# Temporal Convolutional Models of Biomarkers for Disease Diagnosis

**Narges Razavian, David Sontag**

**Computer Science Department, New York University, New York City**

## 1) Introduction

Recent years have seen a dramatic increase in the amount of longitudinal health information that is being recorded about each individual. From vital signs such as heart rate and activity data recorded using mobile phones, to biomarkers that are cheaply and quickly measured from a small sample of a patient's blood, this data holds tremendous promise for precision medicine. Developing techniques to analyze this temporal data is a fundamental challenge for the field. We develop a new machine learning approach, using recent advances in deep learning, for interpreting a patient's longitudinal health data. Our algorithms take as input all of a patient's lab measurements over time, and uses Gaussian processes and temporal convolutional neural networks to jointly learn and optimize variation patterns over lab measurements that accurately diagnose multiple diseases. Our results show that the temporal convolutional models are superior to logistic regression and deep multi-layered perceptron models for the task mapping multiple measurements of biomarkers to diagnoses.

## 2) Methods

We formulate the task of diagnosis as a supervised classification task. Each training sample is composed of input X and output Y. Input X is a set of D biomarkers, where each biomarker is observed at irregular time points. For biomarker d, we have $X_d = x(t_1),\ldots, x(t_N)$ at time points $T_d = t_1,\ldots, t_N$. Output Y is matrix of size M x T encoded as binary labels of M diseases, over T time points. $Y_m = y_1,\ldots, y_T$.

### 2.1) Gaussian Process Regression

We use Gaussian Processes[1] regression, to impute the missing values of X, to convert the irregularly measured biomarker data to fully imputed matrix of size D x T. Gaussian processes regression is a semi-parametric method that assumes a Gaussian prior over the value of all possible functions that can describe a biomarker over time. For each biomarker of each individual, given the observed set of measurements $X = x(t_1),\ldots, x(t_N)$ at time points $T = t_1,\ldots,t_N$, for a new time point $t_{new}$ Gaussian processes give the posterior $P(x(t_{new})| x(t_1),\ldots, x(t_N), t_1,\ldots,t_N) \sim N(\mu_{new}, \sigma_{new})$, where $\mu_{new} = K_{tnew,T} * (K_{T,T} + 2\delta^2 I)^{-1} * X$, and $\sigma_{new} = K_{tnew,tnew} - K_{tnew,T} * (K_{T,T} + 2\delta^2 I)^{-1} * K_{T,tnew}$. K is the kernel matrix defined via a kernel function such as linear, triangular, squared exponential, etc., and $\delta$ is the prior variance of noise per observation. We selected kernel and $\delta$ parameters per lab type separately, via cross validation.

### 2.2) Temporal Convolution and Variation Pattern Discovery

Convolutional neural networks[2] are biologically inspired family of neural networks, in which a number of filters (or variation patterns) are convolved with the input, thus creating a map of where the filters have been most activated. Back-propagation[3] technique allows these filters to be automatically learned form the data in a way that the activations optimize some loss function. We used the negative log likelihood of all diseases as the loss function. Figure 1 shows the architecture of our temporal convolutional network. We formulated the task of mapping biomarkers to diseases as a time series classification, with temporal convolution in a backward window. The temporal convolution layer is shared across different diseases, and size and number of the filters are our hyperparameters. A pooling layer follows the convolution layer to reduce the number of model parameters. The length of pooling layer is another hyperparameter. The fully connected layer projects these patterns into a low dimensional space, where the projections are used to predict each disease separately. We use dropout before fully connected and disease specific layers to make the learned features robust and sparse. Our baselines include a logistic regression and a deep multilayered perceptron.

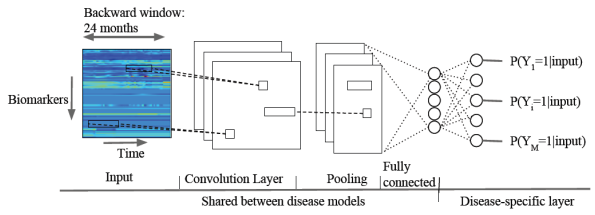


*Figure 1: Architecture for Temporal Convolutional Network*

### 2.3) Data

The source of our data is de-identified lab measurement and diagnosis history records collected by a private insurance company. The

dataset includes 71 commonly measured plasma and urine biomarkers of 30,000 individuals, each individual with at least 10 different biomarkers measured at least 4 times in the past. The dates at which the biomarkers were measured were aggregated by 1 month (taking the maximum of multiple observations over the month as input). The output was presence or absence of disease diagnosis at each month. The disease information in our dataset was encoded as ICD-9 (International Classification of Diseases, version 9). We used the 281 most commonly diagnosed disease/conditions as our target task. All individuals included in the study were required to be enrolled in the system until the last month (July 2013) when the data was collected, and to have at least one biomarker measurement on the last month. We divided the 30,000 individuals into three equally sized train, validation and test datasets. Prior to any calculations, each measurement was normalized across all nonzero measurements of all individuals for the three sets (train, validate, test), so the resulting measurements had zero mean and unit variance per biomarker.

**4) Results and Discussion**

Figure 2 shows the AUC of the diagnosis task, for each disease on the held-out test set, by different models. We only show the diseases where a significant difference was observed. Of the 149 outcomes with significant results, convolution model outperformed other models for 85 outcomes, MLP outperformed other models for 34 outcomes, and logistic regression outperformed for 11 outcomes. For some acute conditions such as pregnancy, pneumonia, and conditions involving immune system (i.e. Lymphoma cancer, HIV), we see convolution models, which specifically learn and utilize variation patterns, outperform the other models more significantly. Most chronic diseases were also observed to benefit from temporal convolution models. Examples include Chronic Kidney Disease, Anemia, Diabetes mellitus Type 2, Rheumatoid arthritis, chronic heart failure, Mitral valve disorder, impotence, among others. Figure 3 shows the learned variation patterns, and as an example, some of the parameters of the model learned for Type 2 Diabetes diagnosis.

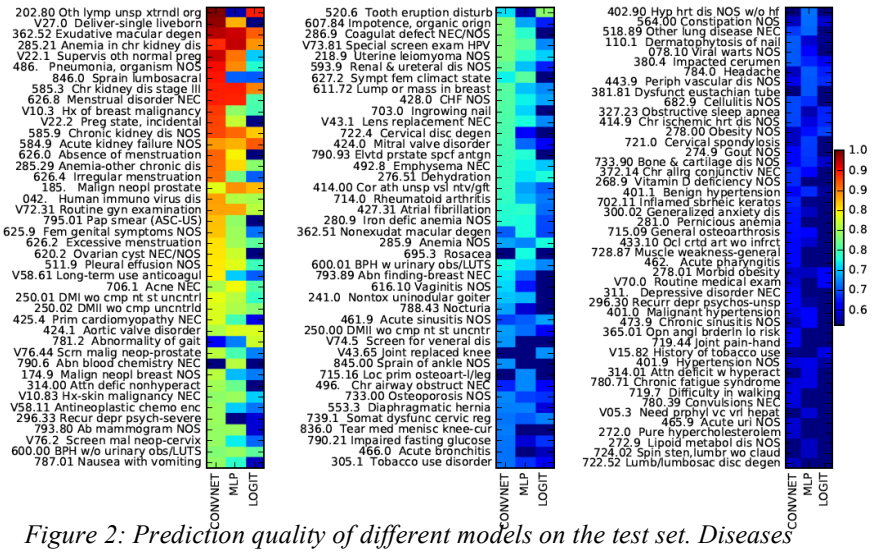


*Figure 2: Prediction quality of different models on the test set. Diseases are sorted by the highest AUC at the last iteration.*

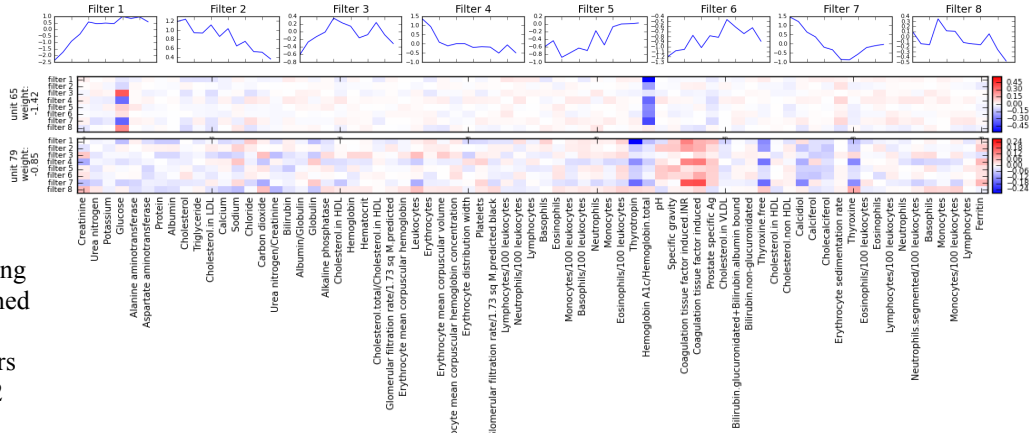


*Figure 3: Learned variation patterns (shared among all diseases), and the top variations on lab measurements predictive of Type 2 diabetes.*

**4) Conclusions**

We presented a novel and general method for data driven joint pattern discovery and prediction from irregularly measured biomarkers over time. Our results showed that not only the quality of the predictions are improved for the task of diagnosis when using convolutional models, but also the learned variation patterns can improve the quality of analysis and feature discovery for some diseases.

**References**


[1] C. E. Rasmussen, *Gaussian processes for machine learning,* 2006.

[2] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition Proceedings of the IEEE, 1998.

[3] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating errors Cognitive modeling 1986