.898 ± 0.0017
	SGD	0.899 ± 0.0020	0.899 ± 0.0018	0.899 ± 0.0027
	RMSProp	0.901 ± 0.0015	0.902 ± 0.0025	**0.906** ± **0.0018**
Recall	**Adam**	0.860 ± 0.0095	0.864 ± 0.0064	0.862 ± 0.0071
	SGD	0.860 ± 0.0094	0.860 ± 0.0086	0.859 ± 0.0082
	RMSProp	**0.866** ± **0.0027**	0.864 ± 0.0024	0.865 ± 0.0046
F1	**Adam**	0.879 ± 0.0055	0.880 ± 0.0038	0.878 ± 0.0040
	SGD	0.878 ± 0.0061	0.878 ± 0.0055	0.878 ± 0.0057
	RMSProp	**0.882** ± **0.0010**	0.882 ± 0.0011	0.881 ± 0.0027
AUC	**Adam**	0.985 ± 0.0011	0.985 ± 0.0010	0.985 ± 0.0006
	SGD	0.986 ± 0.0009	0.986 ± 0.0007	0.986 ± 0.0010
	RMSProp	0.986 ± 0.0005	0.986 ± 0.0002	**0.987** ± **0.0005**

scores ranging from 0.879 to 0.883. RMSProp consistently achieves the highest accuracy across all values of the number of symptoms, with the highest score obtained for the number of symptoms equal to 200 being 0.883. Regarding the Precision metric, all optimizers achieve scores greater than 0.89, indicating a relatively high proportion of correctly predicted positive instances. Similar to accuracy, RMSProp tends to perform slightly better, with the highest Precision of 0.906 for the number of symptoms equal to 200.

Moreover, Recall values vary between 0.859 and 0.866, suggesting that the optimizers have slightly more difficulty identifying all positive instances correctly. However, the differences in the recall scores are relatively small, and RMSProp again tends to perform slightly better for the number of symptoms equal to 100. Besides, the F1-score value ranges from 0.878 to 0.882. Similarly to the previous metrics, RMSProp achieves the highest F1 scores, particularly for the number of symptoms equal to 100, obtaining a value of 0.882. The AUC measures the classifier's overall discriminative power. All optimizers achieve high AUC scores above 0.98, indicating strong performance. RMSProp consistently outperforms the other optimizers, with the highest AUC of 0.987 for the number of symptoms equal to 200. In summary, while all optimizers demonstrate competitive performance across most metrics, RMSProp performs slightly better, especially in ACC, Precision, F1, and AUC. RMSProp may be the preferred optimizer, particularly for the number of symptoms equal to 200.

Table 3. Comparison results of the proposed model SDCANet with other baselines in percent. The best results are marked in bold

Method	ACC	Precision	Recall	F1	AUC
Logistic Regression [9]	0.791	0.799	0.791	0.795	–
Deep Bidirectional LSTM with Tokenizer combined with Sequences [12]	0.873	0.892	0.857	0.874	0.982
SDCANet (Ours)	**0.883**	**0.906**	**0.865**	**0.881**	**0.987**

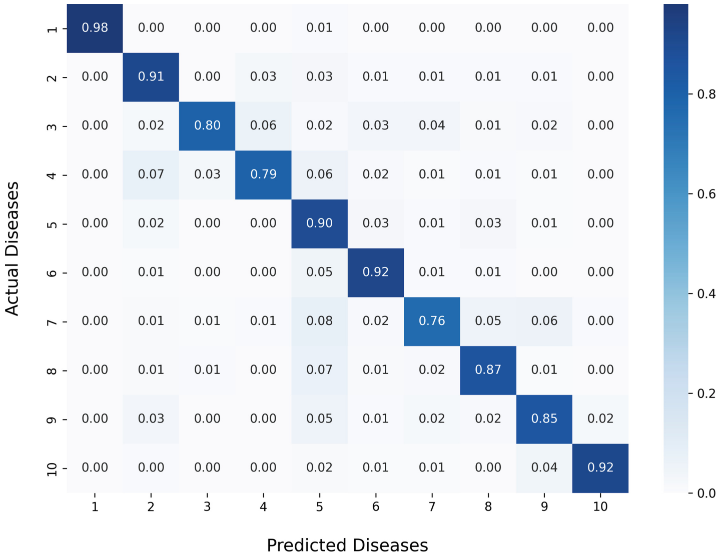

Fig. 3. Confusion matrix on Patient Admission Dataset predicted by SDCANet. Disease name numbers (from 1 to 10) are mapped in Table 1.

Table 3 compares the performance of the proposed model SDCANet with several baseline methods across different evaluation metrics, namely Accuracy (ACC), Precision, Recall, and F1 score. Regarding Accuracy, the proposed model outperforms all the baseline methods from previous studies, achieving an impressive accuracy of 88.3%. From the result reports as shown in Fig. 3, we calculate the confusion matrix, where each predicted outcome is derived from averaging the results across five different folds. Our observations show that the methods demonstrate strong performance in predicting conditions related to Neoplasms (No. 1), Endocrine, Nutritional, and metabolic diseases (No. 2), Diseases of the respiratory system (No. 5), Diseases of the skin and subcutaneous tissue (No. 6), as well as Injury, poisoning and certain other consequences of external (No. 10). All these mentioned diseases get performance over 90%.

Table 4. Ablation Study Results for SDCANet

Method	ACC	Precision	Recall	F1	AUC
SDCANet w/o Demographics	0.873	0.899	0.849	0.872	0.982
SDCANet w/o 1DCNN	0.863	0.887	0.840	0.862	0.983
SDCANet w/o Attention	0.871	0.894	0.849	0.870	0.983
SDCANet	**0.883**	**0.906**	**0.865**	**0.881**	**0.987**

In order to assess the effectiveness of the different components in SDCANet, we conducted several experiments. Table 4 displays the results of these variations. Removing 1DCNN and Attention from SDCANet led to lower outcomes, highlighting their importance in improving the base model. This suggests that 1DCNN, Attention, and Demographic information are crucial for disease prediction. In conclusion, the complete SDCANet outperforms all variations, underscoring the significance of each component in our model.

6 Conclusion

In conclusion, our proposed SDCANet model significantly advanced disease prediction by effectively learning representations of symptoms and patient demographics. SDCANet improved the precise predictions by focusing on relevant symptoms and demographics by combining 1DCNN and Attention mechanism. Leveraging the attention mechanism further enhanced the model's ability to capture the unique traits of each patient, resulting in personalized and accurate disease predictions. Our evaluation demonstrated that SDCANet outperformed baseline methods across evaluation metrics, underscoring its effectiveness in disease prediction. Future research can explore integrating additional features like medications, procedures, and lab tests to enhance the model's accuracy. Additionally, incorporating longitudinal and genomics data can provide a better understanding of disease progression, enabling personalized treatment strategies. These advancements hold the potential to enhance healthcare decision-making and improve patient outcomes in the field of disease prediction.

References

1. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010 (2017) .https://doi.org/10.5555/3295222.3295349
2. Kanchan, B.D., Kishor, M.M.: Study of machine learning algorithms for special disease prediction using principal of component analysis. In: IEEE International Conference on Global Trends in Signal Processing Information Computing and Communication (ICGTSPICC) (2016). https://doi.org/10.1109/ICGTSPICC.2016.7955260

3. Cao, J., Wang, M., Li, Y., Zhang, Q.: Improved support vector machine classification algorithm based on adaptive feature weight updating in the Hadoop cluster environment. PloS ONE **14**(4), e0215136 (2019). https://doi.org/10.1371/journal.pone.0215136
4. Chhogyal, K., Nayak, A.: An empirical study of a simple Naive Bayes classifier based on ranking functions. In: Kang, B.H., Bai, Q. (eds.) AI 2016. LNCS (LNAI), vol. 9992, pp. 324–331. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-50127-7_27
5. Chen, J., Li, D., Chen, Q., Zhou, W., Liu, X.: Diaformer: automatic diagnosis via Symptoms Sequence Generation. In: AAAI Conference on Artificial Intelligence (2021). https://arxiv.org/abs/2112.10433
6. Wu, C., et al.: Neural news recommendation with attentive multi-view learning. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019), AAAI Press, pp. 3863–3869 (2019)
7. Wu, C., Wu, F., Liu, J., He, S., Huang, Y., Xie, X.: Neural demographic prediction using search query. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM 2019). Association for Computing Machinery, New York, NY, USA, pp. 654–662 (2019). https://doi.org/10.1145/3289600.3291034
8. Guilbert, J.J.: The world health report 2006: working together for health. Educ. Health (Abingdon, England) **19**(3), 385–387 (2006). https://doi.org/10.1080/13576280600937911
9. Le, K.D.D., Luong, H.H., Nguyen, H.T.: Patient classification based on symptoms using machine learning algorithms supporting hospital admission. In: Cong Vinh, P., Huu Nhan, N. (eds.) ICTCC 2021. LNICST, vol. 408, pp. 40–50. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-92942-8_4
10. Keniya, R., et al.: Disease prediction from various symptoms using machine learning. SSRN 3661426 (2020). https://doi.org/10.2139/ssrn.3661426
11. Kao, H.-C., Tang, K.-F., & Chang, E. (2018). Context-Aware Symptom Checking for Disease Diagnosis Using Hierarchical Reinforcement Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11902
12. Nguyen, H.T., Dang Le, K.D., Pham, N.H., et al.: Deep bidirectional LSTM for disease classification supporting hospital admission based on pre-diagnosis: a case study in Vietnam. Int. J. Inf. Tecnol. **15**, 2677–2685 (2023). https://doi.org/10.1007/s41870-023-01283-x
13. Milella, F., Minelli, E.A., Strozzi, F., Croce, D.: Change and innovation in healthcare: findings from literature. ClinicoEconomics Outcomes Res. **2021**, 395–408 (2021). https://doi.org/10.2147/CEOR.S301169
14. Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., Karve, A.: Disease prediction using machine learning. Int. Res. J. Eng. Technol. (IRJET) **6**(2019), 831–833 (2019). https://doi.org/10.1126/science.1065467
15. Taunk, K., De, S., Verma, S., wetapadma, A.: A brief review of the nearest neighbor algorithm for learning and classification. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1255–1260. IEEE (2019). https://doi.org/10.1109/ICCS45141.2019.9065747

16. Islam, S.R., Sinha, R., Maity, S.P., Ray, A.K.: Deep learning on symptoms in disease prediction. Mach. Learn. Healthcare Appl. (2021). https://doi.org/10.1002/9781119792611.ch5
17. Tan, Y., et al.: 4SDrug: symptom-based set-to-set small and safe drug recommendation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022), New York, NY, USA, pp. 3970–3980. Association for Computing Machinery (2022). https://doi.org/10.1145/3534678.3539089