1. Configuration and Parameters for the risk prediction tasks

We implement all models used in the experiments using Python and Tensorflow. We train all the models using Adam (Kingma & Ba, 2014) optimizer with dropout regularization and \( \ell_2 \)-regularization. For hyperparameters, we search for the optimal values by cross-validation, within predefined ranges as follows: Mini batch size: \{32, 64, 128, 256\}, annotation subsampling batch size: \{8, 16, 32\}, learning rate: \{0.01, 0.001, 0.0001\}, \( \ell_2 \) regularization: \{0.02, 0.002, 0.0002, 0.0004\}, and dropout rate \{0.1, 0.2, 0.25, 0.3, 0.4, 0.5\}. For hyperparameters \( P, K, \) and \( F \) we use in Cost-Effective instance and feature Selection (CES) algorithm, we search for the optimal value via cross-validation in the following ranges: \( P : \{100, 200\} \), \( K : \{20, 40, 50\} \), and \( F : \{20, 30\} \) for risk prediction tasks and \( P : \{200, 400, 600, 800\} \), \( K : \{50, 100\} \), and \( F : \{20, 30, 40, 50\} \) for real estate forecasting task, and, for fitness squat task, \( F : \{20, 40\} \), \( P : \{20, 40, 60, 80\} \), and \( K : \{6, 10, 20\} \).

2. Additional Experiments for Interpretability

**Model Interpretation** We use the interpretation method of RETAIN (Choi et al., 2016), since our interpretable model is based on it. The most important feature of RETAIN is that it allows us to interpret what the model has learned as follows. What we are interested in is contribution, which shows \( x_d \)'s aggregate effect to the final prediction at time \( t \). Since RETAIN has attentions on both timesteps \((\beta^{(i)})\) and features \((\gamma^{(i)})\), the computation of aggregate contribution takes both of them into consideration when computing the final contribution of an input data point at a specific timestep:

\[
\omega(y, x_d^{(i)}) = \beta^{(i)} w^T (\gamma^{(i)} \odot W_{emb}[; , d]) x_d^{(i)} \tag{1}
\]

where \( w \) is the parameters of an output layer \( h \) to learn.

Each feature contribution \( \omega(y, x_d^{(i)}) \) represents a certain portion of model prediction \( \hat{y} \) for which \( x_d^{(i)} \) is responsible.

### Table 1. 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.

<table>
<thead>
<tr>
<th></th>
<th>Cerebral Infarction</th>
<th>Heart Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s=1 )</td>
<td>0.3142</td>
<td>0.2891</td>
</tr>
<tr>
<td>( s=2 )</td>
<td>0.2992</td>
<td>0.2712</td>
</tr>
<tr>
<td>( s=3 )</td>
<td>0.2648</td>
<td>0.2573</td>
</tr>
</tbody>
</table>

2.1. Quantitative Analysis

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

**Benefits of selecting \( P \) instances** In our Cost-Effective instance and feature Reranking (CER) algorithm, we are given a set of validation instances \( D_{valid} = \{ u_j \}_{j=1}^M \) during training. We first select \( P \) instances that have the highest validation loss \( \mathcal{L}(\Theta, u_j^{val}) \) to comprise \( D'_{valid} = \{ u_j^{val} \}_{j=1}^P \). Searching for the optimal value \( P \) is crucial, such that we select the training instances with large impact on the validation instances that are mis-predicted by the currently trained model. We perform an experiment that shows the difference in percentage of incorrectly interpreted instances that are delivered to the annotator between IAL-NAP with \( P \)-instance selection and without \( P \) (use all validation instances to calculate instance importance) on cerebral infarction and heart failure task.

**Table 2. Type 1 Error Rate at first round \((s=1)\), which shows percentage of features selected from the model that do not match the features selected by the clinicians.**

<table>
<thead>
<tr>
<th></th>
<th>Cerebral Infarction</th>
<th>Heart Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAL-NAP with ( P )</td>
<td>0.3215</td>
<td>0.2862</td>
</tr>
<tr>
<td>IAL-NAP without ( P )</td>
<td>0.2164</td>
<td>0.1792</td>
</tr>
</tbody>
</table>

Table 2 shows the percentage of false positive (Type 1 Error)
in interpretations of $K$ (100) training instances between IAL-NAP with $K$ and without $K$ instance selection. The results shows that IAL-NAP with $P$ validation instances captured more incorrectly interpreted instances than without $P$ selection on cerebral infarction and heart failure tasks, which shows the benefit of $P$ instance selection when it comes to cost-effectively detecting training instances with incorrect interpretation.

### 2.2. Qualitative Analysis

**Risk Prediction** We further perform qualitative analysis on risk prediction tasks. In Figure 1, 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 IAL-NAP attests to the most are interpreted by physicians as key risk factors for accessing a CVD patient. Although Random-NAP failed to ignore GGT which is a relatively less important variable, it accesses the key variable better than other models with IF-RETAIN. For broad clinical descriptions for figure 1, please see the section 4. These examples shows another advantage of our model: It provides human-intuitive interpretations of why it arrives at such a decision that are well-aligned with human knowledge.

Furthermore, in Figure 2, the green bar graph (left) represents the total number of variable corrections that annotator made on CVD task, during evaluation procedure. Out of 34 variables, we visualize the most corrected 10 variables. The blue bar graph (right) shows the rank of variables that yields an attention weight with high variance. For each sample of collected CVD annotations, we list 10 variables with the highest uncertainty and score 10 points to the highest uncertain variable and 1 point to the the lowest and aggregate the all points that each feature received. As a result, it is found that our model, IAL-NAP, precisely identify 70% of the features on which annotators make a correction on and they are cost-effectively prioritized. Feature on red color in blue bars indicates the features matched in green graph.

**Squat Pose Correction Task** In our interactive attention learning framework, professional trainers interactively evaluate visualized attentions generated from the attentional network via annotation attention masks. In figure 3, 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 re-learns how to attend for a given input, as a human annotator guides. (d) shows that attention labels for correcting posture.
3. Detailed Description of Datasets

3.1. Datasets

Squat Posture Correction This dataset consists of 4,000 video frames of squat posture 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 posture and 10 types of typical incorrect postures 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 tights instead of between 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 x and y coordinates, which is expected highly useful when determining attention labels due to its anatomical 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 tights instead of between 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 delivered to multiple annotators, the accumulated sets are aggregated into one annotation matrix by averaging: $m_k = \frac{1}{I} \sum_{i=1}^{I} m_k^{(i)}$.

Electronic Health Records This dataset 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, treatment 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 exercises. Individual medical history was followed with drug prescription history and clinical codes of the 10th revision of the International Classification of Diseases (ICD-10). We determined patients with cardiovascular disease, cerebral infarction, and heart failure tasks by identifying ICD code, C25, on examination and treatment records. Two physicians participated in the experiments for all risk prediction 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 heterogeneous 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 bathrooms, square footage, and the year built), 2) Real estate owner-related information (Tax information or Salary), 3) 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 attention for real-estate price forecast tasks.

4. Detailed Description of Attention Annotation Interface

Annotation interface for risk prediction tasks is shown in Figure 5 and squat pose correction task in Figure 4. Before performing annotation work for risk prediction tasks, we consulted with two physicians in a major hospital to set up annotation rules which are established based on clinical evidence and expertise, and empirical backgrounds. In this section, we describe how annotation rules for cardiovascular disease and cerebral infarction task are determined.

4.1. Cardiovascular Disease

Cardiovascular Disease 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 levels (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 hemoglobin, 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).

**Cerebral Infarction** 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 hyper tension 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 embolic, 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(e 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 hemoglobin, 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).
Figure 4. Online Annotation Interface for Squat Pose Correction Task.

Figure 5. Online Annotation Interface for risk prediction tasks, on which human supervisors 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.
References


