COST-EFFECTIVE INTERACTIVE NEURAL ATTENTION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

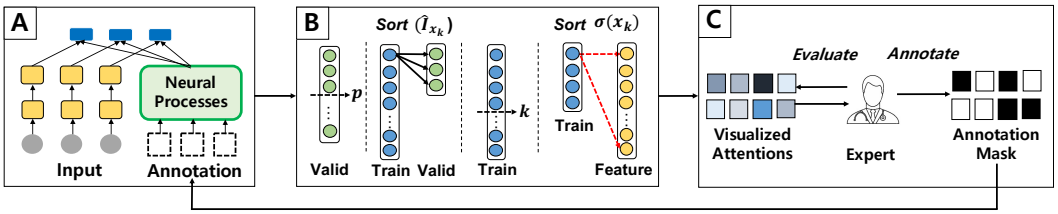
We propose a novel interactive attention learning framework which we refer to as *Interactive Attention Learning (IAL)*, in which the human annotators interactively manipulate the allocated attentions to correct the model's behavior, by updating only the attention-generating model without having to retrain the entire network. For efficient update of the attention generator without retraining, we propose a novel attention mechanism, *Neural Attention Process (NAP)*, which can generate stochastic attentions based on scarce attention-level labels, and can incorporate new training instances without retraining. Further, to minimize human interaction cost, we propose a cost-effective algorithm that selects the most negative training instances that yield *incorrect* and *non-intuitive* interpretation with influence function and re-rank the attentions on the input features by their uncertainties, such that the annotators label the instances and attentions that are more influential to the prediction first. We validate IAL on various datasets from the healthcare and finance domains, on which it significantly outperforms baseline approaches with conventional attention mechanism and random selection of instances when using the same number of annotations, with significantly shorter annotation time per instance owing to attention reranking. Further qualitative analysis shows that IAL also yields interpretations that agree well with human interpretations.

1 INTRODUCTION

Deep neural networks have been the most prevalent tools for predictive modeling tasks nowadays, as they are powerful and can learn complex functions with multiple layers of non-linear transformations without manual engineering of representations. However, the complex nature of the model at the same time makes it very difficult to interpret what they have learned, and brought a new challenge of *interpretability*. Interpreting deep neural networks is crucial to their applications to real-world application domains such as healthcare (Choi et al., 2016; Heo et al., 2018; Ahmad et al., 2018; Sankar et al., 2019), finance (Grath et al., 2018; Wong, 2018), and autonomous driving (Kim & Canny, 2017; Chi & Mu, 2017). For such high-risk tasks, incorrectly learned correlation could result in severe consequences (e.g. mortality, large financial loss, or accidents), and the deployments of such unreliable models could be avoided if the model is interpretable.

Although many recent models proposed diverse solutions to interpretability (Choi et al., 2016; Ahmad et al., 2018; Lage et al., 2018), we face yet another challenge: not all interpretations are correct. Interpretable models should provide *human-understandable* and *intuitive* interpretations (Gilpin et al., 2018; Lage et al., 2018), that conform to the domain knowledge of human experts. For instance, if an interpretable model failed to provide *correct and human-intuitive interpretation* of why it arrived at its decision on diagnosing a patient of heart failure, the final decision maker, the human physician, may not be able to trust the model's decision, even if the actual decision is correct. Often, deep learning models tend to learn tricks that exploit dataset bias, which is another barrier in building trust between humans and machine learning algorithms.

In the cognitive neuroscience perspective, human learning and understanding are built upon two integral parts: *interaction* and *explanation*. Our brain's biological functions are constantly developed by internal reflection (*Back-propagation*) and external explanation (*Human feedback*) during social interactions (Clark et al., 2015). In this sense, interactive learning frameworks could be effective



(A) Neural Attention Process (NAP)

(B) Cost-Effective Selection & Reranking (CESR)

(C) Evaluation & Data-efficient Model update (EDM)

Figure 1: Overview of our Interactive Attention Learning Framework.

means of guiding the model to learn what to focus or ignore, out of training examples and their features, and how to provide desirable explanations for its decision to achieve human-interpretability.

Based on this motivation, we propose an interactive learning framework that allows the users to manipulate the model, by adjusting the provided model interpretations. However, there are several critical challenges that need to be addressed for such an interactive learning framework to be critical. First, to reflect human feedback, the model may need to be retrained, but this is very expensive for deep neural networks. Thus, we need an efficient approach to correct the model's behavior without retraining the entire network. Secondly, requesting and obtaining human feedbacks could be highly expensive (e.g. asking for clinician's annotation), and obtaining feedbacks on already correct interpretations is wasteful. Therefore, we need to selectively provide interpretations on samples and input features that can actually bring in performance improvements of the model. Finally, due to scarcity of human feedbacks, the updated model may overfit, which we need to prevent.

To overcome theses challenges, we propose a novel interactive learning framework which we refer to as *Interactive Attention Learning (IAL)* that can data-efficiently update the trained network without having to train the entire model or overfitting to scarce human feedbacks (See Figure 1). Specifically, we provide the model interpretations in the form of attention allocated on the input variables, and obtain feedbacks from domain experts to correct the model's interpretation by providing a mask on the attention, which is used as supervised labels to update the attention generator without retrainig the main network (Figure 1(c)). However, since the retraining of the attention generator alone could be costly and is prone to overfitting, we propose a novel attention mechanism based on Neural Process, which we refer to as *Neural Attention Process (NAP)*, that can data-efficiently generate attentions with scarce human labels, incorporate additional labeled instances without retraining, and output uncertainty (Figure 1(a)). Further, since obtaining human attention labels for datasets with large number of instances and large number of input variables is costly, we select the examples that has the most negative effect on the generated interpretation using the influence function (Cook & Weisberg, 1980), and then sort the attended input variables by the measured uncertainty in order to interact with the users in the most efficient way (Figure 1(b)).

We validate our interactive attention learning framework on a variety of tasks, including excercise posture correction task, and cerebral infarction risk prediction from electronic health records (EHR), and real-estate price forecast. The experimental results show that our model outperforms the baseline network or naive interactive learning scheme by significant margins with much smaller annotation cost, in terms of number of instances and time to annotate each instance. Further quantitative and qualitative analysis of the learned attention weights shows that our model is able to generate interpretations that align well with the physician's evaluations on the same EHR records.

Our contribution in this paper is threefold:

- We propose a novel interactive attention learning framework with an efficient attention mechanism based on Neural Processes, which enables to efficiently correct the model's understanding with scarce human feedbacks without retraining of the entire network.
- We propose a cost-effective interactive learning algorithm to rank the examples and the attended input variables, in order to maximize the effect of each annotation and thus to minimize the human-machine interaction cost.
- We validate our model on five real-world tasks from three different domains (fitness, healthcare, and finance) for binary, multi-label classification, and regression tasks, and show that our model obtains significant improvements over baselines with less human feedbacks.

2 RELATED WORK

Interpretable machine learning The literature in interpretable machine learning is vast, but we only discuss a few. A popular approach to obtain interpretable model is to build a simple proxy model that mimics the (local) behaviors of a complex model, using either simplified linear models (Ribeiro et al., 2016) or decision trees (Sato & Tsukimoto, 2001; Augasta & Kathirvalavakumar, 2012; Salzberg, 1994). Another approach, specific for neural networks, is analyzing their learned representations (Sharif Razavian et al., 2014; Yosinski et al., 2014) at each unit via visualization. Bau et al. (2017) further consider interpretability of representations in light of their correspondence to semantic concepts, and utilize it for controlling the behavior of generative adversarial networks (Bau et al., 2019). In tihs work, we propose a novel interactive learning framework that can make use of the model's interpretation to iteratively correct the model's understanding, while minimizing the interaction cost via cost-effective instance selection and reranking.

Attention Mechanism Attention mechanism is an effective approach to adaptively select a subset of features (or inputs) in an input-dependent manner, such that the model dynamically focuses on more relevant features for prediction. This mechanism works by input-adaptively generating coefficients for the input features to locate more weights to the features that are found to be relevant for the given input. Attention mechanisms have achieved success with various applications, including image translation (Xu et al., 2015), machine translation (Bahdanau et al., 2015), memory-augmented networks (Sukhbaatar et al., 2015), and visual question answering (Das et al., 2017), and healthcare (Choi et al., 2016; Heo et al., 2018). In this work, we consider attention as a way to both understand what the model has learned and to efficiently correct the model's behavior, using a novel data-efficient attention mechanism based on Neural Process that can generalize well with scarce human labels and can incorporate new labeled instances without retraining,

Neural Processes Neural Processes (NPs) is a neural network-based formulation that combines the benefits of deep neural network and stochastic process, which learns an approximation of a stochastic proces (Garnelo et al., 2018b). NPs allow for global sampling via a latent variable $\mathbf{z}$ to produce different function samples and model the uncertainty for some given context data. (Garnelo et al., 2018a) introduced Conditional Neural Processes (CNPs) which are different from NPs in the sense that CNPs do not sample different functions for the same context points, since it doesn't generate a latent variable for global sampling. Kim et al. (2019) resolves the underfitting problem caused by mean-aggregator, by utilizing the attention mechanism.

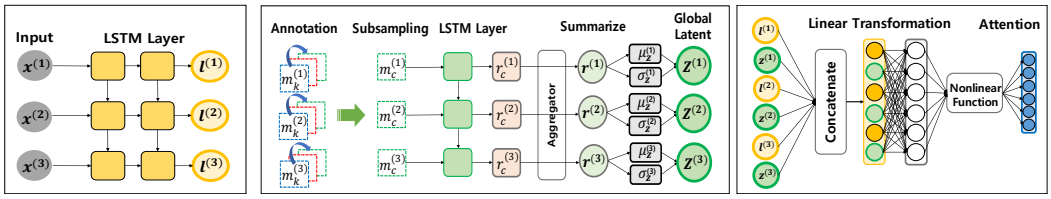
Algorithm 1: Interactive Attention Learning Framework

Input : $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i^{(1:T)}, \mathbf{y}_i\}_{i=1}^N$, $\boldsymbol{\theta} = \{\boldsymbol{\omega}, \boldsymbol{\phi}\}$, S
Output : $\boldsymbol{\theta}$

for $s = 1, ..., S$ **do**
  if $s = 1$ **then**
    Train the network weights $\boldsymbol{\theta}^{(1)}$ by $\text{minimize}_{\boldsymbol{\theta}^{(1)}}\ \mathcal{L}(\boldsymbol{\theta}^{(1)}; \mathcal{D}_{\text{train}}) + \Omega(\boldsymbol{\theta}^{(1)})$
  else
    $(\mathcal{D}_{\text{selected}}^{(s)} = \{\mathbf{x}_k^{(1:T)}, \mathbf{y}_k\}_{k=1}^K, \boldsymbol{\alpha}) \leftarrow \text{CESR}(\boldsymbol{\theta}^{(t-1)})$ ▷ Cost-Effective Selection & Reranking-Algorithm 2
    $\{\mathbf{m}_k\}_{k=1}^K \leftarrow \text{Evaluate}(\mathcal{D}_{\text{selected}}^{(s)}, \boldsymbol{\alpha})$ ▷ Evaluate & get feedback for attention $\boldsymbol{\alpha}$
    $\boldsymbol{\phi}^{(s)} \leftarrow \text{NAP}(\mathbf{x}_s, \{\mathbf{m}_k\}_{k=1}^K, \boldsymbol{\phi})$ ▷ Efficient model update with NAP.
  end
end

3 APPROACH

We now describe our interactive attention learning framework with neural attention process. While our method is generic enough to be applied to any types of prediction tasks, we focus on the case using time-series data. A training dataset is represented as $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i^{(1:T)}, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i^{(1:T)} = [\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(T)}]$ and $\mathbf{x}_i^{(t)} \in \mathbb{R}^D$ is the D-dimensional observation vector for the t-th time step. $\mathbf{y}_i$ is the corresponding label defined according to the task at hands (e.g., a real number for regression). We are interested in building a neural network with two components - an embedding network $\mathbf{f}_{\mathbf{v}}(\cdot; \boldsymbol{\omega})$ embeds $\mathbf{x}_i^{(1:T)}$ into $\mathbf{v}_i^{(1:T)}$, and an attention generating network $\mathbf{f}_{\boldsymbol{\alpha}}(\cdot; \boldsymbol{\phi})$ computes the attention weights. The attention generating networks accelerate the learning by attending to specific parts of the inputs, and more importantly, provides a way to interpret the decision of the model. Thus we are interested in improving the quality of attention by interacting with human experts. More specifically, when the model produces a prediction with an attention, a human supervisor evaluates



(A-1) Embedding Network (A-2) Neural Process (A-3) NP Output Generation

Figure 2: Neural Attention Process (NAP). Embeded outputs $\mathbf{l}$ and the global latent variable $\mathbf{z}$ are generated from the embedding network (A-1) and neural process (A-2), respectively, and the final attention output $\boldsymbol{\alpha}$ is produced by the linear transformation (A-3).

the attention and gives feedback in the form of a binary attention mask $\mathbf{m}$. This attention mask then could be used as supervision to further improve the model by updating the parameter ϕ for $\mathbf{f}_{\boldsymbol{\alpha}}$.

We describe the overall interactive attention learning (IAL) framework in Algorithm 1. We aim to improve the model over S interactions with human supervisors. At the first training stage $(s = 1)$, we initially train the model parameter $\boldsymbol{\theta}^{(1)}$ without the help of human supervisors by minimizing the loss $\mathcal{L}(\boldsymbol{\theta}^{(1)}; \mathcal{D}_{\text{train}}) + \Omega(\boldsymbol{\theta}^{(1)})$ where $\mathcal{L}$ is the loss function for the task and Ω is the regularization function. For each of the following iterations, we select instances and attended input variables, obtain the attention mask $\mathbf{m}$ from the human supervisor, and update the model parameters by minimizing the loss $\mathcal{L}(\boldsymbol{\theta}^{(s)}; \boldsymbol{\theta}^{(s-1)}, \mathbf{m}, \mathcal{D}_{\text{train}}) + \Omega(\boldsymbol{\theta}^{(s)})$.

3.1 Neural Attention Process

In this section, we describe how to effectively update the model with sparse annotations using the variant of neural process (Garnelo et al., 2018b) which we refer to as *Neural Attention Process (NAP)*. Before describing our approach, we briefly explain how attention is applied using RETAIN Choi et al. (2016) as an example. Given the input embedding $\mathbf{v}^{(1:T)}$, attention coefficients for both *timesteps* and *input variables* are constructed as follows:

$$\mathbf{g}^{(1:T)} = \text{RNN}_{\text{time}}(\mathbf{v}^{(1:T)}), \qquad \mathbf{h}^{(1:T)} = \text{RNN}_{\text{var}}(\mathbf{v}^{(1:T)}), \quad (1)$$

$$e^{(t)} = \mathbf{w}_{\text{time}}^{\top}\mathbf{g}^{(t)} + b_{\text{time}} \text{ for } t = 1, \ldots, T, \qquad \mathbf{d}^{(t)} = \mathbf{W}_{\text{var}}\mathbf{h}^{(t)} + \mathbf{b}_{\text{var}} \text{ for } t = 1, \ldots, T, \quad (2)$$

$$\alpha_{\text{time}}^{(1:T)} = \text{Softmax}(e^{(1)}, \ldots, e^{(T)}), \qquad \boldsymbol{\alpha}_{\text{var}}^{(t)} = \tanh(\mathbf{d}^{(t)}) \text{ for } t = 1, \ldots, T, \quad (3)$$

where $\alpha_{\text{time}}^{(1:T)}$ are attention weights applied for time-steps and $\boldsymbol{\alpha}_{\text{var}}^{(1:T)}$ are attention weights for the input variables. We may also consider the stochastic attention as in (Xu et al., 2015). Having $\boldsymbol{\alpha} = \{\alpha_{\text{time}}^{(1:T)}, \boldsymbol{\alpha}_{\text{var}}^{(1:T)}\}$, the model can make predictions as $\hat{\mathbf{y}}_i = \sum_{t=1}^{T} \alpha_{\text{time}}^{(t)} \cdot (\boldsymbol{\alpha}_{\text{var}}^{(t)} \odot \mathbf{v}_i^{(t)})$ where $\odot$ is the element-wise multiplication.

Now we describe the actual algorithm for NAP. Let $\{\mathbf{m}_k^{(1:T)}\}_{k=1}^{K}$ be a set of annotations represented as masks, given for the selected subsamples of the training data. The idea is that, instead of updating the parameter ϕ using these small number of examples, we let network take the *summarization* of the annotation set as an additional input. This approach, when trained properly, can automatically adapt without retraining when a new set of annotations is further given. The overall pipeline of neural attention process is depicted in Figure 2.

Embedding the inputs (A-1) we first embed the input $\mathbf{x}^{(1:T)}$ using LSTM (Hochreiter & Schmidhuber, 1997) into $\mathbf{l}^{(1:T)} = [\mathbf{g}^{(1:T)}, \mathbf{h}^{(1:T)}]$.

Embedding & summarizing the annotations (A-2) Given the set of annotation masks $\{\mathbf{m}_k^{(1:T)}\}_{k=1}^{K}$, we build an intermediate representation $\{\mathbf{r}_k^{(1:T)}\}_{k=1}^{K}$ via another LSTM. Then, for each time step, we build a summarized representation $\bar{\mathbf{r}}^{(t)}$ by a permutation-invariant operation (for instance, average),

$$\bar{\mathbf{r}}^{(t)} = \mathbf{r}_1^{(t)} \oplus \cdots \oplus \mathbf{r}_K^{(t)}. \quad (4)$$

Having $\bar{\mathbf{r}}^{(1:T)}$, we define a distribution for the summary variable $\mathbf{z}$ as Gaussian:

$$\mathbf{z}^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}(\bar{\mathbf{r}}^{(t)}), \boldsymbol{\sigma}^2(\bar{\mathbf{r}}^{(t)})), \;\; \boldsymbol{\mu}(\bar{\mathbf{r}}^{(t)}) = \mathbf{W}_{\boldsymbol{\mu}}\bar{\mathbf{r}}^{(t)} + \mathbf{b}_{\boldsymbol{\mu}}, \;\; \boldsymbol{\sigma}(\bar{\mathbf{r}}^{(t)}) = \text{softplus}(\mathbf{W}_{\boldsymbol{\sigma}}\bar{\mathbf{r}}^{(t)} + \mathbf{b}_{\boldsymbol{\sigma}}). \quad (5)$$

Generating attentions (A-3) Now we generate the attention by a similar procedure to (3), but instead of feeding only, $\mathbf{l}^{(1:T)} = (\mathbf{g}^{(1:T)}, \mathbf{h}^{(1:T)})$, we feed both $\mathbf{l}^{(1:T)}$ and the annotation summarization vector $\mathbf{z}^{(1:T)}$ by concatenation. This allows the network to naturally reflect the information obtained from the summarization $\mathbf{z}^{(1:T)}$ without having to retrain the whole attention network parameter ϕ.

Algorithm 2: Cost-Effective Selection & Reranking (CESR)

Input : $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i^{(1:T)}, \mathbf{y}_i\}_{i=1}^N$, $\mathcal{D}_{\text{valid}} = \{\mathbf{x}_j^{(1:T)}, \mathbf{y}_j\}_{j=1}^M$, $P, K, \boldsymbol{\theta}^{(s-1)}$.

Output : $\mathcal{D}_{\text{selected}}^{(s)} = \{\mathbf{x}_k^{(1:T)}, \mathbf{y}_k\}_{k=1}^K$, $\boldsymbol{\alpha}$.

Evaluate the network with $\mathcal{D}_{\text{valid}}$.
Sort valid points $\{(\mathbf{x}_j^{(1:T)}, \mathbf{y}_j)\}_{j=1}^M$ in the descending order of $\mathcal{L}(\mathbf{x}_j^{(1:T)}, \mathbf{y}_j; \boldsymbol{\theta}^{(s-1)})$.
Select top P-valid points $\mathcal{D}'_{\text{valid}} = \{\mathbf{x}_p^{(1:T)}, \mathbf{y}_p\}_{p=1}^P$
for $i = 1, ..., N$ **do**
  Approximate influence of train points $(\mathbf{x}_i^{(1:T)}, \mathbf{y}_i)$ on loss at $\mathcal{D}'_{\text{valid}}$.
end
Select top K-training points $\mathcal{D}_{\text{selected}}^{(s)} = \{\mathbf{x}_k, \mathbf{y}_k\}_{k=1}^K$ w.r.t. the approximated influence.
Compute the attention weights $\boldsymbol{\alpha}$ for $\mathcal{D}_{\text{selected}}^{(s)}$.
Approximate uncertainty of attention weights $\text{Var}(\boldsymbol{\alpha})$ for $\mathcal{D}_{\text{selected}}^{(s)}$ via *Monte-Carlo sampling*.

Training NAP Even with NAP, we need at least one training procedure to update ϕ so that it can take $\mathbf{z}^{(1:T)}$ as an additional input. For this training, we use two strategies to make the NAP to readily generalize to the future annotations to be given. First, at each training step, we randomly subsample the annotations to comprise random task to train the model. This prevents the model from completely over-fitted to the entire annotation set $\{\mathbf{m}_k^{(1:T)}\}_{k=1}^K$. Secondly, we regularize the summarization vector $\mathbf{z}^{(1:T)}$ by positing a prior distribution. We then train the ELBO in similar fashion to the original neural process objectvie (Garnelo et al., 2018b).

3.2 Cost-Effective Selection and Re-ranking

Acquiring human annotations is highly expensive and thus it is crucial to prioritize the most important instances and attentions (input variables) that negatively affect the model accuracy. The most native approach to identify important negative examples in the training set, would be to retrain the entire model parameter $\boldsymbol{\theta}$ every time, while omitting a single targeted point. To avoid such costly retraining, we utilize influence functions (Cook & Weisberg, 1980) and uncertainty to perform negative instance selection and feature re-ranking respectively in a cost-effective manner.

Influence functions efficiently estimate the effect of removing particular train points on a model without retraining (Koh & Liang, 2017). For notational simplicity, we set train points $\mathbf{s}_1, ..., \mathbf{s}_N$, where $\mathbf{s}_i = (\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_{\text{train}}$[1]. Given a scenario where we approximate the change of the model's predictions by removing a data point $\mathbf{s}$, we can formalize the change as $\hat{\boldsymbol{\theta}}_{-\mathbf{s}} - \hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\boldsymbol{\theta}, \mathbf{s}_i)$ and $\hat{\boldsymbol{\theta}}_{-\mathbf{s}} = \text{argmin}_{\boldsymbol{\theta}} \frac{1}{n-1} \sum_{\mathbf{s}_i \neq \mathbf{s}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{s}_i)$. That is, $\hat{\boldsymbol{\theta}}$ is the parameter that minimizes the empirical risk of the full dataset and $\hat{\boldsymbol{\theta}}_{-\mathbf{s}}$ is the empirical minimizer of the dataset without the train point $\mathbf{s}$. Since upweighting $\mathbf{s}_i$ with $\delta = -\frac{1}{n}$ has the same effect as removing $\mathbf{s}_i$ from the train set (Koh & Liang, 2017), we approximate the influence by upweighting $\mathbf{s}_i$ on the loss at all valid points as follows:

$$\tilde{\mathbf{I}}_{\text{up, loss}}(\mathbf{s}_i, \mathbf{s}^{\text{valid}}) = \sum_{p=1}^P \left| \tilde{\mathbf{I}}_{\text{up, loss}}(\mathbf{s}_i, \mathbf{s}_p^{\text{valid}}) \right| = \sum_{p=1}^P \left| -\nabla_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}}, \mathbf{s}_p^{\text{valid}})^\top H_{\hat{\boldsymbol{\theta}}}^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}}, \mathbf{s}_i) \right| \tag{6}$$

where $H_{\hat{\boldsymbol{\theta}}} = \frac{1}{n} \sum_{i=1}^n \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}(\hat{\boldsymbol{\theta}}, \mathbf{s}_i)$ is the Hessian matrix to compute the second partial derivative of the function as a quadratic approximation with respect to the empirical loss around $\hat{\boldsymbol{\theta}}$ and we set the validation data points $\mathbf{s}_p^{valid} = \{(\mathbf{x}_p, \mathbf{y}_p) | \mathbf{x}_p \in \mathbb{R}^d, \mathbf{y}_p \in \mathbb{R}^1\}_{p=1}^P$ and $\tilde{\mathbf{I}}_{\text{up, loss}}(\mathbf{s}_i, \mathbf{s}^{\text{valid}})$ represents the influence of a train point $\mathbf{s}_i$ on the the sum of loss at all valid points $\{\mathbf{s}_p\}_{p=1}^P$.

Cost-Effective Selection & **Reranking (CESR)** We now describe the procedure illustrated in (B)-Figure 1, in detail. (1) We train the entire network with a training set, sort valid points in descending order of its loss, and select top P-valid points. (2) We then compute influence scores for training points on the loss at P-valid points $\mathcal{D}'_{\text{valid}} = \{\mathbf{x}_p, \mathbf{y}_p\}_{p=1}^P$, and (3) sort the influence scores of training points in descending order of $\tilde{\mathbf{I}}_{\text{up, loss}}(\mathbf{s}_i, \mathbf{s}^{\text{valid}})$ and select top K-train points denoted as $\mathcal{D}_{\text{selected}} = \{\mathbf{x}_k, \mathbf{y}_k\}_{k=1}^K$. (4) Finally, using the approximate uncertainty of each variable's attention $\text{Var}(\boldsymbol{\alpha}(x_s^d))$ with *Monte Carlo dropout* (Gal & Ghahramani, 2016), we sort the set of variables in

[1] We omit the time index in this section.

descending order of the approximated uncertainty. One important aspect of CESR is that the identified set of valid points with the highest losses at evaluation can be effectively used as means for detecting negative train points via influence functions. Algorithm 2 describes the entire procedure.

3.3 Attention Evaluation and Efficient model update

Figure 3: Attention Annotation Interface (Healthcare).

Given the provided set of selected negative train instances from CESR algorithm, (1) we present the visualizations of attentions on our online interactive user interface in Figure 3.3, where the annotators can easily turn on/off the binary mask $\mathbf{m}_k = \{0, 1\}$. The interface visually emphasizes features that the annotator should pay attention to the bar plots or attention map depeding on the given task (see the **appendix** for more information). (2) Annotators examine attention weights to determine whether attention weights are incorrectly allocated, simultaneously checking for the corresponding input values and label. With time-series data, annotators evaluate the delivered attentions $\alpha_{\text{time}}^{(1:T)}$ and $\boldsymbol{\alpha}_{\text{var}}^{(1:T)}$ via an annotation matrix, which are basically matrices of binarized attentions. (3) Accumulated annotations are efficiently updated to NAP without retraining.

4 Experiments

We validate the performance and cost-effectiveness of our interactive neural attention learning, on five datasets from three domains.

1) Fitness - Squat Pose Correction This dataset contains 4,000 video frames of human subject performing squats with 11 multi-labels classification task (e.g., 0: Correct posture, 1: Incorrect-*exaggerated knees-forward movement*, 2: Incorrect-*sitting on the thighs instead of between them*). We extract 14 pairs of key points from joints (e.g., *left shoulder* or *right ankle*) to have a clear picture of which body joints an attentional network attends to for a given instance. The task is to classify whether a person performs the correct posture or 10 types of incorrect posture.

2) Medical Check-ups These datasets are subsets of the electronic health records database of a major hospital, consisting of medical check-up records from 2009 to 2012 (4 time-steps) for patients over age 15 in out-patient units, including around 1.5 million records. We extracted 245,000 patient records from this database, in which each record contains 34 variables including general patient information (e.g., sex and height), vital signs (e.g., systolic pressure and hemoglobin level), and risk-inducing behaviors(e.g., alcohol consumption). The task is to predict the onset of the following disease in one year: 1) *Heart Failure*, 2) *Cerebral Infarction*, 3) *Cardiovascular Disease* (CVD).

3) Realestate Sales Transactions This datasets is a subset of residential sales transaction database from New York City Department of Finance consisting of approximately 70,700 house records with 27,000 sales transaction records over 15 years from 2011 to 2018 (8 time-steps). The dataset includes 3100 housing transactions and each record includes 47 variables (e.g., housing characteristics (number of bathrooms or bed rooms) and macro economic index (interest rates, GDP price index). The task is to forecast a residential property price in NYC in one year.

For all datasets, we generate train/valid/test with the ratio of 70%:10%:20%. For more details on the datasets, network configurations, and hyperparameters used, please **see supplementary file**.

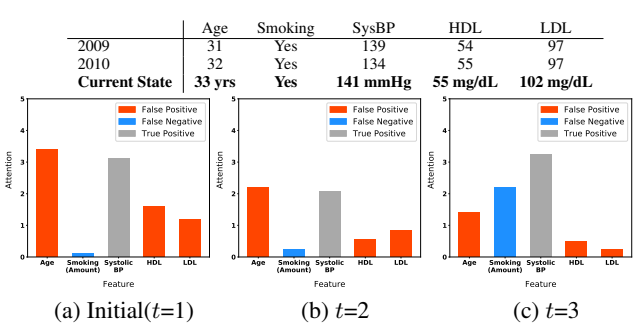
Baselines We now describe the baselines and our models.
1) [RETAIN]: The attentional recurrent neural network model (RETAIN) in (Choi et al., 2016).
2) [RETAIN] Random Selection: RETAIN, but newly trained from a train set with k-train points omitted: k stands for the number of randomly selected train points from the original train set.
3) [RETAIN] IF Selection: RETAIN, but newly trained from a train set with top k-negative points omitted: k is the number of negative points selected via influence function (Koh & Liang, 2017).
4) [IAL] UA-Random: Uncertainty-Aware attentional network (UA) in the interactive attention learning setting. k-data points are randomly selected, delivered to annotators for attention evaluation, and, during retraining, the set of parameters in the attention network are selectively fine-tuned by element-wise multiplication of binary annotation masks ($\mathbf{m}_k$) with attention parameters ($\mathbf{m}_k \odot \alpha$).
5) [IAL] NAP-Random: Our Neural Attention Process model (NAP), in the IAL-Random setting (4). At each iteration, NAP is updated with binary annotation masks without retraining.
6) [IAL+Cost-Effective] Attention Cross Entropy (ACE): Retrain the attention network using a *binary cross entropy loss function* between the attention weight vector $\boldsymbol{\alpha}$ and the attention annotation

		EHR			Squat	Realestate
		Heart Failure	Cerebral Infarction	CVD	Posture	Forecasting
Baselines (RETAIN)	RETAIN	0.7921 ± 0.01	0.6168 ± 0.01	0.6164 ± 0.02	0.8425 ± 0.03	0.2522 ± 0.01
	Random Selection	0.7852 ± 0.02	0.6116 ± 0.02	0.5671 ± 0.01	0.8221 ± 0.05	0.2540 ± 0.01
	IF Selection	0.7984 ± 0.03	0.6182 ± 0.02	0.5882 ± 0.02	0.8363 ± 0.03	0.2434 ± 0.01
IAL (Random Selection)	UA-Random	0.7824 ± 0.01	0.6191 ± 0.01	0.6012 ± 0.02	0.8512 ± 0.00	0.2632 ± 0.02
	NAP-Random	0.8015 ± 0.02	0.6287 ± 0.03	0.6132 ± 0.02	0.8525 ± 0.01	0.2511 ± 0.01
IAL (Cost -effective)	ACE	0.7982 ± 0.04	0.5992 ± 0.03	0.6193 ± 0.02	0.8450 ± 0.03	0.2519 ± 0.01
	NAP-Selective	**0.8157** ± 0.01	**0.6374** ± 0.01	**0.6304** ± 0.02	**0.8562** ± 0.01	**0.2381** ± 0.01

Table 1: The multi-class classification performance on the four electronic health records datasets and one fitness dataset. The reported numbers are mean-AUROC for EHR and mean-Accuracy for squat. In the realestate price forecast, the number indicates mean-**percentage error**, meaning a lower error indicates better performance.

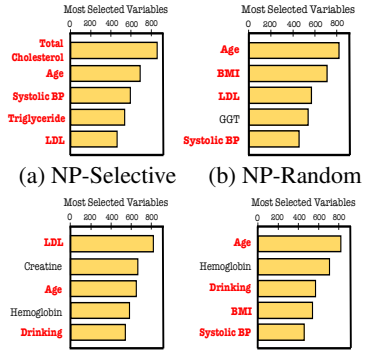
		EHR			Squat
		Heart Failure	Cerebral Infarction	CVD	Posture
NPA	Random Order	0.8082 ± 0.01	0.6274 ± 0.01	0.6224 ± 0.02	0.8519 ± 0.01
	Selective	**0.8159** ± 0.01	**0.6311** ± 0.02	**0.6302** ± 0.01	**0.8538** ± 0.01

Table 2: Accuracy and percentage error on NPA-Selective and NPA-Random-order, in which veriables are randomly ordered. Only top 30% of variables were selected to the annotators.



(a) Initial(t=1) (b) t=2 (c) t=3

Figure 4: Visualization of attention for a selected patient on the Cardio Vascular Disease (CVD) prediction task. Contribution indicates the extent to which each individual feature affects the onset of CVD in 1 year. **Age** - Age, **Smoking** - Whether the patient currently smokes, **SysBP** - Systolic blood pressure, **HDL** - High-density lipoproteins cholesterol, **LDL** - Low-density lipoprotein cholesterol. Bars correspond to attention weights.



(a) NP-Selective (b) NP-Random

(a) IF-Selection (b) ACE

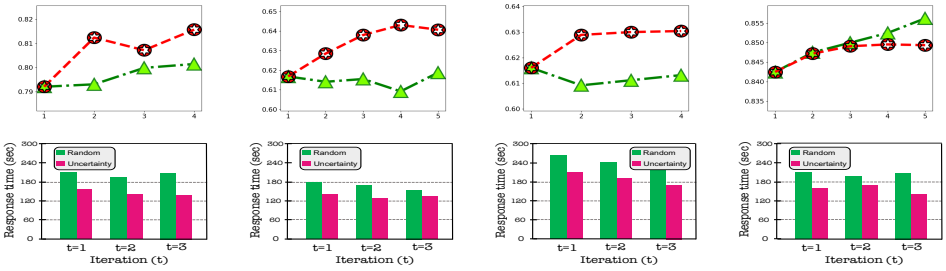
Figure 5: Top 5 variables that are often selected the most by NP-Random, NP-Selective, IF-Selection, and Attention Cross Entropy (ACE) on the CVD task. Variables in **red** stands for important key factor determined by physicians.

$\mathbf{m}_k$ in generating attention weights, which adopts a similar approach with (Choi et al., 2019).
7) [IAL+Cost-Effective] NAP-Selective Our IAL framework with Neural Attention Process (NPA).

4.1 Evaluation of the Final Model Performance

We first examine the prediction accuracy of baselines and our model. Table 1 contains the accuracies of baselines and our model measured with *Area Under the ROC curve (AUROC)* on the risk prediction tasks, *accuracy* on squat posture task with multi-labels, and *percentage error* on real-estate price forecast. Note that IF Selection, which uses influence functions to remove instances with negative influence scores, performs relatively better on most tasks than other RETAIN baselines, but fails to improve on CVD and squat posture task. We observe that UA-Random, whose newly learned knowledge of the human annotations from randomly selected data points, performs worse than NAP-Random on all tasks, which is caused by *overfitting* to a particular example with retraining. Note that NAP-Selective works significantly better than NAP-Random for all tasks, which shows that the effect of attention annotation process cannot have much effect on the model when the instances are *randomly selected*. The Attention Cross Entropy model, which retrained with the binary cross entropy loss function between attention weights and annotation masks, performs worse than NAP-Selective, caused by severe overfitting with scarce annotation samples.

Interpretability and Accuracy of Generated Attentions We further qualitatively analyze the contribution of each feature for a CVD patient (label=1) whose records showed significant changes in attention with the help of physicians in Figure 4. The table (4.1) shows the patient's medical records at the previous (2009, 2010) and the current time-step (2011), yearly registered records. The three graphs show the values of the allocated attentions across three iterations. Our model, NAP-Selective



(a) Heart Failure (b) Cerebral Infarction (c) CVD (f) Squat Posture

Figure 6: (**Top**) Change of accuracy with 40 annotations for each iterations(s) on the all tasks between NAP-Selective (Red) **vs** NAP-Random (Green). (**Bottom**) Mean Response Time (mean-RT) of annotators to evaluate one data sample (Being prioritized by uncertainty **vs** Randomly ordered).

failed to predict the label at s=1 (a), but make a correct prediction at s=1 (c). We visualized five variables that have clinically meaningful changes. Across the change of attentions from (a) to (c), the physicians consider that attentions on age, HDL, and LDL in (a) are *false positive* and smoking as *false negative*, except SysBP as *true positive*. Noting that the patient's age (30) is younger than the median age (50 years-old) of female CVD patient (Garcia et al., 2016), initial NAP-Selective (a) allocated too much weights on age, which led to an overconfident attention model and in turn resulted in the incorrect prediction. However, our model gradually allocated less weights on age over iteration, as it started to learn *what to attend to* from physicians on a disjoint set of training data. Note that attention on smoking highly increased at s=3 (c), which is also clinically guided by a physician for the reason that CVD risk increases by 25% for women who smoke cigarettes (Huxley & Woodward, 2011).

In Figure 5, each bar graph corresponds to top 5 feature variables that are selected most often by each method on the CVD task with 810 EHR records. Interestingly, all variables that NAP-Selective attends to the most are interpreted by physicians as key risk factors for accessing a CVD patient. Although NP-Random failed to ignore GGT which is a relatively a less important variable, it accesses the key variable better than other models with IF-Selection. For broad clinical descriptions for figure 5, please **see supplementary file**.

4.2 Evaluation of the Cost-Effective Selection and Re-ranking

The line graphs in Figure 6 (**top**) shows the change in model accuracy over iterations, with NP-Random and NP-Selective. On the risk prediction and posture estimation tasks, the accuracy of NP-Selective increases over the rounds of interaction, while NP-Random achieves marginal increases only on heart failures and CVD tasks, and degenerates accuracy on others. We further measure the average response time of the annotators with and without reranking of the attented inputs. The bar graphs (**bottom**) show that annotators spend less time with annotation if variables are prioritized by its uncertainty using our uncertainty-based reranking method (red bars), compared to presenting them in the original order (green bars), on all tasks. We further analyze the effect of the re-ranking algorithm by comparing top-10 input variables selected by our re-ranking algorithm, with the top-10 variables whose attentions are corrected by the annotators without variable re-ranking. In Figure 7, We see that 7 out of 10 input variables that are most often selected by our algorithm (colored in red) corresponded to those selected by human annotators, which further demonstrates the effectiveness of the method.

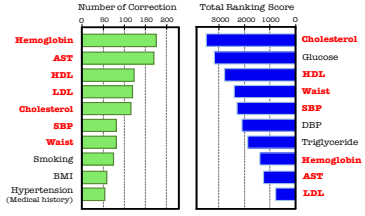


Figure 7: Top 10 variables (Green) ranked by the total number of being checked by physicians and top 10 variables (Blue) ranked by attention uncertainty in the CVD task.

5 Conclusion

We proposed an interactive learning framework through the attention generated by the model, using a novel attention mechanism based on Neural Process that efficiently enables correction of a model's interpretability with scarce human feedback without retraining the entire network. Further, we propose a cost-effective instance selection and attention re-ranking algorithm to minimize the human-machine interaction cost while maximizing its effect. We validated our model on five real-world tasks from the healthcare, fitness, and finance domains and showed that our model significantly outperforms the baselines with fewer annotation cost in terms of the number of training instances and annotation time, while generating more human-interpretable attentions.

REFERENCES

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 559–560. ACM, 2018.

Thomas Almdal, Henrik Scharling, Jan Skov Jensen, and Henrik Vestergaard. The independent effect of type 2 diabetes mellitus on ischemic heart disease, stroke, and death: a population-based study of 13 000 men and women with 20 years of follow-up. *Archives of internal medicine*, 164(13): 1422–1426, 2004.

Keaven M Anderson, Patricia M Odell, Peter WF Wilson, and William B Kannel. Cardiovascular disease risk profiles. *American heart journal*, 121(1):293–298, 1991.

M Gethsiyal Augasta and Thangairulappan Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*, 35(2):131–150, 2012.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

Hailey R Banack and Jay S Kaufman. The "obesity paradox" explained. *Epidemiology*, 24(3): 461–462, 2013.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1901.09887*, 2019.

Eugenia E Calle, Michael J Thun, Jennifer M Petrelli, Carmen Rodriguez, and Clark W Heath Jr. Body-mass index and mortality in a prospective cohort of us adults. *New England Journal of Medicine*, 341(15):1097–1105, 1999.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

Lu Chi and Yadong Mu. Deep steering: Learning end-to-end driving model from spatial and temporal visual cues. *arXiv preprint arXiv:1708.03798*, 2017.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*. 2016.

Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 230. ACM, 2019.

Clark, Ian, and Guillaume Dumas. Toward a neural basis for peer-interaction: what makes peer-learning tick? *Frontiers in psychology*, 2015.

InterAct Consortium et al. The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: the epic-interact study. *Diabetologia*, 56(1): 60–69, 2013.

R Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

John R Downs, Michael Clearfield, Stephen Weis, Edwin Whitney, Deborah R Shapiro, Polly A Beere, Alexandra Langendorfer, Evan A Stein, William Kruyer, Antonio M Gotto Jr, et al. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of afcaps/texcaps. *Jama*, 279(20):1615–1622, 1998.

Robert H Eckel, David A York, Stephan Rossner, Van Hubbard, Ian Caterson, Sachiko T St. Jeor, Laura L Hayman, Rebecca M Mullis, and Steven N Blair. Prevention conference vii: Obesity, a worldwide epidemic related to heart disease and stroke: executive summary. *Circulation*, 110(18): 2968–2975, 2004.

Justin A Ezekowitz, Sharon E Straus, Sumit R Majumdar, and Finlay A McAlister. Stroke: strategies for primary prevention. *American family physician*, 68(12):2379–2386, 2003.

JoAnne Micale Foody, Christopher R Cole, Eugene H Blackstone, and Michael S Lauer. A propensity analysis of cigarette smoking and mortality with consideration of the effects of alcohol. *The American journal of cardiology*, 87(6):706–711, 2001.

Caroline S Fox, Helen Parise, Ralph B DAgostino Sr, Donald M Lloyd-Jones, Ramachandran S Vasan, Thomas J Wang, Daniel Levy, Philip A Wolf, and Emelia J Benjamin. Parental atrial fibrillation as a risk factor for atrial fibrillation in offspring. *Jama*, 291(23):2851–2855, 2004.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Mariana Garcia, Sharon L Mulvagh, C Noel Bairey Merz, Julie E Buring, and JoAnn E Manson. Cardiovascular disease in women: clinical perspectives. *Circulation research*, 118(8):1273–1293, 2016.

Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional neural processes. *CoRR*, abs/1807.01613, 2018a. URL `http://arxiv.org/abs/1807.01613`.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes. *CoRR*, abs/1807.01622, 2018b. URL `http://arxiv.org/abs/1807.01622`.

K Gemes, I Janszky, LE Laugsand, KD Laszlo, S Ahnve, LJ Vatten, and KJ Mukamal. Alcohol consumption is associated with a lower incidence of acute myocardial infarction: results from a large prospective population-based study in norway. *Journal of internal medicine*, 279(4):365–375, 2016.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. IEEE, 2018.

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245*, 2018.

Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. In *Advances in Neural Information Processing Systems*, pp. 909–918, 2018.

Matti Hillbom, Heikki Numminen, and Seppo Juvela. Recent heavy drinking of alcohol and embolic stroke. *Stroke*, 30(11):2307–2312, 1999.

George Hindy, Gunnar Engstrom, Susanna C Larsson, Matthew Traylor, Hugh S Markus, Olle Melander, and Marju Orho-Melander. Role of blood lipids in the development of ischemic stroke and its subtypes: a mendelian randomization study. *Stroke*, 49(4):820–827, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short term memory. *Neural Computation*, 9: 1735–1780, 1997.

Rachel R Huxley and Mark Woodward. Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies. *The Lancet*, 378(9799):1297–1305, 2011.

Sun Ha Jee, Il Suh, Il Soon Kim, and Lawrence J Appel. Smoking and atherosclerotic cardiovascular disease in men with low levels of serum cholesterol: the korea medical insurance corporation study. *Jama*, 282(22):2149–2155, 1999.

Paula Jerrard-Dunne, Geoffrey Cloud, Ahamad Hassan, and Hugh S Markus. Evaluating the genetic component of ischemic stroke subtypes: a family history study. *Stroke*, 34(6):1364–1369, 2003.

Henrik Stig Jorgensen, Hirofumi Nakayama, TS Olsen, and HO Raaschou. Effect of blood pressure and diabetes on stroke in progression. *The Lancet*, 344(8916):156–159, 1994.

Pekka Jousilahti, Erkki Vartiainen, Jaakko Tuomilehto, and Pekka Puska. Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14 786 middle-aged men and women in finland. *Circulation*, 99(9):1165–1172, 1999.

WB Kannel and DL McGee. Diabetes and glucose tolerance as risk factors for cardiovascular disease: the framingham study. *Diabetes care*, 2(2):120–126, 1979.

Ichiro Kawachi, Graham A Colditz, Meir J Stampfer, Walter C Willett, JoAnn E Manson, Bernard Rosner, Frank E Speizer, and Charles H Hennekens. Smoking cessation and decreased risk of stroke in women. *Jama*, 269(2):232–236, 1993.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, S. M. Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *CoRR*, abs/1901.05761, 2019. URL `http://arxiv.org/abs/1901.05761`.

Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1885–1894. JMLR. org, 2017.

Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. In *Advances in Neural Information Processing Systems*, pp. 10159–10168, 2018.

Chong Do Lee, Aaron R Folsom, and Steven N Blair. Physical activity and stroke risk: a meta-analysis. *Stroke*, 34(10):2475–2481, 2003.

Duanping Liao, Richard Myers, Steven Hunt, Eyal Shahar, Catherine Paton, Gregory Burke, Michael Province, and Gerardo Heiss. Familial history of stroke and stroke risk: the family heart study. *Stroke*, 28(10):1908–1912, 1997.

Lynn P Lowe, Philip Greenland, Karen J Ruth, Alan R Dyer, Rose Stamler, and Jeremiah Stamler. Impact of major cardiovascular disease risk factors, particularly in combination, on 22-year mortality in women and men. *Archives of internal medicine*, 158(18):2007–2014, 1998.

James B Meigs, L Adrienne Cupples, and PW Wilson. Parental transmission of type 2 diabetes: the framingham offspring study. *Diabetes*, 49(12):2201–2207, 2000.

Robert M Najarian, Lisa M Sullivan, William B Kannel, Peter WF Wilson, Ralph B D'Agostino, and Philip A Wolf. Metabolic syndrome compared with type 2 diabetes mellitus as a risk factor for stroke: the framingham offspring study. *Archives of internal medicine*, 166(1):106–111, 2006.

Hirofumi Nakayama, HS Jorgensen, HO Raaschou, and Tom Skyhøj Olsen. The influence of age on stroke outcome. the copenhagen stroke study. *Stroke*, 25(4):808–813, 1994.

Antonia C Novello. Surgeon general's report on the health benefits of smoking cessation. *Public Health Reports*, 105(6):545, 1990.

Jaideep Patel, Mahmoud Al Rifai, Maren T Scheuner, Steven Shea, Roger S Blumenthal, Khurram Nasir, Michael J Blaha, and John W McEvoy. Basic vs more complex definitions of family history in the prediction of coronary heart disease: the multi-ethnic study of atherosclerosis. In *Mayo Clinic Proceedings*, volume 93, pp. 1213–1223. Elsevier, 2018.

Michael J Pencina, Ann Marie Navar, Daniel Wojdyla, Robert J Sanchez, Irfan Khan, Joseph Elassal, Ralph B D'Agostino Sr, Eric D Peterson, and Allan D Sniderman. Quantifying importance of major risk factors for coronary heart disease. *Circulation*, 139(13):1603–1611, 2019.

Kathleen Potempa, Martita Lopez, Lynne T Braun, J Peter Szidon, Louis Fogg, and Tyler Tincknell. Physiological outcomes of aerobic exercise training in hemiparetic stroke patients. *Stroke*, 26(1): 101–105, 1995.

Kenneth E Powell, Paul D Thompson, Carl J Caspersen, and Juliette S Kendrick. Physical activity and the incidence of coronary heart disease. *Annual review of public health*, 8(1):253–287, 1987.

Qing Qiao, M Tervahauta, A Nissinen, and J Tuomilehto. Mortality from all causes and from coronary heart disease related to smoking and changes in smoking during a 35-year follow-up of middle-aged finnish men. *European heart journal*, 21(19):1621–1626, 2000.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL `http://doi.acm.org/10.1145/2939672.2939778`.

Paul M Ridker, Robert J Glynn, and Charles H Hennekens. C-reactive protein adds to the predictive value of total and hdl cholesterol in determining risk of first myocardial infarction. *Circulation*, 97 (20):2007–2011, 1998.

Michael Roerecke and Jurgen Rehm. Irregular heavy drinking occasions and risk of ischemic heart disease: a systematic review and meta-analysis. *American journal of epidemiology*, 171(6): 633–644, 2010.

Geoffrey Rose, PJ Hamilton, L Colwell, and MJ Shipley. A randomised controlled trial of anti-smoking advice: 10-year results. *Journal of Epidemiology & Community Health*, 36(2):102–108, 1982.

Meaghan Roy-O Reilly and Louise D McCullough. Age and sex are critical factors in ischemic stroke pathology. *Endocrinology*, 159(8):3120–3131, 2018.

Jean-Bernard Ruidavets, Pierre Ducimetiere, Alun Evans, Michele Montaye, Bernadette Haas, Annie Bingham, John Yarnell, Philippe Amouyel, Dominique Arveiler, Frank Kee, et al. Patterns of alcohol consumption and ischaemic heart disease in culturally divergent countries: the prospective epidemiological study of myocardial infarction (prime). *Bmj*, 341:c6077, 2010.

Steven L Salzberg. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3):235–240, 1994.

Vignesh Sankar, Devinder Kumar, David A Clausi, Graham W Taylor, and Alexander Wong. Sisc: End-to-end interpretable discovery radiomics-driven lung cancer prediction via stacked interpretable sequencing cells. *arXiv preprint arXiv:1901.04641*, 2019.

Makoto Sato and Hiroshi Tsukimoto. Rule extraction from neural networks via decision tree induction. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 3, pp. 1870–1875. IEEE, 2001.

Reena S Shah and John W Cole. Smoking and stroke: the more you smoke the more you stroke. *Expert review of cardiovascular therapy*, 8(7):917–932, 2010.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

Korea Stroke Society. *National clinical guideline for stroke*. Korea Stroke Society, 2016.

Jan A Staessen, Jiguang Wang, Giuseppe Bianchi, and Willem H Birkenhager. Essential hypertension. *The Lancet*, 361(9369):1629–1641, 2003.

Jeffrey D Stanaway, Ashkan Afshin, Emmanuela Gakidou, Stephen S Lim, Degu Abate, Kalkidan Hassen Abate, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159): 1923–1994, 2018.

Neil J Stone, Jennifer G Robinson, Alice H Lichtenstein, C Noel Bairey Merz, Conrad B Blum, Robert H Eckel, Anne C Goldberg, David Gordon, Daniel Levy, Donald M Lloyd-Jones, et al. 2013 acc/aha guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Part B):2889–2934, 2014.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *NIPS*, 2015.

Konstantinos Vemmos, George Ntaios, Konstantinos Spengos, Paraskevi Savvari, Anastasia Vemmou, Theodora Pappa, Efstathios Manios, George Georgiopoulos, and Maria Alevizaki. Association between obesity and mortality after acute first-ever stroke: the obesity–stroke paradox. *Stroke*, 42 (1):30–36, 2011.

Nae-Yuh Wang, J Hunter Young, Lucy A Meoni, Daniel E Ford, Thomas P Erlinger, and Michael J Klag. Blood pressure change and risk of hypertension associated with parental hypertension: the johns hopkins precursors study. *Archives of internal medicine*, 168(6):643–648, 2008.

Alexander Wong. *Waterloo, Canada N2L 3G1 Phone: 519-888-4567 ext. 31299 a28wong@ engmail. uwaterloo. ca*. PhD thesis, University of Waterloo, 2018.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328, 2014.

Salim Yusuf, Steven Hawken, Stephanie Ounpuu, Tony Dans, Alvaro Avezum, Fernando Lanas, Matthew McQueen, Andrzej Budaj, Prem Pais, John Varigos, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): case-control study. *The lancet*, 364(9438):937–952, 2004.

Ivana Zavaroni, Enzo Bonora, Massimo Pagliara, Elisabetta Dall'Aglio, Lucio Luchetti, Giuseppe Buonanno, Piero Angelo Bonati, Marcello Bergonzani, Luigi Gnudi, Mario Passeri, et al. Risk factors for coronary artery disease in healthy persons with hyperinsulinemia and normal glucose tolerance. *New England Journal of Medicine*, 320(11):702–706, 1989.

6 DETAILED DESCRIPTION OF DATASETS AND EXPERIMENTAL SETUP

6.1 DATASETS

Squat Posture Correction This datset consists of 4000 video frames of squat posture our research team collected over 6 months. Professional trainers performed the conventional squat, filed by three Kinect V-2 devices placed at three different angles. Going through discussions with trainers, we determined one correct squat postures and 10 types of typical incorrect posture that non-experience people mistakenly make, which makes posture correction task as multi-label classification task (e.g., 1) *Exaggerated knees-forward movement* or 2) *sitting on the tighs instead of betweeen them*). The average runtime for one video is 5.8 seconds with around 60 frames. For cost-efficiency, we set the frame skip as 3, such that each instance has 14 timesteps. Instead of using raw pixels as input, we extracted 14 pairs of body joints from a human object in frames by using the famous Openpose model (Cao et al., 2017). Extracted body points consist of 14 pairs of $\mathbf{x}$ and $\mathbf{y}$ coordinates, which is expected highly useful when determining attention labels due to its autonomical locality. All data examples have two labels: 1) Label for *class*, 2) Label for *attention*. For example, the data example, labeled as *sitting on the tighs instead of betweeen them* has attention labels: Left hip, Right hip, Left knee, Right Knee. Information about the extracted 14 pairs of body joints and 11 classes are shown in the table. We performed additional experiment on this dataset with respect to interactive attention learning in the next chapter. Three annotators participated in the annotation evaluation procedures. In the case that the same set of negative train points is delievered to multiple annotators, the accumulated sets are aggregated into one annotation matrix by averaging: $\mathbf{m}_k = \frac{1}{I}\sum_{i=1}^{I}\mathbf{m}_k^{(i)}$.

Electronic Health Records This datasets is a subset of electronic healthcare records-based database from healthcare organization, consisting of around 1.5 million records. The database contains demographic information including medical aid beneficiaries, treatmenet information, disease histories, and drug prescription records. In total, 34 features regarding vital signs, social and behavioral factors, medical history, and general information, were extracted from the database over 12 years. Total cholesterol level and fasting glucose level were sampled after overnight fasting and systolic blood pressure and diastolic blood pressure were checked through medical examinations. Also, there were several questionnaires that are designed to identify social and behavioral risk factors, such as smoking habit, alcohol consumption, and time spent on excercise. Individual medical history was followed with drug perscription history and clinical codes of the 10th revision of the International Classification of Diseases (ICD-10). We determined patients with pancreatic cancer by identifying ICD code, C25, on examination and treatment records. On the labeling process, we exclude those who had previous pancreatic cancer-related treatment records as well as pre-existing medical history of pancreatic cancer. Two physicians participated in the experiments with CVD, cerebral infarction, and heart failure tasks, as an annotator.

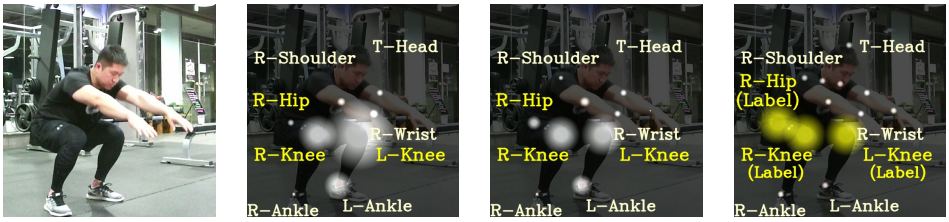
Real-estate Price Forecast in New York City The datasets are the subset of residential sales transaction database from the Department of Finance's Rolling sales files list properties, sold in the last 17 years from 2003 to 2018. We combine the subset from the rolling sales files with another subset extracted from Final Property Assessment Data from all NYC properties. The dataset we processed has a very hetero geneous set of homes spread over five boroughs in New York City: 1) The Bronx, 2) Queens, 3) Brooklyn, 4) Manhattan, 5) Staten Island. Each house is described by a total of 182 attribute variables. These attributes specific to 1) the house-related profiles (Number of bed rooms and bethrooms, square footage, and the year built), 2) Realestate owner-related information (Tax information or Salary), 3) Geographical information (e,g, distance from a hospital, school score in that neighborhood, and the number of hospital facilities within 5 miles), 4) Global economic indicators (e.g., Global copper price, interest rates, total vehicle sales, and Russell 2000). Two real-estate business managers in New York City annotated attentions for real-estate price forecast tasks.

6.2 CONFIGURATION AND PARAMETERS FOR THE RISK PREDICTION TASKS

We trained all the models using Adam (Kingma & Ba, 2014) optimizer with dropout regularization. We set the maximum iteration for Adam optimizer as $10,000$, and for other hyperparameters, we searched for the optimal values by cross-validation, within predefined ranges as fol-

lows: Mini batch size: $\{32, 64, 128, 256\}$, annotation subsampling batch size: $\{8, 16, 32\}$ learning rate: $\{0.01, 0.001, 0.0001\}$, *L-2* regularization: $\{0.02, 0.002, 0.0002, 0.0004\}$, and dropout rate $\{0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5\}$.

7 Benefits of Interactive Attention Learning Framework



(a) Original Image (b) IAL-NAP(S=1) (c) IAL-NAP(S=3) (d) Answer

Figure 8: Attention map for 14 pairs of body joints, generated from RNNs trained for squat pose correction task (Fitness).

Exercise Posture Correction Task In the interactive attention learning framework, professional trainers interactively evaluate visualized attentions generated from the attentional network via annotation attention masks. In figure 8, the size of white circles represents the size of attention weights. For the given the instance (Label 2. Rounding back like c, Attention label: R-Knee, L-Knee, R-Hip, L-Hip), (c) shows that the network evenly generates weights on both left and right knees by allocating more weights on R-Knee over three iterations, compared to the initial iteration (b). An attention network relearned how to attend for a given input, as a human annotator guided (d)shows that attention answer for correcting posture.

	Cerebral Infarction	Fatty Liver
t=1	0.3242	0.2881
t=2	0.2844	0.2752
t=3	**0.2485**	**0.2639**

Table 3: Type 1 Error over three iterations, which shows percentage of features selected from the model that do not match the features selected by the clinicians.

Further Interpretation in the cerebral infarction and fatty liver tasks We further quantitatively compared the accuracy of attentions, using variables selected meaningful by the physicians as ground truth labels (avg. 134 variables per patient). We randomly selected 10 age groups from 40s to 80s for cerebral infarction and fatty liver risk prediction tasks. In table 3, we observe that Type 1 error significantly decreases only with three iterations (from t=1 to t=3).

8 Detailed Description of Attention Annotation Interface

Attention Annotation Interface in the CVD and cerebral infarction tasks Clinical features with high correlation to each task (CVD, Cerebral) were highly annotated according to the value of each features. For example, when the attention is relatively high on LDL value with 90 compared to other features, physician annotated LDL to further change its attention value over the loop. Furthermore, clinical features with low correlation to each task were highly annotated when model weighs them with relatively high attention. Detailed explanations on each features are summarized in Annotation Rules: Cardiovascular Disease and annotation Rules: Cerebral Infarction.

Annotation Rules - Cardiovascular Disease Risk stratification and detailed critea on when the model should attend to each features are summarized below:

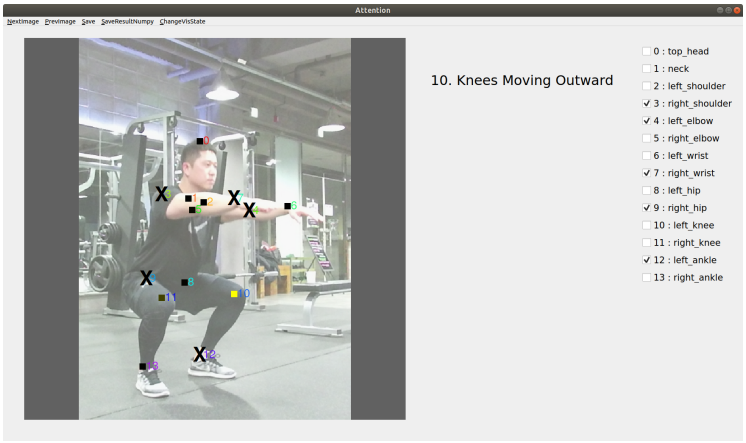


Figure 9: Interactive Attention learning Annotation Interface for **Squat Pose Correction** Task.

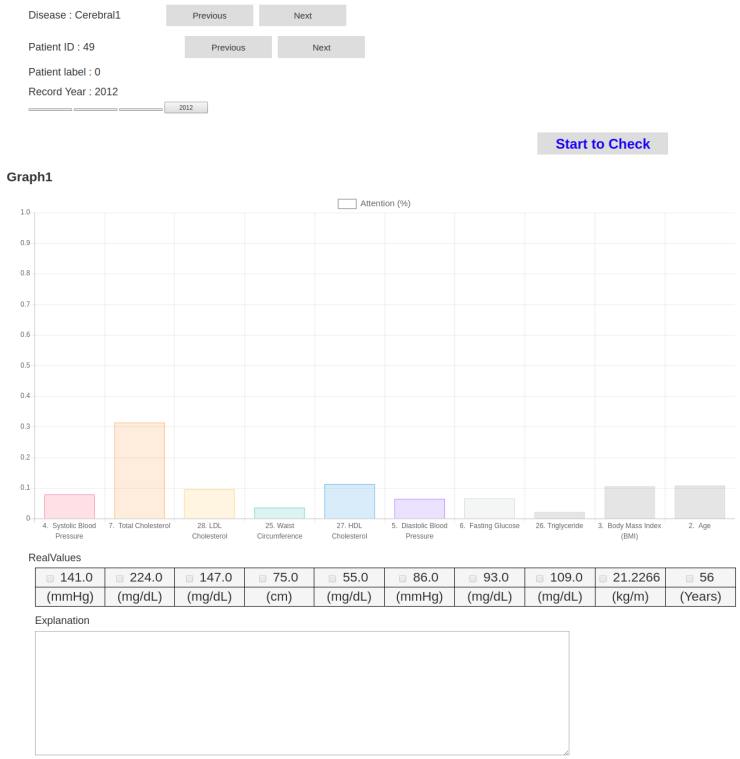


Figure 10: Interactive Attention learning Interface for **EHR datsets**, on which physicians interactively guide the attention network to re-learn how to properly $attend$ to features of a given input. Attentions from the attention network are visualized as above and physicians evaluate them on the web-based attention annotation interface.

First of all, hypertension(systolic blood pressure(SBP) > 140mmHg and diastolic blood pressure(DBP) > 90 mmHg) is quantitatively the most important risk factor of cardiovascular disease (CVD) (Stanaway et al., 2018). Insulin resistance, hyperinsulinemia, diabetic dyslipidemia, and elevated blood glucose are associated with atherosclerotic CVD (Kannel & McGee, 1979; Almdal et al., 2004; Zavaroni et al., 1989). Dyslipidemia, hypercholesterolemia with serum cholesterol $\geq$ 200 mg/dL can be accounted for the attributable risk of CVD (Yusuf et al., 2004; Lowe et al., 1998). Reductions in low-density lipoprotein (LDL) cholesterol levels with the use of statin reduce the risk of CVD

(Downs et al., 1998). Low HDL level (<40mg/dL) raises risk for developing CVD, while high HDL level(>60mg/dL) acts as a protective factor of CVD (Ridker et al., 1998). Obesity(BMI > 30) is associated with a number of risk factors for atherosclerosis, CVD, and cardiovascular mortality. Risk factors for CVD includes diabetic condition of a patient, such as insulin resistance and glucose intolerance (Eckel et al., 2004; Calle et al., 1999).

Among social history of a patient, exposure to tobacco is independent major risk factor, dose-dependently increasing the risk for total atherosclerotic CVD, coronary heart disease(CHD), cerebrovascular disease, heart failure, and mortality (Jee et al., 1999; Qiao et al., 2000; Foody et al., 2001). Smoking cessation is known to be beneficial for preventing CVD; smoking cessation is associated with the reduction in cardiac event rate (Rose et al., 1982), where the risk further decreases with elongation of time since quitting (Novello, 1990). While epidemiologic data indicate that moderate alcohol intake has a protective effect on CHD (Gemes et al., 2016), binge drinking increases the risk for CVD (Roerecke & Rehm, 2010; Ruidavets et al., 2010). Moderate exercise has a protective effect against CHD and all-cause mortality (Powell et al., 1987).

Next, among non-modifiable risk factors, CVD risk increases with aging (over age 45 for men, over age 55 for women), and for the same age patient group, men are more prone to develop cardiovascular disease than women (Jousilahti et al., 1999). Family history of CVD is an independent risk factor for CHD; high risk for the individuals with first-degree relatives who developed atherosclerotic CVD or death from CVD (male relative prior to age 55 and female relative prior to age 65) (Patel et al., 2018; Stone et al., 2014). A wider definition of this significant family history of CVD might also include CVD-related death, stroke, or transient ischemic attack (Patel et al., 2018). History of stroke can also be risk factor of CVD as they both have similar pathophysiology (Anderson et al., 1991). Family history of hypertension(systolic blood pressure > 140mmHg and diastolic pressure > 90 mmHg) (Anderson et al., 1991), diabetes (Anderson et al., 1991) can indirectly be a risk factor of CVD. Additionally, for other features like hemoblogin, urine protein, AST, ALT, GGT, Creatinine and history of pulmonary tuberculosis, there is no proven evidence on the effect of these values with cardiovascular disorder (Pencina et al., 2019).

Annotation Rules - Cerebral Infarction Risk stratification and detailed critea on when the model should attend to each features are summarized below:

History of stroke and transient ischemic attack in the same territory strongly predicts future stroke occurrence (Society, 2016). Hypertension(SBP > 140mmHg and DBP > 90 mmHg) is quantitatively the most common and most important risk factor for stroke with estimated relative risk of 4.0-5.0 and estimated prevalence of 25-40% (Society, 2016; Ezekowitz et al., 2003; Jorgensen et al., 1994). A cardiac evaluation (e.g. echocardiogram) to find out whether patient has cardiac disease, such as atrial fibrillation or other embolic conditions, is important in managing risk factors for stroke (Jorgensen et al., 1994; Ezekowitz et al., 2003). Diabetes itself and diabetic conditions such as insulin resistance, elevated blood glucose increase the likelihood of large and small artery occlusive disease (Jorgensen et al., 1994). Risks for stroke stem not only from increased likelihood of atherogenesis but also from aggravation of other risk factors including hypertension and hyperlipidemia (Najarian et al., 2006). Preventing dyslipidemia by lowering LDL cholesterol and elevating HDL may prevent strokes (Society, 2016; Hindy et al., 2018). Also, compared to those with normal BMI, obese and overweight patients have significantly better early and long-term survival rates, which is called the paradox of obesity (Vemmos et al., 2011; Banack & Kaufman, 2013).

Among social history of a patient, smoking increases the likelihood of CVD, more than doubling the risk of stroke with relative risk of 1.5-2.9 and estimated prevalence of 4-8% (Shah & Cole, 2010; Ezekowitz et al., 2003) which decreases with the cessation of smoking proportional to the period after cessation (Kawachi et al., 1993). Moderate and high level of exercise is associated with reduced risk of stroke (Society, 2016; Potempa et al., 1995; Lee et al., 2003). Epidemiologic data indicate that moderate alcohol intake has a protective effect on stroke. However, binge drinking increases the risk for stroke (Society, 2016; Hillbom et al., 1999).

Aging is nonmodifiable risk factor for ischemic stroke, and also for mortality and morbidity (Society, 2016; Lee et al., 2003; Roy-O Reilly & McCullough, 2018). Also, older subjects are more prone to develop CVD and emblic, thrombotic stroke compared to their younger counterparts (Roy-O Reilly & McCullough, 2018; Nakayama et al., 1994). Individuals with family history of stroke (Liao et al., 1997; Jerrard-Dunne et al., 2003), cardiac conditions(especially atrial fibrillation) (Fox et al., 2004),

and hypertension (Staessen et al., 2003; Wang et al., 2008) possess genetic susceptibility, thus high risk of developing stroke compared to the individuals without family history. Furthermore, family history of type II diabetes in any first degree relative have a two to three-fold increased risk of developing diabetes thus indirectly increasing risk of stroke, compared to individuals without family history (Consortium et al., 2013; Meigs et al., 2000). Additionally, for other features like hemoblogin, urine protein, AST, ALT, GGT, Creatinine and history of pulmonary tuberculosis, there is no proven evidence on the effect of these values with cardiovascular disorder (Society, 2016).

REFERENCES

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 559–560. ACM, 2018.

Thomas Almdal, Henrik Scharling, Jan Skov Jensen, and Henrik Vestergaard. The independent effect of type 2 diabetes mellitus on ischemic heart disease, stroke, and death: a population-based study of 13 000 men and women with 20 years of follow-up. *Archives of internal medicine*, 164(13): 1422–1426, 2004.

Keaven M Anderson, Patricia M Odell, Peter WF Wilson, and William B Kannel. Cardiovascular disease risk profiles. *American heart journal*, 121(1):293–298, 1991.

M Gethsiyal Augasta and Thangairulappan Kathirvalavakumar. Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*, 35(2):131–150, 2012.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

Hailey R Banack and Jay S Kaufman. The "obesity paradox" explained. *Epidemiology*, 24(3): 461–462, 2013.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1901.09887*, 2019.

Eugenia E Calle, Michael J Thun, Jennifer M Petrelli, Carmen Rodriguez, and Clark W Heath Jr. Body-mass index and mortality in a prospective cohort of us adults. *New England Journal of Medicine*, 341(15):1097–1105, 1999.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

Lu Chi and Yadong Mu. Deep steering: Learning end-to-end driving model from spatial and temporal visual cues. *arXiv preprint arXiv:1708.03798*, 2017.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*. 2016.

Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 230. ACM, 2019.

Clark, Ian, and Guillaume Dumas. Toward a neural basis for peer-interaction: what makes peer-learning tick? *Frontiers in psychology*, 2015.

InterAct Consortium et al. The link between family history and risk of type 2 diabetes is not explained by anthropometric, lifestyle or genetic risk factors: the epic-interact study. *Diabetologia*, 56(1): 60–69, 2013.

R Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

John R Downs, Michael Clearfield, Stephen Weis, Edwin Whitney, Deborah R Shapiro, Polly A Beere, Alexandra Langendorfer, Evan A Stein, William Kruyer, Antonio M Gotto Jr, et al. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of afcaps/texcaps. *Jama*, 279(20):1615–1622, 1998.

Robert H Eckel, David A York, Stephan Rossner, Van Hubbard, Ian Caterson, Sachiko T St. Jeor, Laura L Hayman, Rebecca M Mullis, and Steven N Blair. Prevention conference vii: Obesity, a worldwide epidemic related to heart disease and stroke: executive summary. *Circulation*, 110(18): 2968–2975, 2004.

Justin A Ezekowitz, Sharon E Straus, Sumit R Majumdar, and Finlay A McAlister. Stroke: strategies for primary prevention. *American family physician*, 68(12):2379–2386, 2003.

JoAnne Micale Foody, Christopher R Cole, Eugene H Blackstone, and Michael S Lauer. A propensity analysis of cigarette smoking and mortality with consideration of the effects of alcohol. *The American journal of cardiology*, 87(6):706–711, 2001.

Caroline S Fox, Helen Parise, Ralph B DAgostino Sr, Donald M Lloyd-Jones, Ramachandran S Vasan, Thomas J Wang, Daniel Levy, Philip A Wolf, and Emelia J Benjamin. Parental atrial fibrillation as a risk factor for atrial fibrillation in offspring. *Jama*, 291(23):2851–2855, 2004.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Mariana Garcia, Sharon L Mulvagh, C Noel Bairey Merz, Julie E Buring, and JoAnn E Manson. Cardiovascular disease in women: clinical perspectives. *Circulation research*, 118(8):1273–1293, 2016.

Marta Garnelo, Dan Rosenbaum, Chris J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional neural processes. *CoRR*, abs/1807.01613, 2018a. URL `http://arxiv.org/abs/1807.01613`.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes. *CoRR*, abs/1807.01622, 2018b. URL `http://arxiv.org/abs/1807.01622`.

K Gemes, I Janszky, LE Laugsand, KD Laszlo, S Ahnve, LJ Vatten, and KJ Mukamal. Alcohol consumption is associated with a lower incidence of acute myocardial infarction: results from a large prospective population-based study in norway. *Journal of internal medicine*, 279(4):365–375, 2016.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. IEEE, 2018.

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245*, 2018.

Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. In *Advances in Neural Information Processing Systems*, pp. 909–918, 2018.

Matti Hillbom, Heikki Numminen, and Seppo Juvela. Recent heavy drinking of alcohol and embolic stroke. *Stroke*, 30(11):2307–2312, 1999.

George Hindy, Gunnar Engstrom, Susanna C Larsson, Matthew Traylor, Hugh S Markus, Olle Melander, and Marju Orho-Melander. Role of blood lipids in the development of ischemic stroke and its subtypes: a mendelian randomization study. *Stroke*, 49(4):820–827, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short term memory. *Neural Computation*, 9: 1735–1780, 1997.

Rachel R Huxley and Mark Woodward. Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies. *The Lancet*, 378(9799):1297–1305, 2011.

Sun Ha Jee, Il Suh, Il Soon Kim, and Lawrence J Appel. Smoking and atherosclerotic cardiovascular disease in men with low levels of serum cholesterol: the korea medical insurance corporation study. *Jama*, 282(22):2149–2155, 1999.

Paula Jerrard-Dunne, Geoffrey Cloud, Ahamad Hassan, and Hugh S Markus. Evaluating the genetic component of ischemic stroke subtypes: a family history study. *Stroke*, 34(6):1364–1369, 2003.

Henrik Stig Jorgensen, Hirofumi Nakayama, TS Olsen, and HO Raaschou. Effect of blood pressure and diabetes on stroke in progression. *The Lancet*, 344(8916):156–159, 1994.

Pekka Jousilahti, Erkki Vartiainen, Jaakko Tuomilehto, and Pekka Puska. Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14 786 middle-aged men and women in finland. *Circulation*, 99(9):1165–1172, 1999.

WB Kannel and DL McGee. Diabetes and glucose tolerance as risk factors for cardiovascular disease: the framingham study. *Diabetes care*, 2(2):120–126, 1979.

Ichiro Kawachi, Graham A Colditz, Meir J Stampfer, Walter C Willett, JoAnn E Manson, Bernard Rosner, Frank E Speizer, and Charles H Hennekens. Smoking cessation and decreased risk of stroke in women. *Jama*, 269(2):232–236, 1993.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, S. M. Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *CoRR*, abs/1901.05761, 2019. URL `http://arxiv.org/abs/1901.05761`.

Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1885–1894. JMLR. org, 2017.

Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. In *Advances in Neural Information Processing Systems*, pp. 10159–10168, 2018.

Chong Do Lee, Aaron R Folsom, and Steven N Blair. Physical activity and stroke risk: a meta-analysis. *Stroke*, 34(10):2475–2481, 2003.

Duanping Liao, Richard Myers, Steven Hunt, Eyal Shahar, Catherine Paton, Gregory Burke, Michael Province, and Gerardo Heiss. Familial history of stroke and stroke risk: the family heart study. *Stroke*, 28(10):1908–1912, 1997.

Lynn P Lowe, Philip Greenland, Karen J Ruth, Alan R Dyer, Rose Stamler, and Jeremiah Stamler. Impact of major cardiovascular disease risk factors, particularly in combination, on 22-year mortality in women and men. *Archives of internal medicine*, 158(18):2007–2014, 1998.

James B Meigs, L Adrienne Cupples, and PW Wilson. Parental transmission of type 2 diabetes: the framingham offspring study. *Diabetes*, 49(12):2201–2207, 2000.

Robert M Najarian, Lisa M Sullivan, William B Kannel, Peter WF Wilson, Ralph B D'Agostino, and Philip A Wolf. Metabolic syndrome compared with type 2 diabetes mellitus as a risk factor for stroke: the framingham offspring study. *Archives of internal medicine*, 166(1):106–111, 2006.

Hirofumi Nakayama, HS Jorgensen, HO Raaschou, and Tom Skyhøj Olsen. The influence of age on stroke outcome. the copenhagen stroke study. *Stroke*, 25(4):808–813, 1994.

Antonia C Novello. Surgeon general's report on the health benefits of smoking cessation. *Public Health Reports*, 105(6):545, 1990.

Jaideep Patel, Mahmoud Al Rifai, Maren T Scheuner, Steven Shea, Roger S Blumenthal, Khurram Nasir, Michael J Blaha, and John W McEvoy. Basic vs more complex definitions of family history in the prediction of coronary heart disease: the multi-ethnic study of atherosclerosis. In *Mayo Clinic Proceedings*, volume 93, pp. 1213–1223. Elsevier, 2018.

Michael J Pencina, Ann Marie Navar, Daniel Wojdyla, Robert J Sanchez, Irfan Khan, Joseph Elassal, Ralph B D'Agostino Sr, Eric D Peterson, and Allan D Sniderman. Quantifying importance of major risk factors for coronary heart disease. *Circulation*, 139(13):1603–1611, 2019.

Kathleen Potempa, Martita Lopez, Lynne T Braun, J Peter Szidon, Louis Fogg, and Tyler Tincknell. Physiological outcomes of aerobic exercise training in hemiparetic stroke patients. *Stroke*, 26(1): 101–105, 1995.

Kenneth E Powell, Paul D Thompson, Carl J Caspersen, and Juliette S Kendrick. Physical activity and the incidence of coronary heart disease. *Annual review of public health*, 8(1):253–287, 1987.

Qing Qiao, M Tervahauta, A Nissinen, and J Tuomilehto. Mortality from all causes and from coronary heart disease related to smoking and changes in smoking during a 35-year follow-up of middle-aged finnish men. *European heart journal*, 21(19):1621–1626, 2000.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778. URL `http://doi.acm.org/10.1145/2939672.2939778`.

Paul M Ridker, Robert J Glynn, and Charles H Hennekens. C-reactive protein adds to the predictive value of total and hdl cholesterol in determining risk of first myocardial infarction. *Circulation*, 97 (20):2007–2011, 1998.

Michael Roerecke and Jurgen Rehm. Irregular heavy drinking occasions and risk of ischemic heart disease: a systematic review and meta-analysis. *American journal of epidemiology*, 171(6): 633–644, 2010.

Geoffrey Rose, PJ Hamilton, L Colwell, and MJ Shipley. A randomised controlled trial of anti-smoking advice: 10-year results. *Journal of Epidemiology & Community Health*, 36(2):102–108, 1982.

Meaghan Roy-O Reilly and Louise D McCullough. Age and sex are critical factors in ischemic stroke pathology. *Endocrinology*, 159(8):3120–3131, 2018.

Jean-Bernard Ruidavets, Pierre Ducimetiere, Alun Evans, Michele Montaye, Bernadette Haas, Annie Bingham, John Yarnell, Philippe Amouyel, Dominique Arveiler, Frank Kee, et al. Patterns of alcohol consumption and ischaemic heart disease in culturally divergent countries: the prospective epidemiological study of myocardial infarction (prime). *Bmj*, 341:c6077, 2010.

Steven L Salzberg. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3):235–240, 1994.

Vignesh Sankar, Devinder Kumar, David A Clausi, Graham W Taylor, and Alexander Wong. Sisc: End-to-end interpretable discovery radiomics-driven lung cancer prediction via stacked interpretable sequencing cells. *arXiv preprint arXiv:1901.04641*, 2019.

Makoto Sato and Hiroshi Tsukimoto. Rule extraction from neural networks via decision tree induction. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 3, pp. 1870–1875. IEEE, 2001.

Reena S Shah and John W Cole. Smoking and stroke: the more you smoke the more you stroke. *Expert review of cardiovascular therapy*, 8(7):917–932, 2010.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

Korea Stroke Society. *National clinical guideline for stroke*. Korea Stroke Society, 2016.

Jan A Staessen, Jiguang Wang, Giuseppe Bianchi, and Willem H Birkenhager. Essential hypertension. *The Lancet*, 361(9369):1629–1641, 2003.

Jeffrey D Stanaway, Ashkan Afshin, Emmanuela Gakidou, Stephen S Lim, Degu Abate, Kalkidan Hassen Abate, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159): 1923–1994, 2018.

Neil J Stone, Jennifer G Robinson, Alice H Lichtenstein, C Noel Bairey Merz, Conrad B Blum, Robert H Eckel, Anne C Goldberg, David Gordon, Daniel Levy, Donald M Lloyd-Jones, et al. 2013 acc/aha guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Part B):2889–2934, 2014.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *NIPS*, 2015.

Konstantinos Vemmos, George Ntaios, Konstantinos Spengos, Paraskevi Savvari, Anastasia Vemmou, Theodora Pappa, Efstathios Manios, George Georgiopoulos, and Maria Alevizaki. Association between obesity and mortality after acute first-ever stroke: the obesity–stroke paradox. *Stroke*, 42 (1):30–36, 2011.

Nae-Yuh Wang, J Hunter Young, Lucy A Meoni, Daniel E Ford, Thomas P Erlinger, and Michael J Klag. Blood pressure change and risk of hypertension associated with parental hypertension: the johns hopkins precursors study. *Archives of internal medicine*, 168(6):643–648, 2008.

Alexander Wong. *Waterloo, Canada N2L 3G1 Phone: 519-888-4567 ext. 31299 a28wong@ engmail. uwaterloo. ca*. PhD thesis, University of Waterloo, 2018.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320–3328, 2014.

Salim Yusuf, Steven Hawken, Stephanie Ounpuu, Tony Dans, Alvaro Avezum, Fernando Lanas, Matthew McQueen, Andrzej Budaj, Prem Pais, John Varigos, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): case-control study. *The lancet*, 364(9438):937–952, 2004.

Ivana Zavaroni, Enzo Bonora, Massimo Pagliara, Elisabetta Dall'Aglio, Lucio Luchetti, Giuseppe Buonanno, Piero Angelo Bonati, Marcello Bergonzani, Luigi Gnudi, Mario Passeri, et al. Risk factors for coronary artery disease in healthy persons with hyperinsulinemia and normal glucose tolerance. *New England Journal of Medicine*, 320(11):702–706, 1989.