This Patient Looks Like *That* Patient: Prototypical Networks for Interpretable Diagnosis Prediction from Clinical Text

Betty van Aken[1], **Jens-Michalis Papaioannou**[1], **Marcel G. Naik**[2], **Georgios Eleftheriadis**[2], **Wolfgang Nejdl**[3], **Felix A. Gers**[1], **Alexander Löser**[1]

[1] Berliner Hochschule für Technik (BHT),
[2] Charité Berlin,
[3] Leibniz University Hannover

{bvanaken,michalis.papaioannou,gers,aloeser}@bht-berlin.de,
{marcel.naik,georgios.eleftheriadis}@charite.de, nejdl@L3S.de

Abstract

The use of deep neural models for diagnosis prediction from clinical text has shown promising results. However, in clinical practice such models must not only be accurate, but provide doctors with interpretable and helpful results. We introduce ProtoPatient, a novel method based on prototypical networks and label-wise attention with both of these abilities. ProtoPatient makes predictions based on parts of the text that are similar to prototypical patients–providing justifications that doctors understand. We evaluate the model on two publicly available clinical datasets and show that it outperforms existing baselines. Quantitative and qualitative evaluations with medical doctors further demonstrate that the model provides valuable explanations for clinical decision support.

1 Introduction

Medical professionals are faced with a large amount of textual patient information every day. Clinical decision support systems (CDSS) aim to help clinicians in the process of decision-making based on such data. We specifically look at a subtask of CDSS, namely the prediction of clinical diagnosis from patient admission notes. When clinicians approach the task of diagnosis prediction, they usually take similar patients into account (from their own experience, clinic databases or by talking to their colleagues) who presented with typical or atypical signs of a disease. They then compare the patient at hand with these previous encounters and determine the patient's risk of having the same condition.

In this work, we propose ProtoPatient, a deep neural approach that imitates this reasoning process of clinicians: Our model learns prototypical characteristics of diagnoses from previous patients and bases its prediction for a current patient on the similarity to these prototypes. This results in a model that is both inherently interpretable and provides clinicians with pointers to previous prototypical patients. Our approach is motivated by Chen et al. (2019) who introduced prototypical part networks (PPNs) for image classification. PPNs learn prototypical parts for image classes and base their classification on the similarity to these prototypical parts. We transfer this work into the text domain and apply it to the extreme multi-label classification task of diagnosis prediction. For this transfer, we apply an additional label-wise attention mechanism that further improves the interpretability of our method by highlighting the most relevant parts of a clinical note regarding a diagnosis.

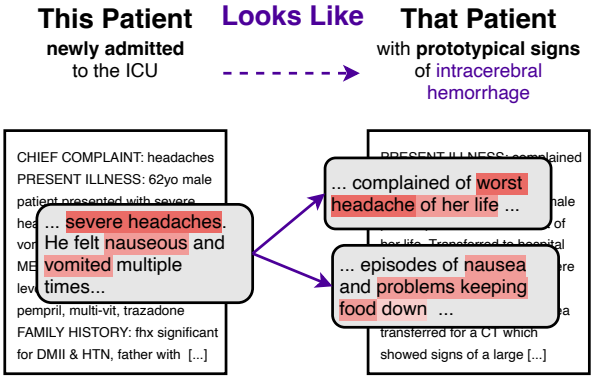

Figure 1: Basic concept of the ProtoPatient method. The model makes predictions for a patient (left side) based on the comparison to prototypical parts of earlier patients (right side).

While deep neural models have been widely applied to outcome prediction tasks in the past (Shamout et al., 2020), their black-box nature remains a large obstacle for clinical application (van

Aken et al., 2022). We argue that decision support is only possible when model predictions are accompanied by justifications that enable clinicians to follow a lead or to potentially discard predictions. With ProtoPatient we introduce an architecture that allows such decision support. Our evaluation on publicly available data shows that the model can further improve state-of-the-art performance on predicting clinical outcomes.

Contributions We summarize the contributions of this work as follows:
1. We introduce a novel model architecture based on prototypical networks and label-wise attention that enables interpretable diagnosis prediction. The system learns relevant parts in the text and points towards prototypical patients that have led to a certain decision.
2. We compare our model against several state-of-the-art baselines and show that it outperforms earlier approaches. Performance gains are especially visible in rare diagnoses.
3. We further evaluate the explanations provided by our model. The quantitative results indicate that our model produces explanations that are more faithful to its inner working than post-hoc explanations. A manual analysis conducted by medical doctors further shows the helpfulness of prototypical patients during clinical decision-making.
4. We release the code for the model and experiments for reproducibility.[1]

2 Task: Diagnosis Prediction from Admission Notes

The task of outcome prediction from admission notes was introduced by van Aken et al. (2021) and assumes the following situation: A new patient p gets admitted to the hospital. Information about the patient is written into an admission note a_p. The goal of the decision support system is to identify risk factors in the text and to communicate these risks to the medical professional in charge. For outcome diagnosis prediction in particular, the underlying model determines these risks by predicting the likelihood of a set of diagnoses C being assigned to the patient at discharge.

Data We evaluate our approach on the diagnosis prediction task from the clinical outcome prediction dataset introduced by van Aken et al. (2021).

[1]Public code repository: https://github.com/bvanaken/ProtoPatient

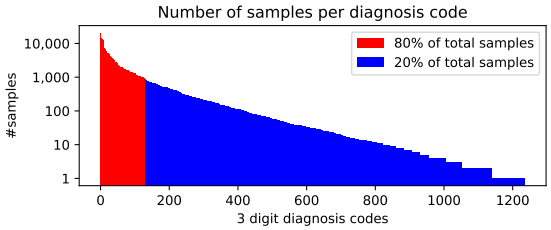

Figure 2: Distribution of ICD-9 diagnosis codes in MIMIC-III training set.

The data is based on the publicly available MIMIC-III database (Johnson et al., 2016). It comprises de-identified data from patients in the Intensive Care Unit (ICU) of the Beth Israel Deaconess Medical Center in Massachusetts in the years 2001-2012. The data includes 48,745 admission notes written in English from 37,320 patients in total. They are split into train/val/test sets with no overlap in patients. The admission notes were created by extracting sections from MIMIC-III discharge summaries which contain information known at admission time such as *Chief Complaint* or *Family History*. The notes are labelled with diagnoses in the form of 3-digit ICD-9 codes that were assigned to the patients at discharge. On average, each patient has 11 assigned diagnoses per admission from a total set of 1266 diagnoses.

Challenges Challenges surrounding diagnosis prediction can be divided into two main categories:

- **Predicting the correct diagnoses** The number of possible diagnoses is large (>1K) and, as shown in Figure 2, the distribution is extremely skewed. Since many diagnoses only have a few samples, learning plausible patterns is challenging. Further, each admission note describes multiple conditions, some being highly related, while others are not. The text in admission notes is also highly context dependent. Abbreviations like *SBP* (i.a. for *systolic blood pressure* or *spontaneous bacterial peritonitis*) have completely different meanings based on their context. Our models must capture these differences and enable users to check the validity of features used for a prediction.

- **Communicating risks to doctors** Apart from assigning scores to diagnoses, for a high-stake task such as diagnosis prediction, a system must be designed for medical professionals to understand and act upon its predictions. Therefore, models must provide faithful explanations for their pre-

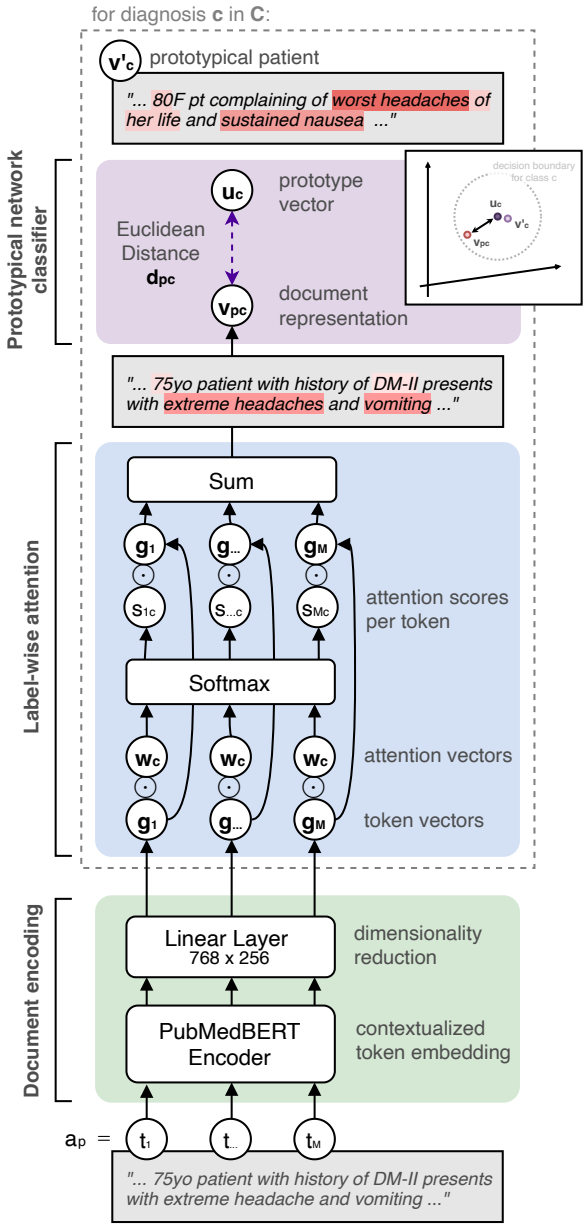

Figure 3: Schematic view of the ProtoPatient method. Starting at the bottom, document tokens get a contextualized encoding and are then transformed into a label-wise document representation $\mathbf{v_{pc}}$. The classifier simply considers the distance of this representation to a learned prototypical vector $\mathbf{u_c}$. The prototypical patient $\mathbf{v'_c}$ is the training example closest to the prototypical vector.

dictions and give clues that enable further clinical reasoning steps by doctors. These requirements are challenging, since interpretability of models often come with a trade-off in their prediction performance (Arrieta et al., 2019).

3 Method: ProtoPatient

To address the challenges above, we propose a novel model architecture called ProtoPatient, which adapts the concept of prototypical networks (Chen et al., 2019) to the extreme multi-label scenario by using label-wise attention and dimensionality reduction. Figure 3 presents a schematic overview. We further show how our model can be efficiently initialized to improve both speed and performance.

3.1 Learning Prototypical Representations

We encode input documents a_p (p indexes patients) into vectors $\mathbf{v_p}$ with dimension D and measure their distance to a learned set of prototype vectors. Each prototype vector $\mathbf{u_c}$ represents a diagnosis $c \in C$ in the dataset. The prototype vectors are learned jointly with the document encoder so that patients with a diagnosis can best be distinguished from patients without it. As a distance measure we use the Euclidean distance $d_{pc} = ||\mathbf{v_p} - \mathbf{u_c}||_2$ which Snell et al. (2017) identified as best suited for prototypical networks. We then calculate the sigmoid σ of the negative distances to get a prediction $\hat{y}_{pc} = \sigma(-d_{pc})$, so that documents closer to a prototype vector get higher prediction scores. We define the loss L as the binary cross entropy (BCE) between $\hat{y}_{pc}$ and the ground truth $y_{pc} \in \{0, 1\}$.

$$L = \sum_p \sum_c BCE(\hat{y}_{pc}, y_{pc}) \quad (1)$$

Prototype initialization Snell et al. (2017) define each prototype as the mean of the embedded support set documents. In contrast, we learn the label-wise prototype vectors end-to-end while optimizing the multi-label classification. This leads to better prototype representations, since not all documents are equally representative of a class, as taking the mean would suggest. However, using the mean of all support documents is a reasonable starting point. We set the initial prototype vectors of a class as $\mathbf{u_{c_{init}}} = \langle \mathbf{v_c} \rangle$, i.e. the mean of all document vectors $\mathbf{v_c}$ with class label c in the training set. We then fine-tune their representation during training. Initial experiments showed that this initialization leads to model convergence in half the number of steps compared to random initialization.

Contextualized document encoder For the encoding of the documents, we choose a Transformer-based model, since Transformers are capable of modelling contextualized token representations. For initializing the document encoder, we use the weights of a pre-trained language model. At the time of our experiments, the PubMedBERT (Tinn et al., 2021) model reaches the best results on a

range of biomedical NLP tasks (Gu et al., 2020). We thus initialize our document encoder with PubMedBERT weights[2] and further optimize it with a small learning rate during training.

3.2 Encoding Relevant Document Parts with Label-wise Attention

Since we face a multi-label problem, having only one joint representation per document tends to produce document vectors located in the center of multiple prototypes in vector space. This way, important features for single diagnoses can get blurred, especially if these diagnoses are rare. To prevent this, we follow the idea of prototypical part networks of selecting parts of the note that are of interest for a certain diagnosis. In contrast to Chen et al. (2019), we use an attention-based approach instead of convolutional filters, since attention is an effective way for selecting relevant parts of text. For each diagnosis c, we learn an attention vector $\mathbf{w_c}$. To encode a patient note with regard to c, we apply a dot product between $\mathbf{w_c}$ and each embedded token $\mathbf{g_{pj}}$, where j is the token index. We then apply a softmax.

$$s_{pcj} = softmax(\mathbf{g_{pj}^T}\, \mathbf{w_c}) \tag{2}$$

We use the resulting scores s_{pcj} to create a document representation $\mathbf{v_{pc}}$ as a weighted sum of token vectors.

$$\mathbf{v_{pc}} = \sum_j s_{pcj}\, \mathbf{g_{pj}} \tag{3}$$

This way, the document representation for a certain diagnosis is based on the parts that are most relevant to that diagnosis. We then measure the distance $d_{pc} = ||\mathbf{v_{pc}} - \mathbf{u_c}||_2$ to the prototype vector $\mathbf{u_c}$ based on the diagnosis-specific document representation $\mathbf{v_{pc}}$.

Attention initialization The label-wise attention vectors $\mathbf{w_c}$ determine which tokens the final document representation is based on. Therefore, when initializing them randomly, we start our training with document representations which might carry little information about the patient and the corresponding diagnosis. To prevent this cold start, we initialize the attention vectors $\mathbf{w_{c_{init}}}$ with tokens informative to the diagnosis c. This way, at training start, these tokens reach higher initial scores s_{pcj}. We consider tokens $\tilde{t}$ informative that surpass a TF-IDF threshold of h. We then use the average of all embeddings $\mathbf{g_{c\tilde{t}}}$ from $\tilde{t}$ in documents corresponding to the diagnosis.

$$\mathbf{w_{c_{init}}} = \langle \mathbf{g_{c\tilde{t}}} \rangle \tag{4}$$

with $\tilde{t} = t : \text{tf-idf}(t) > h$. We found h=0.05 suitable to get 5-10 informative tokens per diagnosis.

3.3 Compressing representations

Label-wise attention vectors for a label space with more than a thousand labels lead to a considerable increase in model parameters and memory load. We compensate this by reducing the dimensionality D of vector representations used in our model. We add a linear layer after the document encoder that both reduces the size of the document embeddings and acts as a regularizer, compressing the information encoded for each document. We find that reducing the dimensionality by one third ($D = 256$) leads to improved results compared to the full-size model, indicating that more dense representations are beneficial to our setup.

3.4 Presenting prototypical patients

For retrieving prototypical patients $\mathbf{v'_c}$ for decision justifications at inference time, we simply take the label-wise attended documents from the training data that are closest to the diagnosis prototype. By presenting their distances to the prototype vector, we can provide further insights about the general variance of diagnosis presentations. Correspondingly, we can also present patients with atypical presentation of a diagnosis by selecting the ones furthest away from the learned prototype.

4 Evaluating Diagnosis Predictions

4.1 Experimental Setup

Baselines We compare ProtoPatient to hierarchical attention models and to Transformer models pre-trained on (bio)medical text, representing two state-of-the-arts approaches for ICD coding and outcome prediction tasks, respectively.

- **Hierarchical attention models** Hierarchical Attention Networks (**HAN**) were introduced by Yang et al. (2016). They are based on bidirectional gated recurrent units, with attention applied on both the sentence and token level. Baumel et al. (2018) built **HA-GRU** upon this concept using only sentence-wise attention,

[2]Model weights from: https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext

	ROC AUC macro	ROC AUC micro	PR AUC macro
HAN (Yang et al., 2016)	83.38 ±0.13	96.88 ±0.04	13.56 ±0.01
HAN + Label Emb (Dong et al., 2021)	83.49 ±0.18	96.87 ±0.12	13.07 ±0.14
HA-GRU (Baumel et al., 2018)	79.94 ±0.57	96.65 ±0.12	9.52 ±1.01
HA-GRU + Label Emb (Dong et al., 2021)	80.54 ±1.67	96.67 ±0.22	10.33 ±1.70
ClinicalBERT (Alsentzer et al., 2019)	80.95 ±0.16	94.54 ±0.93	11.62 ±0.64
DischargeBERT (Alsentzer et al., 2019)	81.17 ±0.30	94.70 ±0.48	11.24 ±0.88
CORe (van Aken et al., 2021)	81.92 ±0.09	94.00 ±1.10	11.65 ±0.78
PubMedBERT (Tinn et al., 2021)	83.48 ±0.21	95.47 ±0.22	13.42 ±0.57
Prototypical Network	81.89 ±0.22	95.23 ±0.01	9.94 ±0.36
ProtoPatient	86.93 ±0.24	**97.32** ±0.00	**21.16** ±0.21
ProtoPatient + Attention Init	**87.93** ±0.07	97.24 ±0.02	17.92 ±0.65

Table 1: Results in % AUC for diagnosis prediction task (1266 labels) based on MIMIC-III data. The ProtoPatient model outperforms the baselines in micro ROC AUC and PR AUC. The attention initialization further improves the macro ROC AUC. ± values are standard deviations. Label Emb: Label Embeddings. Attention Init: Attention vectors initialized as described in Section 3.2.

while adding a label-wise attention scheme comparable to ProtoPatient. Dong et al. (2021) further show that pre-initialized **label embeddings** learned from ICD code co-occurrence improves results for both approaches. We thus evaluate the models with and without label embeddings.[3]

- **Transformers pre-trained on in-domain text** Alsentzer et al. (2019) applied clinical language model fine-tuning on two Transformer models based on the BioBERT model (Lee et al., 2020). **ClinicalBERT** was trained on all clinical notes in the MIMIC-III database, and **DischargeBERT** on all discharge summaries. They belong to the most widely used clinical language models and achieve high scores on multiple clinical NLP tasks. The **CORe** model (van Aken et al., 2021) is also based on BioBERT, but further pre-trained with an objective specific to patient outcomes, which achieved higher scores on clinical outcome prediction tasks. Tinn et al. (2021) introduced **PubMedBERT** which was, in contrast to the other models, trained from scratch on articles from PubMed Central with a dedicated vocabulary. It is currently the best performing approach on the BLURB (Gu et al., 2020) benchmark.

Training We train all baselines on the dataset introduced in Section 2. For training HAN and HA-GRU we use the code and best performing hyperparameters as provided by Dong et al. (2021). We further use their pre-trained ICD-9 label embeddings (for details, see Appendix A.1). For training the Transformer-based models and ProtoPatient, we use hyperparameters reported to perform best for BERT-based models by van Aken et al. (2021) and additionally optimize the learning rate and number of warm up steps with a grid search. We further truncate the notes to a context size of 512. See A.2 for all details on the chosen hyperparameters. We report the scores of all models as an average over three runs with different seeds.

Ablation studies ProtoPatient combines three strategies: Prototypical networks, label-wise attention and dimensionality reduction. We conduct ablation studies to measure the impact of each strategy. To this end, we apply both label-wise attention and dimensionality reduction to a PubMedBERT model using a standard classification head. We further train a prototypical network without label-wise attention and ProtoPatient with different dimension sizes. The results are found in Table 2 and 7.

Transfer to second data set Clinical text data varies from clinic to clinic. We want to test whether the patterns learned by the models are transferable to other data sources than MIMIC-III. We use another publicly available dataset from the i2b2 De-identification and Heart Disease Risk Factors Challenge (Stubbs and Uzuner, 2015) further processed into admission notes by van Aken et al. (2021). The data consists of 1,118 admission notes labelled with

[3]Note that Dong et al. (2021) also propose the H-LAN model, which is a combination of HAN and HA-GRU using label-wise attention on sentence and token level. However, the model is only applicable to smaller label spaces (<100) due to its memory footprint and thus cannot be evaluated on our task.

	ROC AUC macro
Dimensionality reduction	
ProtoPatient 768	83.56 ±0.17
ProtoPatient (our proposed model with D=256)	**86.93** ±0.24
Transformer vs. Prototypical	
PubMedBERT 768	83.48 ±0.21
PubMedBERT 768 + Label Attention	**84.10** ±0.25
ProtoPatient 768	83.56 ±0.17
Label-wise attention	
PubMedBERT 256	83.61 ±0.04
PubMedBERT 256 + Label Attention	**84.68** ±0.52

Table 2: **Ablation studies** comparing different dimension sizes and how a standard Transformer (PubMedBERT) performs with additional label-wise attention.

the ICD-9 codes for *chronic ischemic heart disease*, *obesity*, *hypertension*, *hypercholesterolemia* and *diabetes*. We evaluate models without fine-tuning on the new data to simulate a model transfer to another clinic. The resulting scores are reported in Table 3.

4.2 Results

We present the results of all models on the diagnosis prediction task in Table 1. In addition, we show the macro ROC AUC score across codes depending on their frequency in the training set in Figure 4. We summarize the main findings as follows.

ProtoPatient outperforms previous approaches The results show that ProtoPatient achieves the best scores among all evaluated models. Pre-initializing the attention vectors further improves the macro ROC AUC score. Ablation studies show that all components play a role in improving the results. A prototypical network without label-wise attention is not able to capture the extreme multi-label data. PubMedBERT using a standard classification head also benefits from label-wise attention, but not to the same extent. Combining prototypical networks and label-wise attention thus brings additional benefits. The choice of dimension size is another important factor. Using 768 dimensions (the standard BERT base size) appears to lead to over-parameterization in the attention and prototype vectors. Using 256 dimensions also improves generalization, which is shown in producing the best results on the i2b2 data set in Table 3.

Improvements for rare diagnoses Figure 4 shows that the ROC AUC improvements are particularly large for codes that are rare ($\leq$50 times) in the training set. Prototypical networks are known for their few-shot capabilities (Snell et al., 2017)

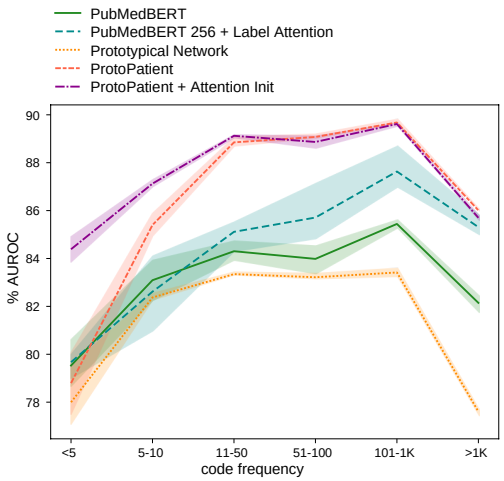

Figure 4: Macro ROC AUC scores regarding the frequency of ICD-9 codes in the training set. ProtoPatient models show the largest performance gain in rare codes ($\leq$100 samples). Attention initialization leads to large improvement for very rare codes ($<$10 samples).

	ROC AUC macro
PubMedBERT	82.11 ±0.12
Prototypical Network	69.65 ±0.22
ProtoPatient 768	85.28 ±0.49
ProtoPatient	**87.38** ±0.20
ProtoPatient + Attention Init	86.72 ±1.52

Table 3: Performance on a second data set based on clinical notes from the **i2b2 challenge** (Stubbs and Uzuner, 2015). ProtoPatient shows the highest degree of transferability. Further metrics shown in Table 8.

which also prove useful in our scenario with mixed label frequencies. For extremely rare codes that appear less than ten times, the attention initialization described in Section 3.2 further improves results. This indicates that the randomly initialized attention vectors need at least a number of samples to learn the most important tokens, and that pre-initializing them can accelerate this process.

PubMedBERT and HAN are the best baselines The pre-trained PubMedBERT and the HAN model achieve the highest scores among the baselines. Interestingly, PubMedBERT outperforms the Transformer models pre-trained on clinical text. This indicates that training from scratch with a domain-specific vocabulary is beneficial for the task. The scores of the HAN model further emphasize the importance of label-wise attention. The addition of label embeddings to HAN and HA-GRU, however, does not add significant improvements in our case.

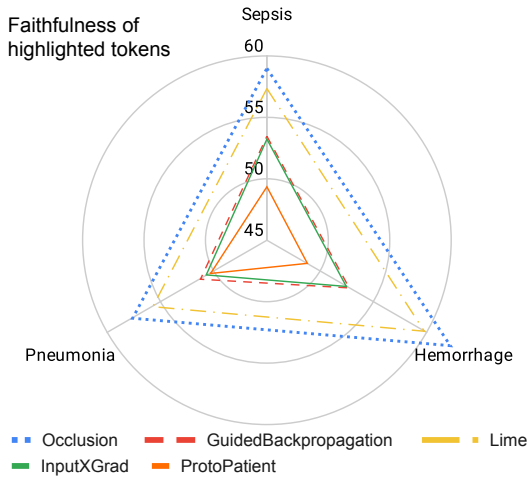

Figure 5: Evaluating faithfulness of highlighted tokens. Lower scores indicate more faithful explanations. ProtoPatient's token highlights are part of the model decision and thus more faithful than post-hoc explanations.

5 Evaluating Interpretability

We evaluate the interpretability of ProtoPatient with quantitative and qualitative analyses as follows.

Quantitative study on faithfulness Faithfulness describes how explanations correspond to the inner workings of a model, a property essential to their usefulness. We apply the explainability benchmark introduced by Atanasova et al. (2020) to compare the faithfulness of ProtoPatient's token highlights to post-hoc explanation methods. Following the benchmark, faithfulness is measured by incrementally masking highlighted tokens, expecting a steep drop in model performance if the tokens are indeed relevant to the model prediction. See B.1 for details. Due to the high computational costs of the evaluation, we limit our analyses to three diagnoses with a high severity to the ICU: Sepsis, intracerebral hemorrhage and pneumonia. We compare against four common post-hoc explanation methods, namely Lime (Ribeiro et al., 2016), Occlusion (Zeiler and Fergus, 2014), InputXGradient (Kindermans et al., 2016), and Gradient Backpropagation (Springenberg et al., 2014), which we apply to the PubMedBERT baseline. Figure 5 shows the results, for which lower scores mean more faithful explanations (i.e. a steeper drop in model performance). We see that ProtoPatient's explanations reach the lowest scores for all three labels, proving that they are more faithful than the post-hoc explanations. This is a result of the interpretable structure of ProtoPatient, in which model decisions are directly based on the highlighted parts. We show these parts, i.e. the tokens that are most frequently highlighted by the model for the three analyzed diagnoses, in B.2.

Manual analysis by medical doctors We conduct a manual analysis with two medical doctors (one specialized, one resident) to understand whether highlighted tokens and prototypical patients are helpful for their decisions. They used a demo application of ProtoPatient[4] and analyzed 20 random patient letters with 203 diagnoses in total. The results are shown in Table 4. The doctors first identified the principal diagnoses and then rated the corresponding prototypical patients presented by the model. Note that some patients have more than one principal diagnosis. In 21 of 23 cases, the prototypical samples were showing typical signs of the respective diagnosis and 17 of them were rated as helpful for making a diagnosis decision. Cases in which they were not helpful included very rare conditions or ones with a strong differ-

[4]Demo URL available at:
https://protopatient.demo.datexis.com

Analysis of prototypical patient cases (principal diagnoses)			
Q1: Prototypical patient shows typical clinical signs			
yes		no	
21		2	
Q2: Highlighted prototypical parts are relevant			
mostly	partially	hardly	
21	2	0	
Q3: Prototypical patient is helpful for diagnosis decision			
yes		no	
17		6	
Analysis of highlighted parts (all diagnoses)			
Q4: Highlighted tokens are relevant for diagnosis (i.e. describe diagnosis, symptoms or risk factors)			
	mostly	partially	hardly
TPs	78	3	7
FPs	50	12	9
FNs	22	10	12
Q5: Important tokens are missing from highlights			
	yes	no	
TPs	17	71	
FPs	13	58	
FNs	2	42	

Table 4: Results of the manual analysis conducted by medical doctors on ProtoPatient outputs. The prototypical patients were analyzed for the principal diagnoses only, while the highlighted parts of the patient letter at hand were analyzed for all diagnoses. Q1..5 denote the questions answered regarding each patient case.

ence to the specific case. They further analyzed the highlighted tokens for all diagnoses and found that they contained mostly relevant information in 150 cases. Examples of highlighted risk factors judged as plausible were *obesity* known to relate to *diabetes type II*, *untreated hypertension* to *heart failure* or a medication history of *anticoagulant coumadin* to *atrial fibrillation*. They also identified cases in which the highlighted tokens were partially or hardly relevant. In these cases, the highlighted tokens often included stop words or punctuation, indicating that the attention vector failed to learn relevant tokens. This was mainly observed in very frequent diagnoses such as *hypertension* or *anemia*, which corresponds to the lower model performance on these conditions (see Figure 4). This is because conditions very common in the ICU are often either not indicated in the clinical note or not labelled, so that the model cannot learn clear patterns regarding their relevant tokens.

6 Related Work

Diagnosis prediction from clinical notes Predicting diagnosis risks from clinical text has been studied using different methods. Fakhraie (2011) analyzed the predictive value of clinical notes with bag-of-words and word embeddings. Jain et al. (2019) experimented with adding attention modules to recurrent neural models. Recently, the use of Transformer models for diagnosis prediction has outperformed earlier approaches. van Aken et al. (2021) applied BERT-based models further pre-trained on clinical cases to predict patient outcomes. However, the black-box nature of these models hinders their application in clinical practice. We therefore introduce ProtoPatient, which uses Transformer representations, but provides interpretable predictions.

Prototypical networks for few-shot learning Prototypical networks were first introduced by Snell et al. (2017) for the task of few-shot learning. They initialized prototypes as centroids of support samples per episode and applied the approach to image classification tasks. Sun et al. (2019) adapted the approach to text documents with hierarchical attention layers. Recently, related approaches based on prototypical networks have been used for multiple few-shot text classification tasks (Wen et al., 2021; Zhang et al., 2021; Ren et al., 2020; Deng et al., 2020; Feng et al., 2023). In contrast to this body of work, we do not train our model in a few-shot scenario using episodic learning. However, our model shows related capabilities by improving results for diagnoses with few available samples.

Prototypical networks for interpretable models Chen et al. (2019) used prototypical networks in a different setup to build an interpretable model for image classification. To this end, they learn prototypical parts of images to mimic human reasoning. We adapt their idea and show how to apply it to clinical natural language. Recently, Ming et al. (2019) and Das et al. (2022) applied the concept of prototypical networks to text classification and showed how prototypical texts help to interpret predictions. In contrast to their work and following Chen et al. (2019), we identify prototypical *parts* rather than whole documents by using label-wise attention. This makes interpreting results easier and enables multi-label classification with over a thousand labels.

Label-wise attention Mullenbach et al. (2018) introduced label-wise attention for clinical text with the CAML model. Since then, the method has been further improved by hierarchical attention approaches (Baumel et al., 2018; Yang et al., 2016; Dong et al., 2021). Label-wise attention has mainly been used for ICD coding, a task related to diagnosis prediction that differs in the input data: ICD coding is done on notes that describe the whole stay at a clinic. In contrast, outcome diagnosis prediction uses admission notes as input and identifies diagnosis *risks* rather than the diagnoses already mentioned in the text. Our method–combining prototypical networks with label-wise attention–is particularly focused on detecting and highlighting those risks to enable clinical decision support.

7 Discussion

7.1 Reflection on the Challenges

Rudin (2019) urges to stop explaining black-boxes and to build interpretable models instead. With ProtoPatient we introduce a model with a simple decision process–*this patient looks like that patient*–that is understandable to medical professionals and inherently interpretable. An exemplary output is shown in Table 5. Our results indicate that the model is able to deal with contextual text in clinical notes, e.g. when identifying *SBP* as a risk factor for sepsis in B.2. In addition, it improves results on rare diagnoses, which are especially challenging for doctors to detect due to lack of experience

Admission note	Relevant parts of admission note	similar to	Parts of prototypical patient notes
PRESENT ILLNESS: Patient is a 35-year-old male pedestrian struck by a bicycle from behind with positive loss of consciousness for 6 minutes at the scene after landing on his head. At arrival at ER patient was confused, had multiple contusions noted on a head CT scan including bilateral frontal and right temporal contusions. His cervical spine and abdominal examinations were negative radiologically. The patient was then transferred to the Emergency Room. Patient had several episodes of vomiting during flight and during the trauma workup. He was assessed and was intubated for airway protection. The patient was given coma score of 9 upon initial assessment. Patient remaining hemodynamically stable throughout the transfer and throughout the workup in the ED. [. . .]	struck by a bicycle . . . loss of consciousness for 6 minutes . . . coma score 9 . . .	⟶	**cerebral hemorrhage** loss of consciousness . . . struck by vehicle . . . with a gcs of 10 . . .
	head CT scan . . . bilateral contusions . . . hemodynamically stable . . .	⟶	**skull fracture** head wound . . . right and left contusions . . . stable blood circulation . . .
	transferred to Emergency Room . . . several episodes of vomiting . . .	⟶	**shock** patient had multiple episodes of vomiting during the day . . .
	patient was confused . . . intubated for airway protection . . .	⟶	**acute respiratory failure** patient was disoriented . . . later intubated for protection. . .

Table 5: Exemplary output of ProtoPatient. The model identifies parts in an admission note that are similar to (i.e. *"look like"*) parts from prototypical patient notes seen during training, leading to the prediction of this diagnosis.

and sensitivity towards their signs. Overall, our approach demonstrates that interpretability can be improved without compromising performance. The modularity of the prototype vectors further allows clinicians to modify the model even after training. This can be done by adding prototypes whenever a new condition is found, or by directly defining certain patients as prototypical for the system.

7.2 Limitations of this work

Our model currently learns relations between diagnoses only indirectly, due to the label-wise nature of the classification. However, considering relations or conflicts between diagnoses is an important part of clinical decision-making. One way to include such relations is the addition of a loss term incorporating diagnosis relations, as proposed by Mullenbach et al. (2018). Another limitation is that the current model only considers one prototype per diagnosis, even though most diagnoses have multiple presentations, varying among patient groups. We therefore propose further research towards including multiple prototypes into the system.

8 Conclusion and Future Work

In this work, we present ProtoPatient which enables interpretable outcome diagnosis prediction from text. Our approach enhances existing methods in their prediction capability—especially for rare classes—and presents benefits to doctors by highlighting relevant parts in the text and pointing towards prototypical patients. The modularity of prototypical networks can be explored in future research. One promising approach is to introduce multiple prototypes per diagnosis, corresponding to the multiple ways diseases can present in a patient. Prototypes could also be added manually by medical professionals based on patients they consider prototypical. Another approach would be to initialize prototypes from medical literature and compare them to those learned from patients.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable feedback. This work was funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) under grant agreements 01MD19003B (PLASS) and 01MK2008MD (Service-Meister), as well as the Federal Ministry of Education and Research (BMBF) under grant agreement 16SV8845 (KIP-SDM).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *CoRR*, abs/1910.10045.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes: case study on icd code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. 2019. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8928–8939.

Anubrata Das, Chitrank Gupta, Venelin Kovatchev, Matthew Lease, and Junyi Jessy Li. 2022. Prototeх: Explaining model decisions with prototype tensors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2986–2997. Association for Computational Linguistics.

Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. 2020. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, pages 151–159. ACM.

Hang Dong, Víctor Suárez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*.

Najmeh Fakhraie. 2011. *What's in a Note? Sentiment Analysis in Online Educational Forums*. University of Toronto (Canada).

Jianzhou Feng, Qikai Wei, and Jinman Cui. 2023. Prototypical networks relation classification model based on entity convolution. *Comput. Speech Lang.*

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*.

Sarthak Jain, Ramin Mohammadi, and Byron C. Wallace. 2019. An analysis of attention over clinical notes for predictive tasks. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 15–21, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. 2019. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 903–913. ACM.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Haopeng Ren, Yi Cai, Xiaofeng Chen, Guohua Wang, and Qing Li. 2020. A two-phase prototypical network model for incremental few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1618–1629. International Committee on Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

Farah Shamout, Tingting Zhu, and David A Clifton. 2020. Machine learning for clinical outcome prediction. *IEEE reviews in Biomedical Engineering*, 14:116–126.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, 58:S20–S29.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485, Hong Kong, China. Association for Computational Linguistics.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Fine-tuning large neural language models for biomedical natural language processing. *CoRR*, abs/2112.07869.

Betty van Aken, Sebastian Herrmann, and Alexander Löser. 2022. What do you see in this patient? behavioral testing of clinical NLP models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 63–73, Seattle, WA. Association for Computational Linguistics.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Wen Wen, Yongbin Liu, Chunping Ouyang, Qiang Lin, and Tong Lee Chung. 2021. Enhanced prototypical network for few-shot relation extraction. *Inf. Process. Manag.*, 58(4):102596.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Jiawen Zhang, Jiaqi Zhu, Yi Yang, Wandong Shi, Congcong Zhang, and Hongan Wang. 2021. Knowledge-enhanced domain adaptation in few-shot relation classification. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2183–2191. ACM.

A Training Details

A.1 Label Embeddings for HAN and HA-GRU

We apply label embeddings to the HAN and HA-GRU network as proposed by Dong et al. (2021). In particular, we use the pre-initialized embeddings provided by the authors. Since they use a larger label set, we map their embedding vectors to the ICD-9 groups we use in our study. The mapping is done by averaging all subcodes for one group. If no code is available for an ICD-9 group, we use a randomly initialized vector.

A.2 Hyperparameter setup

Batch size Since we work with 1266 labels, the label-wise attention calculations limit the batch size that fits into memory. We therefore use a batch size of 20 for all models without label-wise attention, 10 for label-wise attention models reduced to a dimensionality of 256 and 5 for the others. Initial experiments showed that the batch sizes have no influence on model performance in our experiments, only on memory consumption and training duration.

Learning rates We choose different learning rates for the document encoder weights and the prototype and label-wise attention vectors. Since we expect the encoder weights from the pre-trained Transformer models to be already well aligned with clinical language, we choose a small learning rate between 5e-04 and 5e-06. Since the prototypical diagnosis vectors and the label-wise attention vectors need more adjustments to enable the classification task, we search in a range of 5e-02 and 5e-04. We further apply an AdamW (Loshchilov and Hutter, 2017) optimizer and a linear learning rate scheduler with a warm-up period of 1K to 5K steps. We provide the best hyperparameters per model in the public code repository.

B Interpretability Evaluation Details

B.1 Measuring faithfulness

We use the evaluation setup proposed by Atanasova et al. (2020) to measure the faithfulness of ProtoPatient's explanations. The framework evaluates different methods that output saliencies indicating token importance for a model decision. The evaluation then takes place by masking the most salient tokens via multiple thresholds and measuring the model's performance for each one. Thresholds are

Diagnosis	15 most attended words - with medical relation to diagnosis
Sepsis	1. **hypotension** symptom, 2. **sepsis** descriptor, 3. **fever** symptom, 4. **hypotensive** symptom, 5. **fevers** symptom, 6. **septic** descriptor, 7. **lactate** indicator, 8. **shock** descriptor, 9. **bacteremia** descriptor, 10. **febrile** symptom, 11. **vancomycin** medication, 12. **SBP** risk factor, 13. **levophed** medication, 14. **swelling** symptom, 15. **cirrhosis** risk factor
Intracerebral Hemorrhage	1. **hemorrhage** descriptor, 2. **bleed** descriptor, 3. **headache** symptom, 4. **ICH** descriptor, 5. **IPH** descriptor, 6. **CT** diagnostic, 7. **weakness** symptom, 8. **stroke** descriptor, 9. **brain** descriptor, 10. **intracranial** descriptor, 11. **hemorrhagic** descriptor, 12. **intraventricular** descriptor, 13. **hemorrhages** descriptor, 14. **hemiparesis** symptom, 15. **aphasia** symptom
Pneumonia	1. **pneumonia** descriptor, 2. **cough** symptom, 3. **PNA** descriptor, 4. **COPD** risk factor, 5. **infiltrate** symptom, 6. **distress** complication, 7. **fever** symptom, 8. **breath** *ambiguous*, 9. **hypoxia** symptom, 10. **sputum** symptom, 11. **respiratory** complication, 12. **sepsis** complication, 13. **SOB** symptom, 14. **consolidation** symptom, 15. **CAP** descriptor

Table 6: Words from the test set with the highest attention scores assigned by ProtoPatient. All words are directly related to the diagnoses and mostly describe symptoms or direct descriptors (in various forms). The highlights can therefore help doctors to quickly identify important parts within a note and to compare them to prototypical parts.

going from masking only the top 10% of salient tokens in steps of 10pp until 100% of tokens are masked. The final faithfulness score is then calculated as the area under the curve of model performance over all thresholds. As a performance measure, we choose macro ROC AUC to stay consistent with the rest of our experiments. We compare tokens highlighted by ProtoPatient's label-wise attention vectors to four post-hoc explanation methods as described in 5. We apply these methods to the PubMedBERT baseline, corresponding to a typical post-hoc explanation approach for an otherwise black-box model.

B.2 Finding most relevant words per diagnosis

We want to examine which parts of the clinical notes are highlighted by ProtoPatient per diagnosis. To that end, we collect the tokens with the highest attention scores over all training samples per label. We again use the three diagnoses *sepsis*, *intracerebral hemorrhage* and *pneumonia* for a closer analysis. We further map the tokens to their corresponding words. We then let doctors define the words' medical relations to understand which features the model considers important. Table 6 shows that the most attended words are mainly symptoms or descriptors of the condition at hand, which meets the objective of ProtoPatient to point doctors to relevant parts of a note.

	ROC AUC macro	ROC AUC micro	PR AUC macro
Dimensionality reduction			
ProtoPatient 768	83.56 ±0.17	96.65 ±0.03	14.36 ±0.16
ProtoPatient 256	**86.93** ±0.24	**97.32** ±0.00	**21.16** ±0.21
Transformer vs. Prototypical			
ProtoPatient 768	83.56 ±0.17	**96.65** ±0.03	14.36 ±0.16
PubMedBERT 768 + Label Attention	**84.10** ±0.25	**96.66** ±0.17	**19.74** ±1.27
Label-wise attention			
PubMedBERT 256	83.61 ±0.04	95.76 ±0.05	13.35 ±0.25
PubMedBERT 256 + Label Attention	84.68 ±0.52	96.86 ±0.14	17.15 ±1.52
ProtoPatient 256	**86.93** ±0.24	**97.32** ±0.00	**21.16** ±0.21

Table 7: Full results of our ablation studies. Smaller dimension sizes benefit ProtoPatient, while the effect is less notable on PubMedBERT. Adding label-wise attention, however, increases PubMedBERT results clearly. Overall, the combination of prototypical network, label-wise attention, and reduced dimension in ProtoPatient reaches the best results.

	ROC AUC macro	ROC AUC micro	PR AUC macro
PubMedBERT	82.11 ±0.12	85.48 ±0.64	84.38 ±0.54
PubMedBERT 256 + Label Attention	79.78 ±5.30	83.43 ±4.54	84.70 ±2.84
Prototypical Network	69.65 ±0.22	74.31 ±0.19	78.53 ±0.19
ProtoPatient 768	85.28 ±0.49	88.63 ±0.43	87.78 ±0.10
ProtoPatient	**87.38** ±0.20	**90.63** ±0.23	**89.72** ±0.24
ProtoPatient + Attention Init	86.72 ±1.52	89.84 ±1.16	**89.71** ±1.20

Table 8: Full results of the evaluation on i2b2 data with five classes. Note that the baseline PR AUC is much higher for this task than for the task based on MIMIC-III. ProtoPatient models reach the highest scores, indicating that they are more robust towards changes in text style than the PubMedBERT baselines. The PubMedBERT model with label-wise attention, in particular, shows quite inconsistent results regarding different seeds.