DDxT: Deep Generative Transformer Models for Differential Diagnosis

Mohammad Mahmudul Alam
University of Maryland, Baltimore County
m256@umbc.edu

Edward Raff
Booz Allen Hamilton
Laboratory for Physical Sciences
University of Maryland, Baltimore County
Raff_Edward@bah.com

Tim Oates
University of Maryland, Baltimore County
oates@umbc.edu

Cynthia Matuszek
University of Maryland, Baltimore County
cmat@umbc.edu

Abstract

Differential Diagnosis (DDx) is the process of identifying the most likely medical condition among the possible pathologies through the process of elimination based on evidence. An automated process that narrows a large set of pathologies down to the most likely pathologies will be of great importance. The primary prior works have relied on the Reinforcement Learning (RL) paradigm under the intuition that it aligns better with how physicians perform DDx. In this paper, we show that a generative approach trained with simpler supervised and self-supervised learning signals can achieve superior results on the current benchmark. The proposed Transformer-based generative network, named DDxT, autoregressively produces a set of possible pathologies, i.e., DDx, and predicts the actual pathology using a neural network. Experiments are performed using the DDXPlus dataset. In the case of DDx, the proposed network has achieved a mean accuracy of 99.82% and a mean F1 score of 0.9472. Additionally, mean accuracy reaches 99.98% with a mean F1 score of 0.9949 while predicting ground truth pathology. The proposed DDxT outperformed the previous RL-based approaches by a big margin. Overall, the automated Transformer-based DDx generative model has the potential to become a useful tool for a physician in times of urgency.

1 Introduction

Differential Diagnosis (DDx) is referred to the process of systematically identifying a disease from a possible set of pathologies through the process of elimination based on a patient's medical history and physical examinations [5]. During a clinical process, a doctor asks several questions about the patient's symptoms and antecedents (medical history). Based on the response, possible differential diagnoses are narrowed down. If there is uncertainty about the underlying condition, then a medical examination is performed or additional tests are suggested. Given a patient's information and symptoms, an automated system that narrows down the possible pathologies if not identifying the exact one will be of great benefit. In particular, such improvements could help lower-performing doctors or those in under-resourced communities obtain better diagnostic outcomes [25]. Moreover, in times of emergency, an automated system that has access to the patient's medical history and current conditions will be quite valuable.

In recent years, automated diagnosis systems using machine learning have increasingly developed [30, 17, 6, 12]. Existing works have demonstrated the potential of such automated systems in

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

performing complete blood count (CBC) test [1], syndrome detection [15], coronavirus, heart disease, and diabetes detection [14], and more. Previous work such as *Diaformer* [4] also demonstrated success in automated diagnosis using a sequence of explicit and implicit symptoms of a disease. But what is lacking is the details of the symptoms, the patient's previous medical history, and relevant information such as age, and gender. In a DDx process, a doctor would consider all of these information.

In this paper, an automated DDx system is proposed using Transformer [28], named *DDxT*, that would take a sequence of all the patient's necessary information as input to perform DDx by autoregressively generating a set of most likely pathologies and finally, predict the ground truth pathology using a neural network. This sequence of patient information will contain age, gender, medical history, and evidence, i.e., symptoms. Transformer architecture is employed since it is currently state-of-the-art for sequence generation [3]. Asking questions to a patient and acquiring information can easily be done through an automated system. The challenging part is to make an intelligent decision based on the acquired information which will be addressed in this paper. This will be beneficial not only during the time of emergency but also as an assistive tool to the doctor during the diagnosis process.

2 Related Works

Recent works have demonstrated the feasibility of the machine learning-based automated diagnosis system. Such work is presented by [10] where a Transformer-based model is utilized for the differential diagnosis. To perform the task, multi-modal magnetic resonance imaging (MRI) is utilized where a sequence of the brain and spinal cord MRI is processed by the Transformer. Their model performed considerably better than the previous approaches, however, the work is limited only to the diagnosis of demyelinating diseases. Likewise, [24] presented an ensemble approach for automated diagnosis. Their approach involves multiple deep learning-based approaches where the final prediction is the ensemble of all the predictions. On the other hand, [1] proposed an automated complete blood count (CBC) test system which is a very common test in medical diagnosis. Their approach employed YOLO [22] object detection algorithm for blood cell detection.

An image-based classifier is a powerful tool for automated diagnosis. Such a system is presented by [20] where convolutional neural networks (CNN) are utilized to diagnose Coronary Artery Disease (CAD) using Myocardial Perfusion Imaging (MPI). Their system utilizes and compares performance on pre-trained VGG-16 [23] and DenseNet-121 [11] architectures. In the same fashion, [16] developed an automated classification system for fungal keratitis. Their system uses ResNet [8] architecture with fungal hyphae images for binary classification of fungal keratitis. Similarly, [17, 18] both employed EfficientNet [26] for the detection of COVID-19 using X-ray images and Malaria classification from the blood smear images, respectively. In a slightly different manner, [21] adopts vision transformer-based Swin-UNETR [7] model to automatic retinal lesion segmentation from spectral-domain optical coherence tomography (SD-OCT) images.

The rest of the paper is organized as follows. section 3 will cover the proposed method including a description of the dataset, network, and training procedure. Next, section 4 will highlight the results and compare the proposed method to the RL agent-based methods. Finally, we conclude in section 5 with a discussion of the limitations of our approach.

3 Proposed Method

In this paper, differential diagnosis will be performed using a generative Transformer which will take a sequence of patient information as input and predict a sequence of most likely pathologies as differential diagnosis, and finally, the most likely pathology will be predicted using a classifier. In the following subsections, a brief description of the dataset, proposed network architecture, and training process will be discussed.

3.1 Dataset

For differential diagnosis, along with evidence, i.e., symptoms, patient's antecedents (medical history) and personal details such as age and sex are necessary information. DDXPlus [27] dataset is such a dataset that contains synthetically generated 1.3M patient information where each sample contains

patient details, evidence, ground truth differential diagnoses, and the ground truth condition. The dataset has a total of 49 pathologies that cover various age groups, sexes, and patients with a broad spectrum of medical history. We note that our work assumes the fidelity of the data since obtaining diagnostic data and medical history from patients comes at high expense, legal hurdles, ethics review, and slow collection rate [19]. Such challenges are beyond the scope of our study.

The dataset is preprocessed so that it can be processed by the Transformer. Each patient's information in the dataset contains age, sex, initial evidence, evidence (symptoms), ground truth differential diagnosis, and ground truth pathology. The age is categorized into 8 groups in the following way: `[less than 1)`, `[1-4]`, `[5-14]`, `[15-29]`, `[30-44]`, `[45-59]`, `[60-74]`, and `[above 75]`. Sex is represented by `M` for male and `F` for female. The Initial and rest of the evidence were acquired by back-and-forth questioning with a patient. Differential diagnosis contains a set of likely pathologies with a probability score for each pathology based on the evidence. Therefore, the ground truth DDx output sequence is organized in descending order of the probability score of each pathology, i.e., the order of prediction is significant and the pathology with a higher probability needs to be predicted first. Finally, the ground truth pathology is what the patient actually has.

Special tokens are incorporated to facilitate the learning process. Particularly, `<bos>` indicates the beginning of the sequence, `<sep>` token is used to separate each type of information, and to indicate the end of sequence `<eos>` is used. Since all sequences need to be equal in size, `<pad>` token is used in shorter sequences to fill out the sequence up to the maximum length, and longer sequences are truncated. Each patient's information is preprocessed as follows. First `<bos>` is used to initiate a sequence. Next, age, sex, initial evidence, and evidence all are stacked together using `<sep>` in between. Finally, the end of the sequence is indicated by `<eos>` token. To cover the unknown words in special circumstances, `<unk>` token is included in the vocabulary.

3.2 Network Architecture

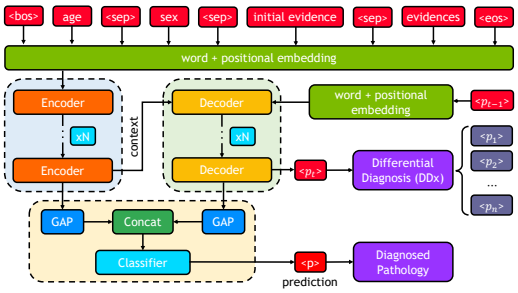

Figure 1: The block diagram of the proposed deep generative network architecture. The encoder blocks are shadowed with blue and the decoder blocks are shadowed with green. The classifier section is shadowed with orange. DDx is the set of n pathologies from <p_1> to <p_n> and the classifier predicts the diagnosed pathology <p>.

After preprocessing the dataset, a vocabulary is built using all the unique tokens. The input string is split into words and using the generated vocabulary, each word is replaced with the associated index of the word in the vocabulary. The encoder vocabulary length is 436 and the decoder vocabulary length is 54 (49 pathologies + 5 special tokens). Next, these integer values are utilized to gather the associated word embedding and added with positional embedding so that the order of the word in a sequence is recognized by the network. Similarly, the decoder input tokens are also preprocessed, and word and positional embedding are applied. The Transformer architecture consists of encoder and decoder blocks. Each of the blocks contains a self-attention mechanism, a brief description of which is provided in Appendix A. The encoder will process the patient's information and feed the context to the decoder. The decoder will be initialized with the p_0=`<bos>` token which will iteratively take previously generated pathology p_{t-1} as input and use tokens p_0 to p_{t-1} to generate a new possible pathology p_t until it reaches `<eos>` token. Both encoder and decoder are repeated $N(N = 6)$ times which will help recognize richer context. The decoder output is the DDx, a sequence of most likely pathologies.

The final layer of the encoder holds the processed context information of the evidence, i.e., symptoms and relevant patients' information, and the final layer of decoder holds the information of all the possible likely pathologies. Therefore, combining both features will be quite advantageous in predicting the actual pathology. As a result, Global average pooling (GAP) is applied to both the encoder and decoder features, concatenated, and fed to a classifier. The classifier is a two-layer neural network. The first layer contains the same number of features as the encoder or decoder, and the second layer has the same number of logits as the number of classes in the dataset. Both layers

are preceded by layer normalization [2]. In between the layers, GELU activation [9] is used. The classifier predicts the ground truth pathology among the most likely DDx. The full block diagram of the network architecture is presented in Figure 1.

3.3 Training

During training, the input size of the encoder and decoder must be fixed. Therefore, the maximum sequence length for the encoder is set to be 80, and the maximum sequence length for the decoder is set to be 40 by truncating or adding `<pad>` tokens. The built vocabulary has 436 unique tokens thus the vocab size for the word embedding is set to be 436. For the embedding layers, the feature size is set to 128 and the feature size of the multi-layer perceptron (MLP) of the encoders and decoders is increased 4 times. In the self-attention layers, 4 heads are used and the encoder and decoder are repeated 6 times. A categorical cross-entropy loss is employed for both the decoder output and the classifier which are added together to compute the final loss. To regularize the network, Dropout with a rate of 0.1 and layer normalization are employed. The loss function is optimized using the Adam [13] optimizer and trained for a total of 20 epochs. The initial learning rate is set to 10^{-3} with an exponential decay learning rate scheduler of the decay rate of $\gamma = 0.95$.

4 Results

The proposed network predicts a sequence of most likely pathologies, i.e., DDx, and the actual pathology among the DDx. Both the predicted DDx sequence and the predicted pathology are compared with the ground truth DDx sequence and pathology. The ground truth DDx sequence is organized in descending order of probability distribution of most likely pathologies. As a result, the positional embedding plays an important role in maintaining the correct prediction order leading to a better performance. The ground truth DDx sequence is compared with the predicted sequence elementwise and the mean result is computed. For evaluation, Accuracy, Precision, Recall, and F1 scores are considered. In the following subsections, a comparison of the proposed method with the RL agent-based automated diagnosis methods is performed. Subsequently, the performance of DDx pathology sequence generation and pathology classification will be analyzed and discussed.

4.1 Comparison

The baseline models that perform automatic diagnosis using the DDXPlus dataset are Reinforcement Learning (RL)-based agents. Adaptive Alignment of Reinforcement Learning and Classification (AARLC) presented by [29] is such a system that employs an RL-based agent to adaptively acquire the patient's symptoms and subsequently uses a classifier to predict the pathology. The process continues iteratively thus generating a DDx sequence of pathologies. Similarly, the baseline automatic symptom detector (BASD) [27] utilizes an RL-based agent to gather evidence and an MLP classifier to predict the pathology. Table 1 shows the comparison of the performance of the proposed DDxT model with the baseline RL agent-based models.

Table 1: Our DDxT improves Precision and thus F1 score significantly, showing value in a generative approach to retrieving accurate diagnoses over the prior RL agent-based approaches. The best results for each metric are highlighted in **bold**.

Method	GTPA@1	DDP	DDR	DDF1	GM
AARLC	99.21	69.53	**97.73**	0.7824	87.68
BASD	97.15	88.34	85.03	0.8369	90.03
DDxT	**99.98**	**94.84**	94.65	**0.9472**	**96.45**

The comparison is performed in terms of top-1 ground truth pathology accuracy (GTPA@1), Precision, Recall, and F1 score of DDx denoted as DDP, DDR, and DDF1 following the convention of [27]. AARLC gets the highest recall score but a much lower precision score, lesser than the BASD model, therefore, a lower F1 score. On the other hand, DDxT has a balanced performance in terms of both precision and recall. As a result, it achieved the new highest F1 score of 0.9472 in the DDXPlus

dataset. Additionally, the proposed method outperforms the previous approaches in terms of top-1 accuracy. Moreover, the accuracy, precision, and recall are combined by geometric mean (GM) also shown in Table 1 to compare the effectiveness of each method. DDxT achieves the best result of 96.45 with a big margin over the previous RL agent-based methods.

4.2 DDx Pathology Sequence Generation

The predicted DDx sequence is compared with the ground truth DDx sequence. Since the ground truth sequence is organized in descending order by the probability distribution, predictions are compared element-wise and the mean result is computed per sequence. To evaluate all the metrics, a confusion matrix is built. In DDx pathology sequence generation, the proposed method achieved 99.82% mean accuracy and a mean F1 score of 0.9472%. The confusion matrix of the generated pathology sequence is presented in Appendix B which demonstrates the robustness of the proposed generative method. The accuracy, precision, recall, and F1 score of all the pathology classes are also presented in Appendix B. Among all the pathologies, the highest F1 score of 0.9946 is achieved for **Myasthenia gravis**, and the minimum F1 score of 0.8643 is achieved for **Pancreatic neoplasm**.

4.3 Pathology Classification

The proposed network takes the processed feature of the encoder and decoder using a GAP, concatenating them together and feeding them into a classifier for the final pathology classification, i.e., given the list of evidence and set of pathologies (DDx), the final classifier will predict the actual pathology among the most likely DDx pathologies. The results of pathology classification are also evaluated in terms of accuracy, precision, recall, and F1 score. Since the classifier has both encoder and decoder information, it shows significant robustness in classification where the network achieved a mean accuracy of 99.98% with a mean F1 score of 0.9949. Additionally, the mean precision and recall scores achieved are 99.61% and 99.44%, respectively. The minimum F1 score achieved 0.8567 is for the **Acute rhinosinusitis**. The confusion matrix of classification along with metric scores for all the pathologies are presented in Appendix C.

Some conditions, like **Unstable angina**, **Acute rhinosinusitis**, and **Chronic rhinosinusitis** obtain lower precision for varying recall rates. These conditions may need to be considered distinctly in the case of the condition's likelihood to a given population, the risk of the condition itself, and other factors to decide if such conditions are useful to detect in this fashion. Separately, the vast majority of conditions can be detected and a conservative threshold may be used to increase confidence in deployment while expecting a limited reduction in missed diagnoses.

5 Conclusion and Limitations

In this paper, an automated system of autoregressively generating DDx pathologies and predicting the actual pathology among them is presented. The proposed network uses a Transformer architecture where patient information and evidence are processed by the encoder. Next, the decoder generates a set of likely pathologies. The pathology sequences are generated in the most likely to the least likely order. Afterward, the features of both the encoder and decoder are concatenated and fed to a smaller classifier for the final prediction of the most likely pathology. Experimental results on the DDXPlux dataset demonstrate the feasibility and robustness of DDxT where it achieved a mean accuracy of 99.98% and a mean F1 score of 0.9472 in the DDx. Moreover, while predicting the pathology it achieved a mean accuracy of 99.98% with a mean F1 score of 0.9949. Nevertheless, the proposed system does not acquire and assumes the preexistence of the evidence. Therefore, the performance of the system has a dependency on the correct rendering of accurate information by an authorized user or another automated system. Besides, pathologies such as **Acute rhinosinusitis**, **Chronic rhinosinusitis** that displayed lower precision scores with varying recall need to be addressed distinctly. However, in general, the proposed system performed the desired goal of automating differential diagnosis and has the potential to emerge as an assistive tool for the physician during the diagnosis process.

References

[1] Mohammad Mahmudul Alam and Mohammad Tariqul Islam. Machine learning approach of automatic identification and counting of blood cells. *Healthcare technology letters*, 6(4):103–108, 2019.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4432–4440, 2022.

[5] Chad E Cook and Simon Décary. Higher order thinking about differential diagnosis. *Brazilian Journal of Physical Therapy*, 24(1):1–7, 2020.

[6] Hossam Faris, Maria Habib, Mohammad Faris, Haya Elayan, and Alaa Alomari. An intelligent multimodal medical diagnosis system based on patients' medical questions and structured symptoms for telemedicine. *Informatics in Medicine Unlocked*, 23:100513, 2021.

[7] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pages 272–284. Springer, 2022.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[10] Chuxin Huang, Weidao Chen, Baiyun Liu, Ruize Yu, Xiqian Chen, Fei Tang, Jun Liu, and Wei Lu. Transformer-based deep-learning algorithm for discriminating demyelinating diseases of the central nervous system with neuroimaging. *Frontiers in immunology*, 13, 2022.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[12] Emrah Irmak. Covid-19 disease diagnosis from paper-based ecg trace image data using a novel convolutional neural network model. *Physical and Engineering Sciences in Medicine*, 45(1):167–179, 2022.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Naresh Kumar, Nripendra Narayan Das, Deepali Gupta, Kamali Gupta, and Jatin Bindra. Efficient automated disease diagnosis using machine learning models. *Journal of Healthcare Engineering*, 2021, 2021.

[15] Hongyin Luo, Shang-Wen Li, and James Glass. Knowledge grounded conversational symptom detection with graph memory networks. *arXiv preprint arXiv:2101.09773*, 2021.

[16] Jian Lv, Kai Zhang, Qing Chen, Qi Chen, Wei Huang, Ling Cui, Min Li, Jianyin Li, Lifei Chen, Chaolan Shen, et al. Deep learning-based automated diagnosis of fungal keratitis with in vivo confocal microscopy images. *Annals of Translational Medicine*, 8(11), 2020.

[17] Gonçalo Marques, Deevyankar Agarwal, and Isabel de la Torre Díez. Automated medical diagnosis of covid-19 through efficientnet convolutional neural network. *Applied soft computing*, 96:106691, 2020.

[18] Gonçalo Marques, Antonio Ferreras, and Isabel de la Torre-Diez. An ensemble-based approach for automated medical diagnosis of malaria using efficientnet. *Multimedia tools and applications*, 81(19):28061–28078, 2022.

[19] Catherine Ordun, Alexandra N Cha, Edward Raff, Byron Gaskin, Alex Hanson, Mason Rule, Sanjay Purushotham, and James L Gulley. Intelligent sight and sound: A chronic cancer pain dataset. In *NeurIPS*, 2021.

[20] Nikolaos I Papandrianos, Anna Feleki, Elpiniki I Papageorgiou, and Chiara Martini. Deep learning-based automated diagnosis for coronary artery disease using spect-mpi images. *Journal of Clinical Medicine*, 11(13):3918, 2022.

[21] Daniel Philippi, Kai Rothaus, and Mauro Castelli. A vision transformer architecture for the automated segmentation of retinal lesions in spectral domain optical coherence tomography images. *Scientific Reports*, 13(1):517, 2023.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] MV Sunena Rose and NV Sobhana. Automated diagnosis of diseases using integrated machine learning approaches. In *International Conference on Soft Computing and Pattern Recognition*, pages 195–204. Springer, 2021.

[25] Elena Ashtari Tafti. *Technology, Skills, and Performance: The Case of Robots in Surgery*. Institute for Fiscal Studies, 2022.

[26] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[27] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[29] Hongyi Yuan and Sheng Yu. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *arXiv preprint arXiv:2112.00733*, 2021.

[30] Quan Zhang, Yuliang Liu, Guohua Liu, Geng Zhao, Zhigang Qu, and Weiming Yang. An automatic diagnostic system based on deep learning, to diagnose hyperlipidemia. *Diabetes, metabolic syndrome and obesity: targets and therapy*, pages 637–645, 2019.

A Transformer

Transformer [28] is an attention-based model that uses its input to generate query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ to compute self-attention given in Equation 1 where d is the feature dimension. The self-attention layer is accompanied by multi-layer perceptron, layer normalization, and skip connections to form an attention block. The blocks are repeated multiple times. The sequence of the patient's information will be processed on a stack of blocks called the encoder which gives context to the decoder, another stack of decoder blocks that produces the output. A look-ahead mask is applied to the decoder to generate output autoregressively so that the current prediction p_t only relies on the previous predictions from p_0 to p_{t-1}.

$$\texttt{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V} \tag{1}$$

B DDx Comprehensive Results

The confusion matrix for the DDx sequence generation is presented in Figure 2. The dataset has a total of 49 pathologies. So, the confusion matrix is of the size of 49×49. The ground truth classes are shown in the row and the predicted labels are in the column. Visually, the proposed method has very few false positives and false negatives while generating the most likely pathologies.

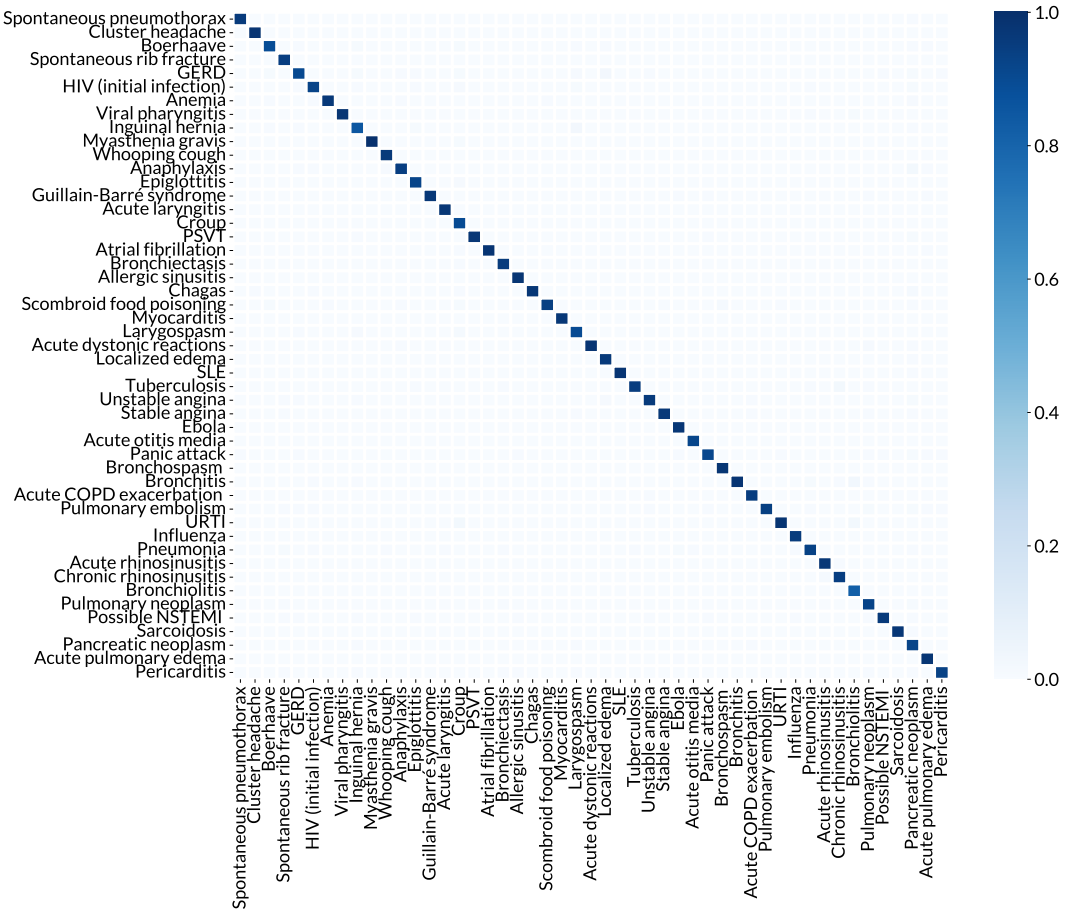

Figure 2: Confusion matrix for the DDx sequence generation where each element, i.e, pathology of the sequence is evaluated elementwise. Each row indicates the ground truth class and each column indicates the predicted class.

The accuracy, precision, recall, and F1 score of all the pathology classes are also presented in Table 2. Among all the pathologies, the highest F1 score of 0.9946 is achieved for **Myasthenia gravis**, and the lowest F1 score of 0.8643 is achieved for **Pancreatic neoplasm**.

Table 2: Classification results of the Differential Diagnosis (DDx) where the sequence of pathologies are generated given a set of evidence.

Pathology	**Acc. (%)**	**Prec. (%)**	**Rec. (%)**	**F1**
Spontaneous pneumothorax	99.91	96.11	95.65	0.9588
Cluster headache	99.92	96.69	97.73	0.9720
Boerhaave	99.77	93.04	89.01	0.9098
Spontaneous rib fracture	99.95	93.07	94.38	0.9372
GERD	99.56	91.54	90.44	0.9098
HIV (initial infection)	99.67	91.28	92.30	0.9179
Anemia	99.67	96.89	96.46	0.9668
Viral pharyngitis	99.94	98.44	97.57	0.9800
Inguinal hernia	99.84	90.58	84.94	0.8767
Myasthenia gravis	99.96	99.53	99.40	0.9946
Whooping cough	99.99	97.67	95.87	0.9676
Anaphylaxis	99.59	94.68	94.42	0.9455
Epiglottitis	99.92	94.13	91.81	0.9296
Guillain-Barré syndrome	99.78	97.43	96.83	0.9713
Acute laryngitis	99.92	96.93	97.10	0.9701
Croup	99.93	85.13	89.49	0.8726
PSVT	99.86	98.38	97.20	0.9778
Atrial fibrillation	99.85	98.26	98.03	0.9815
Bronchiectasis	99.85	94.42	95.06	0.9474
Allergic sinusitis	99.99	96.73	97.47	0.9710
Chagas	99.69	97.05	96.91	0.9698
Scombroid food poisoning	99.56	95.13	93.94	0.9453
Myocarditis	99.78	97.01	96.76	0.9689
Larygospasm	99.92	91.21	89.60	0.9040
Acute dystonic reactions	99.86	98.81	97.78	0.9829
Localized edema	99.95	94.54	95.76	0.9515
SLE	99.92	98.28	98.01	0.9814
Tuberculosis	99.76	95.39	94.62	0.9500
Unstable angina	99.63	93.21	95.54	0.9436
Stable angina	99.76	95.28	96.31	0.9579
Ebola	99.98	94.62	96.96	0.9578
Acute otitis media	99.95	96.46	91.67	0.9400
Panic attack	99.65	94.24	91.40	0.9280
Bronchospasm	99.92	95.42	97.52	0.9646
Bronchitis	99.85	98.48	97.70	0.9809
Acute COPD exacerbation	99.95	96.88	94.66	0.9576
Pulmonary embolism	99.68	96.89	94.13	0.9549
URTI	99.88	96.28	97.69	0.9698
Influenza	99.81	94.05	95.11	0.9458
Pneumonia	99.67	92.32	93.32	0.9282
Acute rhinosinusitis	99.95	96.64	96.69	0.9667
Chronic rhinosinusitis	99.90	94.25	94.52	0.9438
Bronchiolitis	100.00	92.59	81.97	0.8696
Pulmonary neoplasm	99.73	89.91	92.44	0.9116
Possible NSTEMI	99.60	94.99	95.73	0.9536
Sarcoidosis	99.90	97.90	96.90	0.9740
Pancreatic neoplasm	99.56	81.10	92.52	0.8643
Acute pulmonary edema	99.80	94.53	97.47	0.9598
Pericarditis	99.72	92.57	93.25	0.9291
Mean	99.82	94.84	94.65	0.9472

C Pathology Classification Comprehensive Results

The confusion matrix of the pathology classification is presented in Figure 3. The proposed method showed a robust performance while classifying pathology from encoder and decoder features which can be observed from the figure.

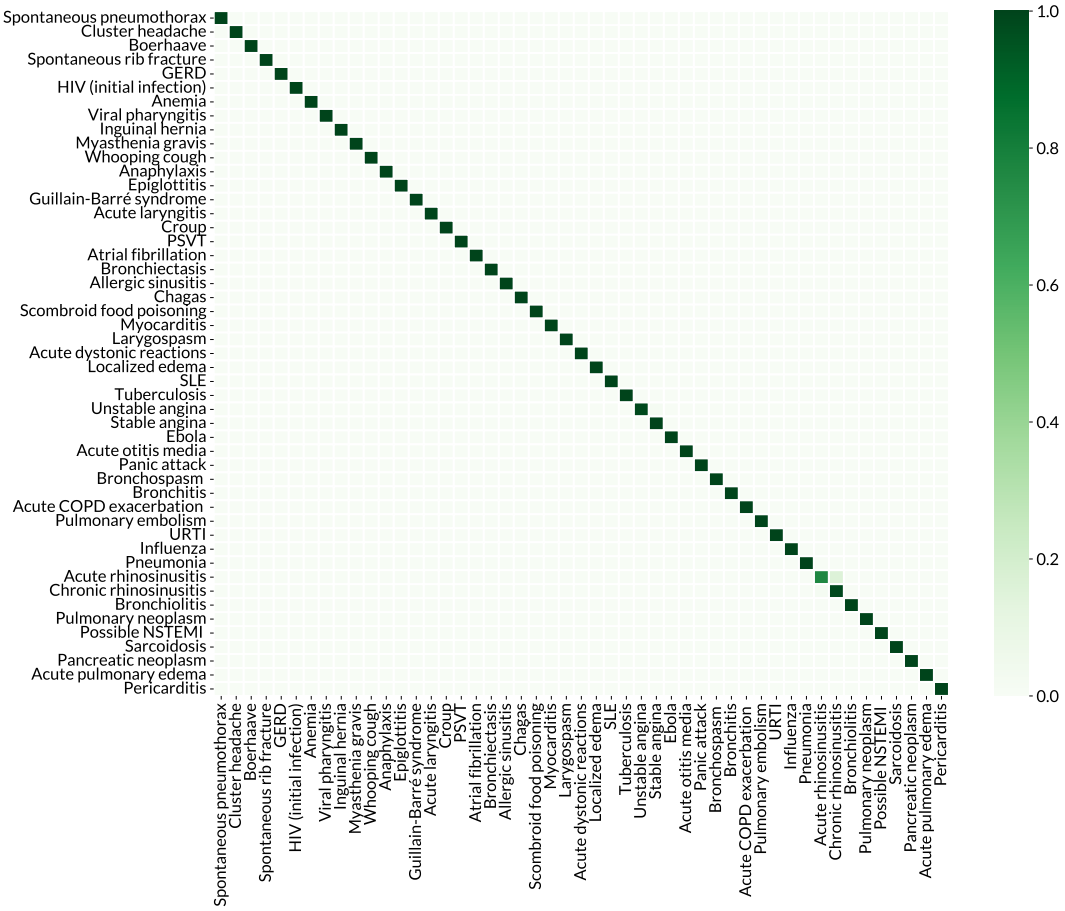

Figure 3: Confusion matrix of the final pathology classification. True labels are shown in the row and the predicted labels are shown in the column.

The proposed method has achieved a mean accuracy of 99.98% with a mean F1 score of 0.9949. Additionally, the mean precision and recall scores achieved are 99.61% and 99.44%, respectively. The accuracy, precision, recall, and F1 score of all the classes for pathology classification are also presented in Table 3.

Table 3: Results of the pathology classification the classifier processed the encoder and the decoder information to predict the most likely pathology among the predicted set of differential diagnoses.

Pathology	**Acc. (%)**	**Prec. (%)**	**Rec. (%)**	**F1**
Spontaneous pneumothorax	100.00	100.00	100.00	1.0000
Cluster headache	100.00	100.00	100.00	1.0000
Boerhaave	100.00	100.00	100.00	1.0000
Spontaneous rib fracture	100.00	100.00	100.00	1.0000
GERD	100.00	100.00	100.00	1.0000
HIV (initial infection)	100.00	100.00	100.00	1.0000
Anemia	100.00	100.00	100.00	1.0000
Viral pharyngitis	99.97	99.72	99.77	0.9975
Inguinal hernia	100.00	100.00	100.00	1.0000
Myasthenia gravis	100.00	100.00	100.00	1.0000
Whooping cough	100.00	100.00	100.00	1.0000
Anaphylaxis	100.00	100.00	100.00	1.0000
Epiglottitis	100.00	100.00	100.00	1.0000
Guillain-Barré syndrome	100.00	100.00	100.00	1.0000
Acute laryngitis	99.97	99.41	99.29	0.9935
Croup	100.00	100.00	100.00	1.0000
PSVT	100.00	100.00	100.00	1.0000
Atrial fibrillation	100.00	100.00	100.00	1.0000
Bronchiectasis	100.00	100.00	100.00	1.0000
Allergic sinusitis	100.00	100.00	100.00	1.0000
Chagas	100.00	100.00	100.00	1.0000
Scombroid food poisoning	100.00	100.00	100.00	1.0000
Myocarditis	100.00	100.00	100.00	1.0000
Larygospasm	100.00	100.00	100.00	1.0000
Acute dystonic reactions	100.00	100.00	100.00	1.0000
Localized edema	100.00	100.00	100.00	1.0000
SLE	100.00	100.00	100.00	1.0000
Tuberculosis	100.00	100.00	100.00	1.0000
Unstable angina	99.96	99.96	98.37	0.9916
Stable angina	99.97	98.07	100.00	0.9902
Ebola	100.00	100.00	100.00	1.0000
Acute otitis media	100.00	100.00	100.00	1.0000
Panic attack	100.00	100.00	100.00	1.0000
Bronchospasm	100.00	100.00	100.00	1.0000
Bronchitis	100.00	100.00	100.00	1.0000
Acute COPD exacerbation	100.00	100.00	100.00	1.0000
Pulmonary embolism	100.00	100.00	100.00	1.0000
URTI	100.00	100.00	100.00	1.0000
Influenza	100.00	100.00	100.00	1.0000
Pneumonia	100.00	100.00	100.00	1.0000
Acute rhinosinusitis	99.65	97.36	76.49	0.8567
Chronic rhinosinusitis	99.65	86.30	98.62	0.9205
Bronchiolitis	100.00	100.00	100.00	1.0000
Pulmonary neoplasm	100.00	100.00	100.00	1.0000
Possible NSTEMI	100.00	100.00	99.97	0.9998
Sarcoidosis	100.00	100.00	100.00	1.0000
Pancreatic neoplasm	100.00	100.00	100.00	1.0000
Acute pulmonary edema	100.00	100.00	100.00	1.0000
Pericarditis	100.00	100.00	100.00	1.0000
Mean	99.98	99.61	99.44	0.9949