Predictive network with leveraging clinical measures as auxiliary task

Xiangrui Li
Department of Computer Science
Wayne State University
Detroit, MI
Email: xiangruili@wayne.edu

Dongxiao Zhu
Department of Computer Science
Wayne State University
Detroit, MI
Email: dzhu@wayne.edu

Phillip Levy
Department of Emergency Medicine and Cardiovascular Research Institute
Wayne State University
Detroit, MI
Email: plevy@med.wayne.edu

Abstract—Deep neural networks (DNNs) have made impressive improvements for predictive modeling in various fields. Successful network building for predictive tasks usually requires abundant data, for effectively learning high-level, non-additive information from raw input features. The merit of high-level learning makes DNN promising in clinical research, but the need for abundant data hampers DNN applications. To address this challenge, we propose the use of Auxiliary-Task-Augmented Network (ATAN), a predictive model (for primary target) with introducing auxiliary tasks as regularization. ATAN leverages clinically relevant measures as auxiliary targets and learns the clinical relevance explicitly. We focused on applying ATAN in a clinical dataset of hypertension collected from a vulnerable demographic population (African-American). Performance of ATAN was compared with various popular methods to demonstrate its effectiveness. In addition, we analyzed the learned parameters for finding risk factors and interpreting the model.

I. INTRODUCTION

Predictive modeling is an important task in clinical research, as it can unveil associations between risk factors and asymptomatic disease phenotypes enabling prediction of patients most likely to benefit from early intervention. Traditional methods, such as linear regression, are often used but render predictions based only on the low-level features (i.e. demographics, blood pressure, cholesterol level, glucose level et al.), simplifying the inherent, complex biological mechanism of disease progression as additive effects of low-level features. As a result, nonlinear relations are excluded leading to information losses, and potentially less accurate performance.

On the contrary, deep neural networks (DNNs) have made impressive improvements for complex predictive tasks in natural language processing and computer vision. The key factor for their success is that DNNs are capable of learning high-level information from low-level input features. This merit of DNN makes it very promising for predictive modeling in clinical research, enabling capture of non-additive effects among various low-level features. Successful DNNs require abundant labeled data to avoid overfitting. However, collecting labeled data in clinical practice is expensive and time-consuming as the labeling process requires dedicated personnel, explicit definitions, and a formalized mechanism for data compilation. Consequently, we only have limited available labeled data, precluding application of DNN methods to various challenging, but important clinical prediction problems.

The main danger for of using DNNs to analyze small datasets is overfitting, resulting in the potential for limited generalizability. To this end, the overfitting problem in DNNs can be effectively alleviated by various regularization methods including dropout, early stopping and L2 regularization [8]. For predictive modeling in clinical research, this problem can be further mitigated by defining primary targets that are relevant within the context of other associated clinical measures.

For a motivating example, African-Americans with hypertension are at high risk for left ventricular hypertrophy (LVH) and left ventricular mass indexed to body surface area (LVMI) is used as a measure of structural heart damage. In clinical practice, there are many African-Americans with hypertension and it is quite challenging to predict which patient is at greatest risk. However, detection of LVH has important implications on patient care and targeted treatment can decrease the likelihood of developing further cardiovascular complications. At the same time, accurate prediction of those at greatest risk is critical as definitive testing to diagnose LVH, including echocardiography, cardiac computed tomography, and cardiac magnetic resonance imaging (CMR), is quite expensive and testing every African American with hypertension would be cost prohibitive. We recently completed an NIH funded study of African-Americans with hypertension that focused on identification of LVH using LVMI on CMR for diagnosis.

With the considerations described above, herein we propose auxiliary-task augmented network (ATAN), a model that predicts the primary target with introducing clinically relevant measures as auxiliary predictive tasks. With auxiliary tasks, ATAN takes advantage of benefits from multi-task learning (MTL) — an approach of regularization and implicit data augmentation. Without assumption of homogeneous feature representation for all tasks in classic multi-task frameworks, ATAN explicitly models the shared feature representation for all tasks, as well as task-specific representation, and combines them together using a weighting mechanism to capture the clinical relevance. By the weighting mechanism, we conceptually quantify the relevance between the primary and auxiliary target. While LVMI was collected as the primary target in our
test case, many other variables from CMR that are clinically related with LVMI are also available (see Section IV-A), providing auxiliary targets that can be utilized in predictive modeling. To demonstrate the effectiveness of our method, we apply ATAN to our hypertension dataset and compare its predictive performance with different popular methods (i.e., k-nearest neighbor, linear regression, Lasso et al. [6]). We also interpret our model by analyzing the learned parameters (a.k.a. weights) to identify clinically relevant risk factors.

The rest of this paper is organized as follows. In Section II, we review related works in DNNs and multi-task methods. Section III describes details of our proposed method. In Section IV, experimental results are shown with model analysis for interpretation. Finally, we conclude this paper with discussion in Section V.

II. RELATED WORK

Deep neural network DNNs have been successfully applied in complex problems such as image classification and natural language processing ([2], [10]). In the fields of health informatics and bioinformatics, DNN applications have recently been flourishing due to DNNs’ capacity of capturing latent non-additive interactions among low-level input features.

Based on the large electronic health record (EHR), [3] leverages longitudinal patient health history and builds a recurrent neural network (RNN) using gated recurrent unit (GRU) for early prediction of onset heart failure; [13] combines unsupervised pre-training with deep belief networks to train a classifier for decision making in healthcare. In [23], a feed-forward fully connected network is proposed for gene annotation. In [16], DNN is used for high-level feature learning and the learned feature representations are further combined with traditional method for predictive modeling as well as risk factor prioritization. In medical image processing, [15] proposes convolutional neural network (CNN) with cascade architecture to predict the location of left ventricular region in cardiac magnetic resonance images; [12] builds CNN using dilated convolutions for detecting secondary structure in protein density maps.

Multi-task learning MTL for predictive modeling is a machine learning paradigm that jointly learns a model for multiple tasks [31]. When these tasks are related to each other, the joint learning can lead to improvement in the generalized predictive performance by leveraging information contained in other tasks. This perspective inspires a learning strategy that many single-task problems often introduce auxiliary tasks and thereafter be transformed into multi-task problems. With this approach for the target (single) task, MTL can be viewed as a method of regularization and implicit data augmentation. See [1], [8], [24], [31] for more details on MTL. There have been many successful applications of MTL in biomedicine. For example, [26], [30] apply MTL in conjunction with linear model to predict Alzheimer’s disease progression; [18] uses MTL for a multi-label task in a clinical text classification problem; [20] propose group Lasso under the MTL paradigm to detect population heterogeneity at the genetic level.

DNN MTL DNN-based MTL (DMTL) has been attracting much interest recently as DNNs can be conveniently adapted to MTL: DNN can have multiple output neurons for multiple tasks. Aside the high-level learning of DNN, another advantage of DMTL over linear-based modeling is that the hierarchical feature learning provides a flexible way of capturing relevance among multiple tasks. These merits lead to many applications for various problems such as speech recognition and computer vision ([14], [25], [29], [32]). Particularly in bioinformatics and health informatics, [22] uses DMTL to predict protein interactions between HIV and human proteins; [27] integrates DNN feature learning into SVM-based MTL for diagnosis of Alzheimer’s disease. For general considerations in DMTL, we refer to [1] and [24] for details.

The proposed ATAN is a DNN approach of predictive modeling for our primary target. ATAN introduces other clinically relevant features as auxiliary tasks for enjoying the benefits of MTL as regularization. We simultaneously learn task-specific feature representations and shared representations to conceptually capture the clinical relevance among primary and auxiliary targets.

III. METHOD

In this section, we present the proposed method ATAN. We start with the fully connected network for predictive tasks. Then we describe our approach of utilizing auxiliary targets under the multi-task learning framework. Finally, we present a method adapted from [7] to our case for finding the risk factors.

Notations: We use $\mathbb{R}$ to represent the set of real numbers. Vectors and matrices are denoted as bold letters and scalars as unbold letters.

A. Basic Feed-forward Network

We implemented fully connected feed-forward neural network (FNN) as the building block of our model. But FNN itself can serve as a predictive model. The mathematical formulation of FNN is described as follows.

Let $(x, y) \in \mathbb{R}^p \times \mathbb{R}$ be a sample, where $x$ is the input $p$-dimensional feature and $y$ be the target. A FNN for regression task consists of one layer that takes $x$ as input, followed by $k$ hidden layers for learning feature representations:

$$h_1 = \sigma(W_1 x + b_1),$$
$$h_i = \sigma(W_i h_{i-1} + b_i), \quad i = 2, \ldots, k$$  \hspace{1cm} (1)

and one output layer that make predictions using the highest level feature representations:

$$\hat{y} = Wh_k + b,$$

where $W_i$ and $W$ are the weight matrices with compatible dimensions, $b_i$ and $b$ the bias terms and $h_i$ the hidden state; $\sigma(\cdot)$ is the element-wise activation function, such as sigmoid, ReLU or tanh; $\hat{y}$ is the prediction of the network.

When there are multiple targets, FNN can be modified effortlessly to accommodate the case. In the output layer,

$$\hat{y} = Wh_k + b,$$
where \( \hat{\mathbf{y}} = (\hat{y}_1, \cdots, \hat{y}_T) \) is the predictive vector of \( T \) targets. Notice that in this formulation of multi-task learning, an implicit assumption is made that all tasks share the feature representation.

To learn the model parameters, gradient descent based methods along with back propagation are used to minimize the least square loss function.

### B. ATAN Structure

Our central task is to build a predictive model for the primary target that clinicians care about but costly to label (LVMI in our motivation example). We denote the primary target by \( y^c \). In practice, we observe that in addition to the primary measure \( y^c \), there are often other measures available, denoted as \( y^a \), that are clinically relevant to \( y^c \). This relevance provides additional information for the primary task. By incorporating \( y^a \) as the secondary target, the predictive model can benefit from multi-task learning as described in Section II.

However, one pitfall with introducing auxiliary tasks is that the “relevance” is not universally defined. Multi-task methods generally either make assumptions, or define the relevance based on the domain-specific knowledge. ATAN circumvents this issue assuming that the primary target \( y^c \) and the secondary target \( y^a \) are jointly regulated by the shared and task-specific biological mechanisms. Translating this assumption in modeling, ATAN learns a feature representation that can be decomposed into a weighted sum of the shared and task-specific feature representation. Fig. 1 presents the high-level overview of ATAN structure.

**Learning feature representations** Let \((x, y^c, y^a) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}\) be a sample. We first learn the shared and task-specific feature representations using feed-forward DNN (FDNN) as follows:

\[
\begin{align*}
\mathbf{h}^s &= \text{FDNN}^s(x), \\
\mathbf{h}^c &= \text{FDNN}^c(x), \\
\mathbf{h}^a &= \text{FDNN}^a(x),
\end{align*}
\]

where \( \text{FDNN}(\cdot) \) is calculated by (1) and the activation function therein is the element-wise sigmoid function \( 1/(1+\exp(-x)) \). For notational convenience, we use \( E^s \), \( E^c \) and \( E^a \) to represent the set of parameters for \( \text{FDNN}^s \), \( \text{FDNN}^c \) and \( \text{FDNN}^a \) respectively.

Based on our assumption, the final feature representations \( h^{fc} \) and \( h^{fa} \) for the primary and auxiliary task respectively are decomposed as a weighted sum as follows:

\[
\begin{align*}
h^{fc} &= \alpha_1 \mathbf{h}^c + \alpha_2 \mathbf{h}^s, \\
h^{fa} &= \beta_1 \mathbf{h}^a + \beta_2 \mathbf{h}^s,
\end{align*}
\]

where \( \{\alpha_1, \alpha_2\} \) and \( \{\beta_1, \beta_2\} \) are the attention weights that conceptually quantify the contributions of the shared and task-specific feature representations. Note that in this formulation, \( \mathbf{h}^c \), \( \mathbf{h}^c \) and \( \mathbf{h}^s \) are of the same dimension. As a side note, another strategy to combine the task-specific and shared feature representations is through vector concatenation \( h^{fc} = [h^c, h^s] \). But this approach could introduce more parameters for each \( h \) having enough representation power. We hence prefer the weighted sum approach when only limited amount of data is available.

To calculate \( \{\alpha_1, \alpha_2\} \) and \( \{\beta_1, \beta_2\} \), we adopt a strategy that represents the compatibility of shared and task-specific feature representations:

\[
\begin{align*}
\alpha_1 &= \frac{1}{2} \cosd(h^c, h^s), \\
\alpha_2 &= 1 - \alpha_1, \\
\beta_1 &= \frac{1}{2} \cosd(h^a, h^s), \\
\beta_2 &= 1 - \beta_1,
\end{align*}
\]

for the primary task, and

\[
\begin{align*}
\beta_1 &= \frac{1}{2} \cosd(h^c, h^a), \\
\beta_2 &= 1 - \beta_1,
\end{align*}
\]

for the auxiliary task, where \( \cosd(\cdot, \cdot) \) is the cosine-distance between two vectors \( \cosd(v_1, v_2) = v_1 \cdot v_2 / (||v_1|| ||v_2||) \).
, $|| \cdot ||_2$ is the euclidean norm of a vector. Since we use sigmoid as the activation function, $\{\alpha_1, \alpha_2\}$ and $\{\beta_1, \beta_2\}$ are positive and hence proper weights. Note that this strategy biases toward the shared feature representation and forces it to make at least half contribution (i.e. $\alpha_2, \beta_2 \geq 0.5$) to the final feature representation for ATAN enjoying benefits of multitask learning.

**Model Training** The predictions for $y^c$ and $y^a$ is calculated using the final feature representations:

$$
\hat{y}^c = W^c h^E + b^c,
\hat{y}^a = W^a h^E + b^a,
$$

where $W^c$ and $W^a$ are dimension-compatible vectors, $b^c$ and $b^a$ are bias terms.

The objective function is a weighted sum of the least squared loss functions for the primary and auxiliary tasks:

$$
\min_{E^c, E^a, W^c, W^a, b^c, b^a} \sum_{i=1}^{n} (y_{i}^c - \hat{y}_i^c)^2 + \omega(y_{i}^a - \hat{y}_i^a)^2,
$$

where the summation is over n training samples, $\omega$ is a parameter controlling the weight of the auxiliary task. We use standard gradient descent algorithm to solve Problem (8). Note that ATAN is not restricted to one auxiliary target and multiple auxiliary targets can be incorporated straightforwardly.

**C. Analyzing Weights**

Model interpretability and accuracy are both critically important in in clinic practice. While DNNs is generally difficult to interpret, we can still analyze the learned weights to see how contributions of input features propagate through the network. This approach was initially proposed in [7]. Here we adapt this method to fit the proposed ATAN model.

To keep notations uncluttered, let us take an example, as shown in Fig. 2, to see how contributions of input features to the target can be recursively calculated from the output end of DNNs. Assume the last 3 layers of FDNN are of size 2, 3 and 1 respectively; $W_1 = (w_{1i}^1)_{2 \times 3}$ and $W_2 = (w_{1i}^2)_{1 \times 3}$ are the two weight matrices between layers. Let $(g_1, g_2)$ and $(h_1, h_2, h_3)$ be the two hidden layers, $y$ the output layer.

For a hidden neuron $h_t$ ($t = 1, 2, 3$), its contribution $C_t$ to the target $y$ can be computed as in linear regression:

$$
C_{ty} = \frac{|w_{ti}^2|}{\sum_{i=1}^{n} |w_{ti}^2|}.
$$

For $g_k$ ($k = 1, 2$), its contribution $C_{kt}$ to $h_t$ is

$$
C_{kt} = \frac{|w_{tk}^1|}{\sum_{i=1}^{n} |w_{ti}^1|}.
$$

Then the contribution $Q_{kty}$ of $g_k$ through $h_t$ to the target $y$ is defined as

$$
Q_{kty} = C_{kt} C_{ty}.
$$

Summing over all hidden neurons, the total contribution $Q_{ky}$ for $g_k$ is given by

$$
Q_{ky} = \sum_{t=1}^{3} Q_{kty}.
$$

For input features $x_i$’s, we can recursively apply the strategy above to calculate the contributions.

Within the proposed ATAN model, the contribution of each input features to the primary target $y^c$ can be propagated through the task-specific network FDNN$^c$ and the shared network FDNN$^a$. Hence if we assume $Q_{ky^c}$ and $Q_{kby^c}$ are the contributions of feature $x_k$ through FDNN$^c$ and FDNN$^a$ to $y^c$ respectively, the overall contribution $Q_{ky^c}$ for $x_k$ is just the weighted sum given by

$$
Q_{ky^c} = \alpha_1 Q_{kby^c} + \alpha_2 Q_{ky^c},
$$

which provides us a heuristic approach for interpreting ATAN, $\alpha_1$ and $\alpha_2$ are given by (5).

**IV. APPLICATION**

In this section, we apply ATAN to our clinical dataset of African Americans with hypertension at-risk for LVH and compare it with various baseline models to demonstrate its effectiveness. We also analyzed ATAN with its learned parameters to rank risk factors. This provides us some insight for understanding the progression of and potential targeted early intervention for cardiovascular disease.

**A. Data Information and Preprocessing**

Our study dataset was derived from a cohort of African American patients who presented to the emergency department of a single center (Detroit Receiving Hospital) between October 2011 and November 2014 with a known history of hypertension and elevated systolic blood pressure (> 160 mm Hg). Previous studies have shown that there are disparities among hypertension patients with some who are at greater risk of LVH, yet no single variable has sufficient power for explaining the disparity. This motivates us to use a DNN model that is capable of capturing complex feature interactions.

As noted, LVH was determined using LVM as measured on CMR using a cut-point of 89 g/m$^2$ in men and 73 g/m$^2$ in women. Along with LVM, other measures from CMR are also available, such as wall thickness, left ventricular stroke volume to body surface area (LVSVI) and left ventricular end-diastolic volume indexed to body surface area (LVEDVI). These measures are clinically related to LVM and could be used as the auxiliary targets in ATAN.
For data preprocessing, we first remove samples with missing LVMI and measures whose missingness is greater than 10%. The remaining dataset consists of 155 samples and 65 measures, where 59 measures are used as features and the remaining 6 including LVMI and other CMR results are used as targets. For missing values, we use “multiple imputation chained equations” (MICE) [28] for imputation. We also standardize feature values to have zero mean and unit variance.

Table I, II and Fig. 3 show the detailed information about the used dataset and descriptive statistics of LVMI respectively.

### B. Experiment Implementation

We implement ATAN in Pytorch [21]. In ATAN, we select one CMR measure as the auxiliary target. We experiment with posterior wall thickness and LVEDVI separately for our auxiliary task. The architecture of ATAN is of 4 layers, within which 2 hidden layers are used to learn the shared and task-specific feature representations. We also restrict the dimension of hidden layers to 80 and 40 for FDNN, FDNN and FDNN for experimental convenience. ATAN is trained using standard gradient descent in conjunction with L2 regularization.

For performance comparison, we also implement various baseline models in scikit-learn library [17]. These baselines include k-nearest neighbors (KNN), random forest (RF), support vector regression (SVR), regularized linear regression (Ridge and Lasso) and also the multi-task Lasso (MTLasso). We also implement a 4-layer FNN model (MLP-4) for predicting LVMI only. The hidden size is chosen to match the feature representation level and dimension of ATAN.

The complete dataset is divided into training and testing sets by a split 125/30. For ATAN and MLP-4, we split out 90 samples from the training set to actually train the models and the remaining 25 samples is used for validation and parameter selection. For baselines, we use three-fold cross-validation on the training set for parameter selection. The evaluation metrics are finally reported on the testing set. We repeat this procedure 10 times.

For model evaluations, we use the following three metrics. Assume that $y^c = (y^c_1, \ldots, y^c_n)$ is the true vector of the primary target of $n$ samples, $\hat{y}^c = (\hat{y}^c_1, \ldots, \hat{y}^c_n)$ the vector of predicted values:

- **Mean squared error (MSE)**
  
  $$
  \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y^c_i - \hat{y}^c_i)^2.
  $$

- **Explained variance score (EVS)**
  
  $$
  \text{EVS} = 1 - \frac{\text{Var}(\hat{y}^c) - \text{Var}(y^c)}{\text{Var}(y^c)},
  $$

  where $\text{Var}(\cdot)$ is the variance.

- **Median absolute error (MAE)** is a more robust error than MSE that compute the median of absolute predictive errors:
  
  $$
  \text{MAE} = \text{Med}(|y^c - \hat{y}^c|),
  $$

  where $\text{MED}(\cdot)$ outputs the median of a vector.

### C. Results

#### Using entire feature set

We first evaluate models using all features in the hypertension data. The predictive performance on the test data is reported in Table III. In the table, ATAN-1 and multi-task lasso (MTLasso) use LVMI and LVEDVI as the tasks and ATAN-2 uses posterior wall thickness as the auxiliary task. From the table, we have following observations.

- Among all models, ATAN with LVEDVI as the auxiliary target (i.e. ATAN-1) achieves the best predictive performance and ATAN with posterior wall thickness (i.e. ATAN-2) the second best in terms of MSE, EVS and MAE. For example, ATAN-1 and ATAN-2 provide approximately 5% and 4% improvements over Lasso in MSE respectively.
By examining the predictive performance, we find that predictions often fail at the left and right tails in the sample distribution (results not shown). Previous study ([9], [11]) has discovered the correlation between LVMI and calcium metabolism and shown that patients with LVH (i.e. large LVMI values) have significant higher serum calcium level than those without LVH. For our dataset, we find that in the right tail of data distribution (LVMI value > 120), correlation between calcium and LVH is 0.79 and the two-tail correlation test is significant with \( p \)-value = 0.0004. However, correlation between calcium and LVH is 0.006 in the entire dataset, and -0.100 for LVMI < 120. This implies that LVH prevalence is different among different patient subgroups, and predictive models may fail to capture the disparities when we only have limited amount of data.

**Using demographics and lab results only** As we see from Table I, many features are functional measures of the cardiovascular system. However in practice, much previous study has focused on the relations between LVMI and lab results along with demographics, as these relations are more informative on the disease progression and often readily available in clinical practice. In this section we hence exclude features that are functional measures for the cardiovascular system and only use demographics and lab results as the input features, resulting in 34 features remain in the experiment. We follow the same experiment procedure as in the previous section and results are reported on 10 runs.

Table IV shows the performance only using lab results and demographics. We see that neural network models (MLP-4, ATAN-1 and ATAN-2) have comparable performances and are better than other methods, implying that capturing high-level information would benefit predictive modeling. Comparing
with Table III, predictive performances overall degrade. This is possibly due to that functional measures are expected to be more informative for predicting LVMI.

D. Interpreting ATAN via Analyzing Weights

Interpretability is as important as accuracy in clinical research. In this section, we use the heuristic approach proposed in Section III-C to calculate the contributions of features to targets, from which we can find risk factors for better understanding of disease progression.

Fig. 4 shows the top-20 features using the complete feature set. From the table, sex is indeed an important predictor: sample means of LVMI is 85.21 for female, and 95.78 for male; the two-sample t-test shows that the difference between female and male is significant with \( p \)-value less than 0.0001.

Aside from sex (which already is known to be a key determinant of LVMI), several features with significant contributions are cardiac measures, such as ejection duration, LV ejection fraction. This is intuitive as heart structure and function are inherently related.

Fig. 5 presents the top-15 features out of demographics and lab results. We see from the figure that both systolic and diastolic blood pressure contributes most for predicting LVMI. The relationship between hypertension and LVH was
the basic premise of this work, and the fact that elevated blood pressure corresponds with LVMI is not surprising [4]. However, our interest in this modeling exercise was to see if ATAN, could identify more subtle associations. Indeed, other contributory features from lab results were identified including renin, potassium, vitamin D, calcium, parathyroid hormone, creatinine et al. These top-ranked features accord with previous studies ([5], [9], [19]), supporting that this analysis of feature contribution through weight propagation provides a heuristically reasonable approach for interpreting DNN models.

V. DISCUSSION

In this paper, we present a novel DNN predictive model, ATAN, that functions by introducing multi-task learning as a regularization method. ATAN learns high-level latent information from low-level input features, as well as flexibly leveraging other information contained in the clinically relevant targets. Our experiments with one auxiliary target show that DNNs can offer great improvements for predictive modeling in clinical research when only limited labeled data are available. Moreover, ATAN can be easily extended to multiple auxiliary targets. However, ATAN does not exploit the information from the unlabeled data. Hence, for future work, we plan to develop DNN models combining multi-task learning paradigm with semi-supervised learning, which fully exploits different sources of information for better predictive performance.

REFERENCES