Predictive Deep Network with Leveraging Clinical Measure 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 apply ATAN in a clinical dataset of hypertension collected from a vulnerable demographic population (African-American) to demonstrate its effectiveness.**

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. Recently 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 often 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 labeled data, precluding DNN applications to various challenging, but important clinical prediction problems.

The main danger for 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 regularization methods including dropout, early stopping and L2 regularization [5]. 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.

*Corresponding author.

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, 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 and cardiac magnetic resonance imaging (CMR), is quite expensive. We recently completed an NIH funded study of African-Americans with hypertension that focused on identification of LVH using LVMI 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. 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. 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 performance with different methods.

II. Related Work

Deep neural network 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), [2] leverages longitudinal patient health history and builds a recurrent neural network (RNN) using gated recurrent unit (GRU) for early prediction of onset heart failure; [8] combines unsupervised pre-training with deep belief networks to train a classifier for decision making in healthcare. In [15], a feed-forward fully connected

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

network is proposed for gene annotation. In [10], 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.

Multi-task learning MTL for predictive modeling is a machine learning paradigm that jointly learns a model for multiple tasks [20]. 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], [5], [16] and [20] for more details on MTL.

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 ([9], [17], [21]). Particularly in bioinformatics and health informatics, [14] uses DMTL to predict protein interactions between HIV and human proteins; [18] integrates DNN feature learning into SVM-based MTL for diagnosis of Alzhermer's disease. For general considerations in DMTL, we refer to [1] and [16] for details.

III. Method

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

$$\begin{aligned} \boldsymbol{h_1} &= \sigma(\boldsymbol{W_1 x} + \boldsymbol{b_1}), \\ \boldsymbol{h_i} &= \sigma(\boldsymbol{W_i h_{i-1}} + \boldsymbol{b_i}), \quad i = 2, \cdots, k \end{aligned} \tag{1}$$

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

$$\hat{\boldsymbol{y}} = \boldsymbol{W h_k} + \boldsymbol{b},$$

where $\boldsymbol{W_i}$ and $\boldsymbol{W}$ are the weight matrices with compatible dimensions, $\boldsymbol{b_i}$ the bias terms and $\boldsymbol{h_i}$ the hidden state; $\sigma(\cdot)$ is the activation function; $\hat{\boldsymbol{y}}$ is the prediction of the network.

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. However, one pitfall with introducing auxiliary tasks is that the "relevance" is not universally defined. 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. Fig. 1 presents the overview of ATAN structure.

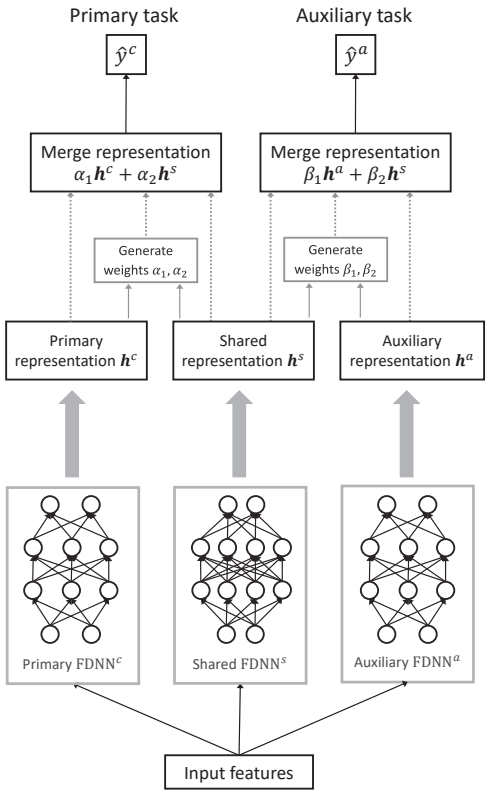

Fig. 1. Overview of ATAN with one auxiliary task.

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

$$\begin{aligned} \boldsymbol{h}^s &= \text{FDNN}^s(\boldsymbol{x}), \\ \boldsymbol{h}^c &= \text{FDNN}^c(\boldsymbol{x}), \\ \boldsymbol{h}^a &= \text{FDNN}^a(\boldsymbol{x}), \end{aligned} \tag{2}$$

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

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

$$\boldsymbol{h}^{fc} = \alpha_1 \boldsymbol{h}^c + \alpha_2 \boldsymbol{h}^s, \tag{3}$$

$$\boldsymbol{h}^{fa} = \beta_1 \boldsymbol{h}^a + \beta_2 \boldsymbol{h}^s, \tag{4}$$

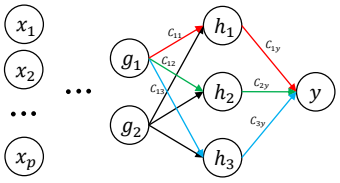

Fig. 2. A simple example showing the recursively procedure of calculating the contribution from g_1 to the target y via h_1 (red), h_2 (green) and h_3 (blue) using weight propagation [4]. The total contribution Q_{1y} from g_1 to y is $Q_{1y} = C_{11}C_{1y} + C_{12}C_{2y} + C_{13}C_{3y}$. Recursively applying this strategy can calculate the contributions of input feature x_i's.

where $\{\alpha_1, \alpha_2\}$ and $\{\beta_1, \beta_2\}$ are the weights that conceptually quantify the contributions of the shared and task-specific feature representations.

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:

$$\begin{aligned} \alpha_1 &= 0.5\text{cosd}(\boldsymbol{h}^c, \boldsymbol{h}^s), \quad \alpha_2 = 1 - \alpha_1, \\ \beta_1 &= 0.5\text{cosd}(\boldsymbol{h}^a, \boldsymbol{h}^s), \quad \beta_2 = 1 - \beta_1, \end{aligned} \tag{5}$$

where $\text{cosd}(\cdot, \cdot)$ is the cosine-distance between two vectors.

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

$$\begin{aligned} \hat{y}^c &= \boldsymbol{W}^c \boldsymbol{h}^{fc} + b^c, \\ \hat{y}^a &= \boldsymbol{W}^a \boldsymbol{h}^{fa} + b^a. \end{aligned} \tag{6}$$

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

$$\min_{E^s, E^c, E^a, \boldsymbol{W}^c, \boldsymbol{W}^a, b^c, b^a} \sum_{i=1}^{n} (y_i^c - \hat{y}_i^c)^2 + \omega (y_i^a - \hat{y}_i^a)^2, \tag{7}$$

where the summation is over n training samples, ω is a parameter controlling the weight of the auxiliary task. We use standard gradient descent algorithm to solve Problem (7).

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 [4]. 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; $\boldsymbol{W_1} = (w_{ij}^1)_{3\times 2}$ and $\boldsymbol{W_2} = (w_{ij}^2)_{1\times 3}$ 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:

$$C_{ty} = \frac{|w_{1t}^2|}{\sum_{i=1}^{3} |w_{1i}^2|}.$$

Table I. Details of hypertension dataset. LVMI is the primary target and other measures in CMR serve as the candidates for auxiliary tasks.

# Sample	155	
# Features	59	• Demographics • Lab results (calcium, eGFR et al.) • Heart functioning (LV ejection rate et al.) • Acoustic and electrocardiography (Cornell product et al.) • Applanation tonometry (Central tension time index et al.)
# CMR results	6	• LVMI • Heart wall thickness (septal, posterior and anterior) • LVSVI • LVEDVI

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

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

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

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

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^s. Hence if we assume $Q_{ky^c}^c$ and $Q_{ky^c}^s$ are the contributions of feature x_k through FDNN^c and FDNN^s to y^c respectively, the overall contribution Q_{ky^c} for x_k is just the weighted sum given by

$$Q_{ky^c} = \alpha_1 Q_{ky^c}^c + \alpha_2 Q_{ky^c}^s.$$

IV. Application

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). LVH was determined using LVMI as measured on CMR using a cut-point of 89 g/m2 in men and 73 g/m2 in women. Along with LVMI, 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 LVMI 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

Table II. Predictive performance along with standard deviations on the testing data. For MTLasso and ATAN, performance on LVMI is reported. ATAN-1 uses LVEDVI as auxiliary target and ATAN-2 uses posterior wall thickness. For MSE and MAE, smaller is better; for EVS, larger is better.

Method	MSE	EVS	MAE
KNN	254.140 (52.329)	0.230 (0.155)	11.071 (1.861)
RF	227.803 (34.387)	0.261 (0.129)	10.778 (2.284)
SVR	294.831 (67.502)	0.083 (0.015)	10.530 (2.735)
Ridge	257.496 (26.513)	0.143 (0.277)	12.464 (1.742)
Lasso	205.983 (28.081)	0.337 (0.109)	11.079 (1.975)
MTLasso	213.398 (31.025)	0.310 (0.136)	11.177 (2.015)
MLP-4	204.104 (20.869)	0.338 (0.126)	10.323 (1.789)
ATAN-1	**195.820** (25.991)	**0.372** (0.090)	**10.149** (1.392)
ATAN-2	198.320 (22.427)	0.356 (0.122)	10.185 (1.812)

Table III. Predictive performance along with standard deviations on the testing data. Only demographics and lab results are used as input features. For MSE and MAE, smaller is better; for EVS, larger is better.

Method	MSE	EVS	MAE
KNN	298.568 (51.941)	-0.011 (0.151)	10.874 (2.007)
RF	255.045 (31.077)	0.123 (0.197)	10.029 (0.826)
SVR	281.800 (44.624)	0.057 (0.014)	10.023 (0.949)
Ridge	288.446 (54.822)	0.001 (0.327)	10.350 (1.488)
Lasso	253.616 (29.377)	0.137 (0.143)	9.322 (1.048)
MTLasso	252.770 (34.322)	0.142 (0.128)	9.231 (1.186)
MLP-4	243.327 (30.834)	0.172 (0.146)	9.290 (1.668)
ATAN-1	238.585 (29.211)	0.191 (0.124)	9.017 (1.638)
ATAN-2	237.482 (31.289)	0.195 (0.127)	9.172 (1.658)

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) [19] for imputation. We also standardize feature values to have zero mean and unit variance. Table I shows the details about the used dataset.

B. Experiment Implementation

We implement ATAN in Pytorch [13]. 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^c, FDNN^s and FDNN^a for experimental convenience. ATAN is trained using standard gradient descent in conjunction with L2 regularization.

For comparison, we also implement various baseline models in scikit-learn library [11]. 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.

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 35 samples is used for validation. 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 $\boldsymbol{y}^c$ is the true vector of the primary target of n samples, $\hat{\boldsymbol{y}}^c$ the vector of predicted values:

- Mean squared error (MSE): $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i^c - \hat{y}_i^c)^2$.
- Explained variance score (EVS) computes the explained variation that a model accounts for the data:
$$\text{EVS} = 1 - \frac{\text{Var}(\boldsymbol{y}^c - \hat{\boldsymbol{y}}^c)}{\text{Var}(\boldsymbol{y}^c)}.$$
- Median absolute error (MAE): $\text{MAE} = \text{Med}(|\boldsymbol{y}^c - \hat{\boldsymbol{y}}^c|)$, 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 II. 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.

· Compared with Lasso, the vanilla DNN (*i.e.* MLP-4) has almost identical performances. Although DNN is expected to capture the complex relations among features for predicting LVMI, DNN may not efficiently learn feature representation for such small data. However, ATAN-1 and ATAN-2 that introduce auxiliary tasks based on domain knowledge bring some improvements in predictive modeling. This confirms that multi-task framework is an effective regularization approach.

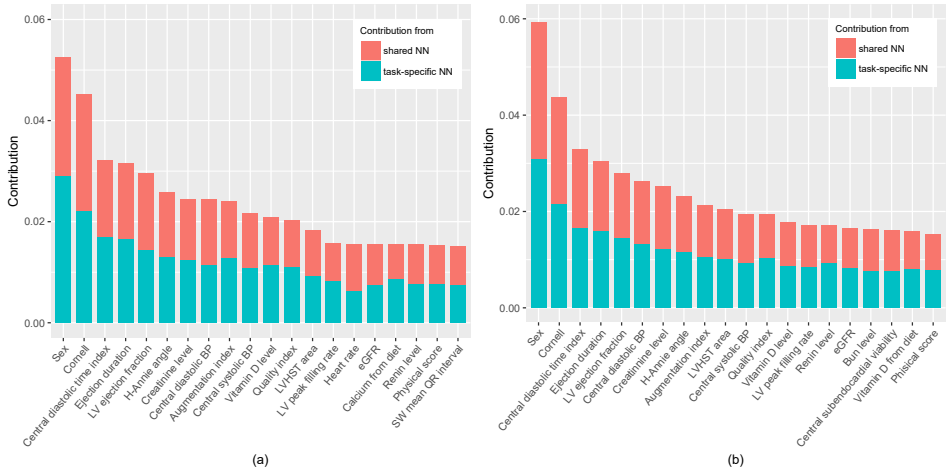

Fig. 3. Top-20 important features for the complete set of features. Auxiliary target: (a) LVEDVI (b) posterior wall thickness.

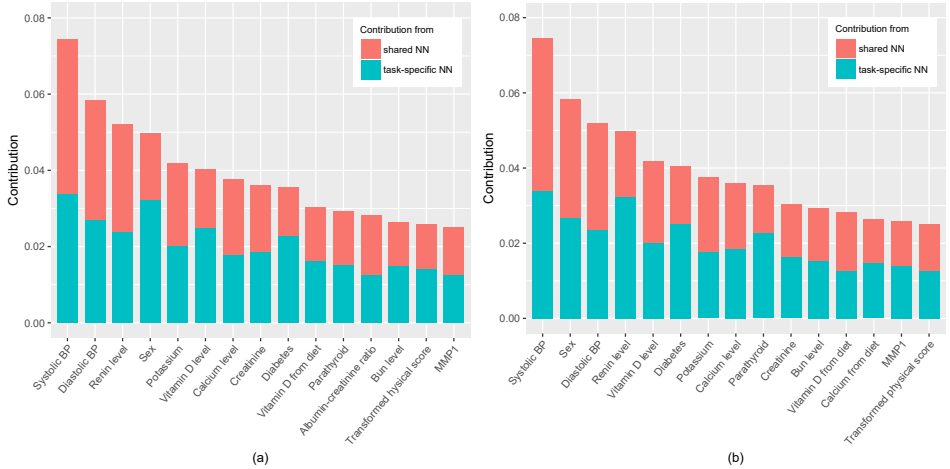

Fig. 4. Top-15 important features for only lab results as features. Auxiliary target: (a) LVEDVI (b) posterior wall thickness.

· Introducing auxiliary task may not always work. The multi-task Lasso (*i.e.* MTLasso) has worse performance than Lasso. One reason accountable for this case is that MTLasso assumes that all tasks share the same feature structure. This is however unlikely for LVMI and LVEDVI being regulated by the same set of features.

· ATAN-1 and ATAN-2 performs better than MTLasso, with margins 8% and 7% (MSE), 20% and 15% (EVS), 9% and 8.8% (MAE). This is possibly due to less restrictive assumption of ATAN for defining "relevance" between two targets: ATAN conceptually capture the "relevance" by learning feature representations shared by tasks. This implies that effectively modeling task relevance is an integral part in MTL.

By examining the predictive performance, we find that predictions often fail for large LVMI in the sample distribution (results not shown). Previous study ([6], [7]) 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 level and LVMI is 0.79 and the two-tail correlation test is significant ($p = 0.0004$). 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.

Table III 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 II, 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. 3 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. 4 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. 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 ([3], [6], [12]), 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.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. 1637312 and 1451316. Cohort collection was supported under grant NIH/NIMHD 5 R01 MD005849.

References

[1] R. Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[2] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2016.

[3] A. H. El-Gharbawy, V. S. Nadig, J. M. Kotchen, C. E. Grim, K. B. Sagar, M. Kaldunski, P. Hamet, Z. Pausova, D. Gaudet, F. Gossard, et al. Arterial pressure, left ventricular mass, and aldosterone in essential hypertension. *Hypertension*, 37(3):845–850, 2001.

[4] T. D. Gedeon. Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems*, 8(02):209–218, 1997.

[5] I. Goodfellow, Y. Bengio, and A. Courville. Deep learning, 2016.

[6] A. Helvacı, B. Çopur, and M. Adaş. Correlation between left ventricular mass index and calcium metabolism in patients with essential hypertension. *Balkan medical journal*, 30(1):85, 2013.

[7] J. Li, N. Wu, Y. Li, K. Ye, M. He, and R. Hu. Cross-sectional analysis of serum calcium levels for associations with left ventricular hypertrophy in normocalcemia individuals with type 2 diabetes. *Cardiovascular diabetology*, 14(1):43, 2015.

[8] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu. Deep learning for healthcare decision making with emrs. In *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, pages 556–559. IEEE, 2014.

[9] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang. Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *HLT-NAACL*, pages 912–921, 2015.

[10] M. Z. Nezhad, D. Zhu, X. Li, K. Yang, and P. Levy. Safs: A deep feature selection approach for precision medicine. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 501–506. IEEE, 2016.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] A. Piovesan, N. Molineri, F. Casasso, I. Emmolo, G. Ugliengo, F. Cesario, and G. Borretta. Left ventricular hypertrophy in primary hyperparathyroidism. effects of successful parathyroidectomy. *Clinical endocrinology*, 50(3):321–328, 1999.

[13] Pytorch. http://pytorch.org.

[14] Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston. Semi-supervised multi-task learning for predicting interactions between hiv-1 and human proteins. *Bioinformatics*, 26(18):i645–i652, 2010.

[15] D. Quang, Y. Chen, and X. Xie. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31(5):761–763, 2014.

[16] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[17] M. L. Seltzer and J. Droppo. Multi-task learning in deep neural networks for improved phoneme recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6965–6969. IEEE, 2013.

[18] H.-I. Suk and D. Shen. Deep learning-based feature representation for ad/mci classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 583–590. Springer, 2013.

[19] I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[20] Y. Zhang and Q. Yang. A surey on multi-task learning. *arXiv preprint arXiv:1707.08114*, 2017.

[21] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.