# Contemporaneous Health Monitoring and Biomarker Discovery by Integration of Patient Data and Disease Knowledge

Banafsheh Samareh Abolhasani

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Shuai Huang, Chair

Linda Ng Boyle

Archis Ghate

Program Authorized to Offer Degree:

Industrial & Systems Engineering

University of Washington

## Abstract

Contemporaneous Health Monitoring and Biomarker Discovery
by Integration of Patient Data and Disease Knowledge

Banafsheh Samareh Abolhasani

Chair of the Supervisory Committee:

Professor Shuai Huang

Industrial & Systems Engineering

Technological innovations have given rise to data-rich environments that support the use of heterogeneous sensor measurements to monitor complex healthcare systems. Despite these advancements, however, there remains little understanding of how patient health evolves in real clinical settings and how changes in health condition generate manifestations that are captured by the data. To address this knowledge gap, my research aims to build disease trajectory modeling that can reconstruct the evolving patient's health condition over time, termed the contemporaneous health index (CHI), by combining data and the natural history model of the disease. This global index may help increase the continuity of care, facilitate patient-provider communication, and assist with a range of clinical decision makings. However, lack of deep understanding of the disease and its progression, existence of patient heterogeneity, inherent uncertainty in predictive models and the emergence of large amounts of complex, and unstructured data, all could compromise prediction capabilities.

In this dissertation, we developed innovative methodologies that reflect the progression of the underlying patient condition by learning personalized models, quantifying the uncertainty of

those models, and creating effective biomarker engineering pipelines to analyze large amounts of complex data for effective incorporation into the calculation of CHI. We first, proposed a novel clinical data fusion framework, named DL-CHI: a dictionary learning-based CHI that quantifies the severity of the deterioration process over time and represents monotonic progression patterns with a systematic optimization formulation. DL-CHI mitigated the heterogeneity of the patients by incorporating dictionary learning to create personalized models for individual patients. We then developed the UQ-CHI framework: an uncertainty quantification-based model of CHI to further enhance the disease trajectory modeling with uncertainty quantification by considering imperfect and continuous delivery of knowledge via probabilistic nature of maximum entropy discrimination (MED) principle. Finally, we proposed effective biomarker engineering pipelines to enable possible extensions of the CHI for building trajectory models from complex data (video, audio, text, and mobile sensor reading data).

We applied the proposed methodologies to real-world applications, including Alzheimer's disease (AD), surgical site infection (SSI), depression and human activity recognition using wearable sensors.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Professor Shuai Huang, for the guidance and support throughout my Ph.D. journey. His insightful guidance and approaches helped me solving challenging research problems. Professor Huang has been supportive and has given me the chance to pursue various projects and helped my transition from an individual to a collaborative mindset – Thank you for giving me the opportunity to work with and learn. In addition, I would like to extend my sincere gratitude to my committee members, Professor Linda Boyle, and Professor Archis Ghate for their continued support – I appreciate your insightful comments and encouragement and untiring support throughout my journey.

I would also like to thank our collaborator, Dr. Heather Evans and Dr.Bill Lober for the amazing opportunity to work on the assessing surgical site infection surveillance technologies project – Thank you for letting me be part of your amazing team and helping me closely work with health professionals.

I am very grateful to the ISE department faculty & staff and all of those who helped me towards the successful completion of my Ph.D. studies. I would also like to thank my fellow friends at B14 graduate lab. I have been very lucky to work and study in such a warm and supportive environment – Thank you for your friendship and support.

To Sara – Thank you for encouraging me to apply to ISE-UW. You supported me to keep pursuing my passion no matter what.

To Jamshid and Tania — Thank you for helping me in numerous ways during various stages

of my life and being there whenever I needed a family.

Last but not least, I want to thank Farzaneh and Bardia — Thank you for encouraging me in all of my pursuits and inspiring me to follow my dreams. Bardia, I can't thank you enough for encouraging me throughout my life journeys. Farzaneh, you have made countless sacrifices to help me get to this point. Without you, I doubt that I would be in this place today. I'm forever in your debt.

# DEDICATION

To Masoud, Farzaneh, and Bardia.

# Chapter 1

# INTRODUCTION

## *1.1 Motivation*

The emergence of big data era in routine health care provides an unprecedented opportunity for patient monitoring and characterization of clinical conditions. Current knowledge of disease comes from clinical experience and laboratory experimentation, but a systematic collection of observations from patients remains rare. This has led to a limited understanding of how the disease evolves in real clinical settings and has hampered the true power of big data analytics in the health sciences. Healthcare system characteristics present challenges that call for specialized techniques capable of integrating data, extracting intelligent knowledge, and coordinate monitoring by multiple sensors. Among the many challenges in using these methods, is an incomplete understanding of the disease that could ultimately contribute to model development (e.g., monotonic progression) that fully leverages complex new types of data (e.g., text, image, audio, video, mobile sensor reading data, etc) and readily handles the inherent complexity/variety in clinical decision support (e.g., uncertainty). Some recent developments in statistics and machine learning, including dictionary learning, Bayesian learning, deep learning, as well as some traditional methodologies (e.g., multi-level models), hold potential in addressing these challenges in a statistically powerful and computationally efficient way. This dissertation concentrates on developing novel statistical methods for helping overcome these challenges. In the following, we illustrate the objectives and organization of this dissertation.

## *1.2 Research objectives*

### *1.2.1 Disease trajectory modeling and contemporaneous health monitoring*

We developed novel formulations for contemporaneous patient risk monitoring by exploiting emerging data-rich environments, where longitudinal data that reflects the degeneration of the health condition could be continuously collected. Our formulations considered the existence of patient heterogeneity and uncertainty as two unavoidable features that could affect model prediction capabilities.

***DL-CHI: a dictionary learning-based CHI model to address heterogeneity***

It is recognized that the heterogeneity of the patient population (even of the same diagnostic group) can play a significant role in clinical outcomes [2], for reasons such as the severity of the disease, responsiveness to the care, or the presence of multiple comorbid conditions. Although it is believed that a personalized model can be built for each patient according to their specific data, such models require a significant amount of labeled training samples, which is generally not available in clinical settings. Furthermore, multivariate longitudinal clinical measurements are temporal realizations of an underlying disease progression, therefore it is challenging to model this data to ensure robust predictions. The fact that these data are usually sampled at irregular time points introduces an additional layer of complexity. Towards this goal, we developed a knowledge-driven contemporaneous health index DL-CHI, which represents an index for degenerative disease monitoring [3]. This index precisely reflects the underlying patient condition across the course of the condition's progression and accounts for disease complexity and patient heterogeneity by drawing on recent theoretical developments in statistics and machine learning.

***UQ-CHI: an uncertainty quantification-based model of CHI to quantify imperfect and continuous delivery of knowledge***

A degree of uncertainty is always involved in decision-support systems. This uncertainty can be traced back to sample errors, system noise, or insufficient sample size that is encountered during the modeling process. In clinical predictions, it is necessary to deal with

such uncertainty in an effective manner. If the model parameters are not well constrained, the resulting predictions will likely represent an unacceptable degree of uncertainty that would lead healthcare professionals and caregivers to make uninformed, and potentially erroneous, medical decisions. To address this concern, we developed a methodology to further enhance the disease trajectory modeling with uncertainty quantification named UQ-CHI: an uncertainty quantification-based contemporaneous health index for degenerative disease monitoring [4]. This framework informs optimal decision-making given imperfect and continuous delivery of knowledge and can ultimately lay the foundation for further decision-making such as prediction and monitoring.

### *1.2.2 Biomarker engineering for data-rich healthcare environments*

Many biomarker measurement techniques can be used to represent the progression of complex disease, however, they have to be in a form that allows algorithms for effective modeling. Building trajectory models from complex data (video, audio, text, and mobile sensor reading data) require advanced data-driven approaches that integrate multiple sources and draw meaningful conclusions that support decision-making. Therefore, we developed biomarker engineering pipelines by using domain knowledge of the data to serve as a benchmark for building CHI models. These pipelines enable possible extensions of CHI to other healthcare applications such as mental health monitoring and human activity recognition.

***Multi-modal biomarker engineering to detect depression***

We examined how domain knowledge of the disease and data could be used to create biomarkers that make machine learning models more accurate and useful. Particular healthcare system characteristics call for specialized data fusion models to predict patient's conditions, address data-rich environments, and extract intelligent knowledge from multiple, interacting sensors. For example, for complex diseases such as depression, a single sensor is typically insufficient for characterizing various types of manifestations. Studies show that depression can manifest in multiple modalities such as a patient's vocal acoustic, linguistic and facial patterns [5, 6, 7]. Hence, the explicit modeling of heterogeneous signals provides a much better

characterization of the condition of a system compared to relying solely on measures from an individual signal. We previously proposed an application scenario of a multi-modality methodology-based prediction, intending to discover promising biomarkers that were predictive of depression for better treatment evaluation and monitoring [8, 9]. Here, we apply a multi-modality prediction model to combine the audio, video, and text modalities.

***Unknown context discovery through an uncertainty quantification-based of deep neural networks***

With the knowledge-driven contemporaneous health index, warning signals occurring throughout disease progression can be automatically sent to the clinic team via integration of information from the sensor reading data (e.g., medical sensor, mobile device, and wearables). A healthcare application may bring about severe network traffic as patients, biomedical signals, sampling rates all increase. Using contextual information to characterize the situation of patients (i.e., location, physical condition, disease type, etc.) may facilitate the clinic team to provide better treatment for the patient [10, 11], and might be a solution to solve the interoperability problem linking sensor makers to healthcare service providers. However, contextual information can vary from person to person, and it can change over time for the same person. Therefore, we require approaches to generate health indices from sensor reading data that are able to effectively reveal unknown contexts.

Towards this goal, we focused on context-aware systems as a part of wearable computing to improve activity recognition and to detect useful digital biomarkers that are predictive or indicative of patient's conditions. We developed a deep context-aware mobile activity recognition system that detected the uncertainty of possible unknown contexts. This was accomplished by determining the potential distribution mismatch between known and unknown contexts by combining the feature extraction power of deep learning with the learning power of maximum entropy learning (MEL) to define a probabilistic mechanism for unknown context discovery.

### *1.3 Organization of the dissertation*

This dissertation is organized according to the following sections: the first part of this dissertation focuses on the disease trajectory modeling of DL-CHI: a dictionary learning-based contemporaneous health index for degenerative disease monitoring (Chapter 2). We then developed a probabilistic generalization of CHI, UQ-CHI: an uncertainty quantification-based contemporaneous health index for degenerative disease monitoring (Chapter 3). Next, we developed effective pipelines for complex data by presenting a multi-modality framework to detect depression from communication, specifically the linking of computer vision, signal processing, and sentiment analysis (Chapter 4). Finally, we developed an uncertainty quantification-based mixture of deep neural networks for context-aware activity recognition to enable CHI for sensor reading data (Chapter 5). The conclusion of the dissertation is presented in Chapter 6.

Chapter 2

# DL-CHI: A DICTIONARY LEARNING BASED CONTEMPORANEOUS HEALTH INDEX FOR DEGENERATIVE DISEASE MONITORING

Effective monitoring of degenerative patient conditions is crucial for many clinical decision-making problems. Leveraging the nowadays data-rich environments in many clinical settings, in this work we propose a novel clinical data fusion framework that can build a contemporaneous health index (CHI) for degenerative disease monitoring to quantify the severity of deterioration process over time. Our framework specifically exploits the monotonic progression patterns of the target degenerative disease conditions such as the Alzheimer's disease (AD) and articulate these patterns with a systematic optimization formulation. Further, to address the patient heterogeneity, we integrate CHI with dictionary learning to build sets of overcomplete bases to represent the personalized models efficiently. Numerical performances on two real-world applications show the promising capability of the proposed DL-CHI model.

## *2.1 Introduction*

In this work, we concern the problem of patient risk monitoring that is to characterize the trajectory over the course of progression. Although there is no universal definition of the concept "patient condition", it has been a crucial concept in the communications between clinicians and frequently referenced by healthcare providers. Developing a precise contemporaneous longitudinal index (CHI) that can faithfully reflect the underlying patient condition across the course of the condition's progression holds great value for facilitating a range of clinical decision-makings. For instance, it will help early detection of patient

deterioration to help reduce the number of serious incidents, i.e., it is reported that 11% of serious incidents are a function of deterioration not acted upon mainly due to the failure to recognize the sign of deterioration [12, 13]. It will also help enhance the continuity of care since a longitudinal perspective of the patient condition can be provided for clinicians and healthcare providers. Also, it may ultimately lead to development of control system engineering that can implement adaptive interventions for better healthcare management [14, 15, 16, 17], with a global representation of the dynamic condition in evolution.

Towards this goal, technological innovations are emerging in many healthcare applications, which have given rise to a data-rich environment where an abundance of longitudinal clinical measurements that reflect the degeneration of the health condition can be continuously collected. For example, to monitor the surgical site infection (SSI), daily wound measurements, such as the temperature, granularity, distance of the wound, could be acquired to assess the condition of the wound, together with other non-wound related but important clinical signals such as heart rate, morning body temperature, and NG tube presence, etc. However, particular data characteristics present challenges that call for specialized data fusion models to predict patient conditions using the multivariate longitudinal data. For instance, as these multivariate longitudinal data are actually temporal realizations of an underlying disease progression in different dimensions, how to leverage our knowledge of the disease progression process to fuse the data is a challenge. Also, the fact that these data are usually sampled at irregular time points adds in another layer of complexity. And even we could fuse the data properly, the existence of patient heterogeneity multiplies the complexity of the problem that calls for a generic framework to personalize the model based on individual's characteristics implicitly embedded in data.

To tackle those challenges, we propose a novel framework, named as DL-CHI that focuses on a particular category of disease conditions that follow a monotonic disease progression process. In our previous work [1], we have developed a Contemporaneous health index (CHI) that fuses the irregular multivariate longitudinal time series data to quantify the severity of degenerative

disease conditions to fit the monotonic degradation process of the disease condition. However, CHI is designed for average user and ignores the patient heterogeneity, and therefore limits their applicability in real-world applications. For example, it is known that patients of AD suffer from very diverse and heterogeneous progression processes [18, 19, 20]. A possible remedy is to build personalized model on an individual's basis. However, this demands a great amount of labeled training samples, which are very likely not feasible in many clinical settings.

Thus, this motivates us to develop the DL-CHI framework by integrating CHI with dictionary learning [21, 18]. The basic idea shared by the dictionary learning algorithms is that the input signal is approximated with a sparse linear combination of a few dictionary elements or basis [22]. DL has been used in many signal processing applications, such as signal reconstruction [23], face recognition [24], and healthcare [25, 26]. The dictionary basis provides a succinct representation that can span the space of the personalized models to capture the patient heterogeneity and reveal the hidden structures in the data (in a similar spirit as principal component analysis). It has been shown that the performance of a classification task can be improved by learning a sparsifying dictionary from the data set. [27, 28]. The reason is that the sparsifying dictionary actually plays a role in the regularization of the model learning, as the dictionary basis vectors are numerical representations of patient heterogeneity. Translating this wisdom into DL-CHI, our basic idea is to first learn individual models through the CHI formulation, and then, reconstruct the model parameters of the learned individual models via supervised dictionary learning. Each column of the dictionary represents a basis vector. As such, each individual model is represented as a sparse linear combination of the basis vectors.

The work is organized as follows. In Section 2, related work in the literature will be reviewed and discussed. In Section 3, the proposed analytic framework will be presented, and the corresponding computational algorithm will be derived. In Section 4, the proposed method will be implemented and validated using two real-world applications; one is for monitoring of

brain health in AD and the other is monitoring of SSI. We will conclude the study in Section 5. Note that, in this work, we use lower-case letters, e.g., $x$, to represent scalars, bold-face lower-case letters, e.g., $\boldsymbol{v}$, to represent vectors, and bold-face upper-case letters, e.g., $\boldsymbol{W}$, to represent matrices.

## *2.2 Related works*

### *2.2.1 The CHI model*

The CHI model is developed in [1] which specifically utilizes the monotonicity of disease progression to enhance the data fusion of multivariate clinical measurements taken at irregular time points. In this section, we will first briefly present the basic formulation of the CHI model, and then, present the DL-CHI model that integrates CHI with dictionary learning for personalized models.

The CHI model was motivated by the common characteristics of many degenerative conditions such as AD which shows monotonic progression trajectory. For example, for AD, a number of biomarkers have been developed to measure the degeneration of the neural systems, including the neuroimaging modalities such as PET and MRI scans [29, 30]. It is typical to see that, along with the disease progression, the brain volumes shown in the MRI scans continues to shrink over time. The same phenomenon could be observed on the PET scans with the persistent decrease of metabolic activities. Those monotonic patterns indicate that the disease progression, once started, tends to be worse and worse.

The task of CHI is to translate multivariate longitudinal measurements into a contemporaneous health index $h_{n,t}$ that captures patient condition changing over the course of progression. Note that different individuals could be measured with different length of time and at different time locations. As we target degenerative conditions, CHI should be monotonic, i.e., $h_{n,t_1} \geq h_{n,t_2}$ if $t_1 \geq t_2$, if we assume that higher index represents more severe condition. Since CHI is a latent construct that is not directly measurable, clinical variables associated with it can be measured over time, which provide us data to learn it.

Denote the training set by $\mathbf{x}_{n,t} = [x_{n,1,t}, ..., x_{n,d,t}]^T \in \mathbb{R}^d$ collected from $N$ patients. Here, each measurement $x_{n,i,t}$ is the value of the $i$th variable for the $n$th subject for a given time $t$, where $t \in \{1, \ldots, T_n\}$ is the time index. Converting the measurements $\mathbf{x}_{n,t}$ into $h_{n,t}$ needs a mathematical model for $h_{n,t} = f(\mathbf{x}_{n,t})$. Here, for simplicity and interpretability, we start with the linear models, i.e., $h_{n,t} = \mathbf{x}_{n,t} \cdot \mathbf{w}$, where $\mathbf{w} \in R^d$ is a vector of weight coefficients to combine the $d$ variables. Denote the total number of positive and negative samples by $N^+$ and $N^-$ respectively, i.e., $N^+ := |\{n|y_n = 1\}|$ and $N^- := |\{n|y_n = -1\}|$.

The formulation of the CHI learning framework is shown in below:

$$
\begin{aligned}
\min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + && \text{(2.1a)}\\
& \beta \sum_{n\in\{1,\cdots,N\}} \max(0, 1 - y_n(\mathbf{x}_{n,T_n}^\top \mathbf{w} + b)) + && \text{(2.1b)}\\
& \alpha \sum_{\substack{n\in\{1,\cdots,N\}\\ t\in\{1,\cdots,T_n-1\}}} \max(0, 1 - \mathbf{z}_{n,t}^\top \mathbf{w}) + && \text{(2.1c)}\\
& \frac{\lambda}{2}\left(\frac{1}{N^+}\sum_{n\in\{N^+|y_n=1\}} ((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^+)^T\mathbf{w})^2\right) + && \text{(2.1d)}\\
& \frac{\lambda}{2}\left(\frac{1}{N^-}\sum_{n\in\{N^-|y_n=-1\}} ((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^-)^T\mathbf{w})^2\right) && \text{(2.1e)}\\
& + \gamma\|\mathbf{w}\|_1. && \text{(2.1f)}
\end{aligned}
$$

Items in (3.1) can be explained as follows:

- The first term (3.1a) and second term (3.1b) is the SVM formulation that aims to utilize the label information to enhance the discriminatory power of CHI. Here, $y_n \in \{1, -1\}$ is the label of the $n$th sample that indicates if the $n$th subject is diseased or not.
- The term (3.1c) is invented to enforce the monotonicity of the learned health index, i.e., $h_{n,t_1} \geq h_{n,t_2}$ if $t_1 \geq t_2$. Here, $\mathbf{z}_{n,t}$ is the difference of two successive data vectors

$\mathbf{z}_{n,t} := \mathbf{x}_{n,t+1} - \mathbf{x}_{n,t}$.

- Items (3.1d) and (3.1f) are invented to encourage the homogeneity of CHI within the group that has the same health status. Here, $\bar{\mathbf{x}}^{+}_{T_n}$ and $\bar{\mathbf{x}}^{-}_{T_n}$ represent the center of data vectors at time $T_n$ for all positive and negative samples, respectively, that are,

$$\bar{\mathbf{x}}^{+}_{T_n} := \frac{1}{N^+} \sum_{n \in \{n|y_n=1\}} \mathbf{x}_{n,T_n}$$
$$\bar{\mathbf{x}}^{-}_{T_n} := \frac{1}{N^-} \sum_{n \in \{n|y_n=-1\}} \mathbf{x}_{n,T_n}.$$

- The last term, (3.1f), is the $L_1$-norm penalty that is used to encourage sparsity of the features.

Note that the proposed formulation generalized many existing models, such as SVM, sparse SVM, LASSO, etc. The CHI model could be efficiently solved by using the block coordinate descent algorithm that is illustrated in Appendix A.1.

### *2.2.2 Dictionary learning*

Developing models like CHI helps us to capture changes in various aspects of the disease trajectory. But as CHI assumes the same model for the whole population, it ignores heterogeneity of degenerative diseases and therefore limits its applicability in real-world applications that have shown great patient heterogeneity [31, 32]. Recently, it has been shown that learning a dictionary can overcome the above limitations [33, 24, 34]. The basic idea of dictionary learning algorithms is to approximate training samples as a sparse linear combination of the few dictionary elements. Hence, dictionary learning algorithm can be considered as a way to represent low-dimensional structure of high-dimensional data.

DL was applied to many applications and achieved state-of-the-art performances, such as image denoising [23] and inpainting [35], clustering [36, 37], classification [38, 39] etc. It is known that the conventional DL framework was designed for a reconstruction task instead of adapting to classification. It is believed that classification performance will be further

improved if we carefully learn a classification oriented dictionary. For instance, in [22] a sparse representation based classification (SRC) method was proposed for robust face recognition, and achieved very impressive results. SRC treats the original data set as a dictionary, wherein the class-specific training sets are sub-dictionaries contributing to discrimination. Inspired by SRC, Yang et al. proposed a meta-face learning [24] to learn an adaptive dictionary for each class, and Ramirez et al. [27] added another term to derive more delicate classification-oriented dictionaries.

The use of dictionary learning for personalization of prediction models is also achieved by proposing novel transfer learning approaches. For example, in [17] personalization task was performed in two phases; learning user-specific source classifiers, and learning a distribution-to-classifier mapping via implementing dictionary learning. Another approach is to perform multi-modal task-driven dictionary learning algorithm under the joint sparsity constraint to enforce collaborations among multiple homogeneous/heterogeneous sources of information. In task driven formulation, the multi-modal dictionaries are learned simultaneously with their corresponding classifiers. The resulting multi-modal dictionaries can generate discriminative sparse codes from the data that are optimized for a given task such as binary or multi-class classification [40].

There are various dictionary learning algorithms that are effective for classification tasks [41, 42, 43, 44]. Zhang and Li proposed discriminative K-SVD to simultaneously achieve a dictionary which has a good representation power while supporting optimal discrimination of the classes [43]. The name K-SVD refers to updating a dictionary with $K$ vectors. A collection of training vectors corresponding to the dictionary vector in its approximation are taken by minimizing the Frobenius norm of the approximation error by solving for the dictionary vector at each iteration. This algorithm starts with an initial dictionary and initial sparse code coefficients and then one dictionary vector is updated at each iteration. The corresponding sparse coefficient is changed before proceeding to update the next dictionary vector. The minimization is done through singular value decomposition (SVD). Another example is the

iterative least squares dictionary learning algorithms (ILS-DLA) presented in [41, 42], where assumes known sparse code coefficients at each iteration and derives the best possible dictionary using either the orthogonal matching pursuit (OMP) or Focal Under-determined System Solver (FOCUSS). ILS-DLA method deploys a second order update which makes it nearly impractical in reasonable dimensions due to its matrix inversion step. Another example is the recursive least squares dictionary learning algorithm RLS-DLA, which is an online alternation of ILS-DLA. In the online alternation each training signal is processed one at a time to improve the dictionary. One of the larger challenges with ILS-DLA and K-SVD is to find a good initial dictionary. The online nature of RLS-DLA prevents getting stuck in a local minimum close to the initial dictionary contrary to the K-SVD and ILS-DLA. RLS-DLA uses the forgetting factor to improve the convergence properties of the algorithm, and hence makes the algorithm less dependent on the initial dictionary. However, RLS-DLA method requires to permute the order of training vectors and adapt the forgetting factor to satisfy the randomness and convergence properties of the online nature of the algorithm.

There are several properties that should be considered in the search for a successful dictionary training algorithm. *Flexibility*: The algorithm should be flexible enough to run with various sparse approximation algorithm such as pursuit algorithm which involves finding the best projections of input signal onto the span of an overcomplete dictionary $\mathbf{D}$. The flexibility property would enable different choices in favor of run-time constraints. Usually methods that are flexible enough would separate the dictionary updates with sparse coding stage. *Adaptivity*: An overcomplete dictionary $\mathbf{D}$ either can be chosen as a pre-determined set of functions, or designed to iteratively getting updated to better fit the data. Choosing a pre-specified dictionary is appealing because it is simpler and may lead to a fast algorithm. However, the dictionary that leads to the best representation for each member in this set, under strict sparsity constraints is needed. Such dictionaries have the potential to outperform the commonly used pre-specified dictionaries. *Efficiency*: A dictionary learning algorithm should lead to a numerically efficient and fast convergence. For example, ILS-DLA, has a

second-order update which makes it nearly impractical in reasonable dimensions.

K-SVD algorithm is flexible and works with any pursuit algorithm. In addition, it leads to the best representation for each training vector. Given the merits of DL in overcoming heterogeneity of models, and the classification performance, here we used the idea of DL and developed the DL-CHI framework using K-SVD dictionary learning algorithm. Therefore, we reconstructed our model parameter of each individual sample to be linear combination of dictionary elements. We further, compared our methodology with CHI, and other dictionary models K-SVD, ILS-DLA and RLS-DLA. Note, that DL-CHI formulation is personalized and not designed for average users unlike the above methods.

## *2.3 The proposed DL-CHI model*

### *2.3.1 Rational and formulation*

To extend CHI for personalized models, our approach is built on the dictionary learning framework [45]. As we have mentioned, the dictionary learning aims to identify a set of representative vectors that could characterize the low-dimensional structure embedded in a high-dimensional vector space [46, 47, 48]. Particularly, here, taking the model parameter vectors of all the individuals as the high-dimensional vector space, we seek a dictionary to represent these model parameter vectors. The dictionary will be learned from data, and it helps regularize the learning of the models since it requires the model parameter vectors to be (sparse) linear combination of the dictionary bases. The whole pipeline of this DL-CHI model is shown in Figure 2.3.1.

From this point of view, the Dictionary learning could be viewed as a trade-off made between two extremes. In one extreme, there is only one model for all the individuals, i.e., the "one-size-fits-all" model. On the other extreme, there is one distinct model for all the individuals and these models are all independent with each other. As a trade-off, dictionary learning exploits the dependency and difference of the individuals simultaneously.

To fulfill this idea, here, we denote the set of model parameter vectors of all the individuals as

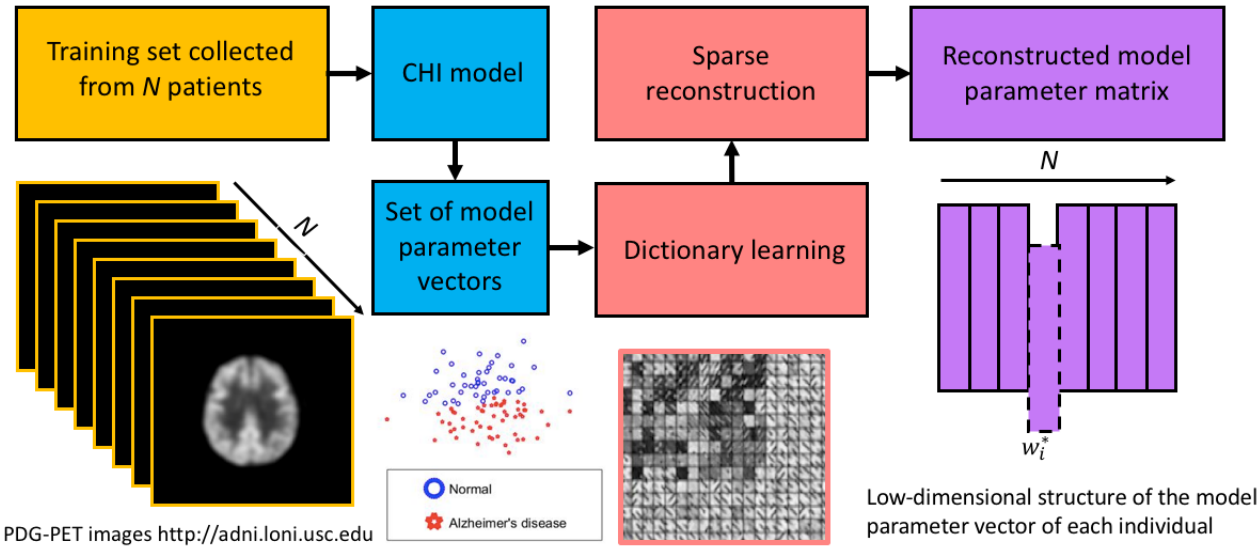



Figure 2.3.1: A conceptual overview of the DL-CHI method

$\mathbf{W}^* = [\mathbf{w}_1^* \ldots, \mathbf{w}_i^*, \ldots, \mathbf{w}_N^*]$, where $\mathbf{w}_i^*$ represents weight coefficient vector of the $i^{th}$ patient learned from the CHI model. Using dictionary learning, we aim to find an overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{d \times k}$ that contains $k$ independent columns referred as the basis vectors, $\{d_i\}_{i=1}^k$. A model parameter vector $\mathbf{w}^*$ can be represented as a linear combination of these basis vectors, satisfying the approximation condition $\mathbf{w}^* \approx \mathbf{Da}$, where $\mathbf{a}$ is the coefficient vector which can be considered as the representation of $\mathbf{w}^*$ over the dictionary $\mathbf{D}$.

In order for $\mathbf{D}$ to be flexible and robust to noise, we set the dictionary to be overcomplete ($k > d$). On the other hand, given any $\mathbf{w}^*$ with a overcomplete dictionary, we need to find the smallest set of basis vectors from the dictionary to represent it. When we set the dictionary to be overcomplete, an infinite number of solutions are available for the representation, hence constraints on the solution must be set. The solution with the fewest number of nonzero coefficients in $\mathbf{a}$ to represent $\mathbf{w}^*$ is certainly an appealing representation. This strategy is called sparse coding that is often used in dictionary learning representations. In this setting, sparse coding amounts to computing the following:

$$\begin{aligned}\{\mathbf{A},\mathbf{D}\} &= \min_{\mathbf{D},\mathbf{A}} \sum_{i}^{N} \|\mathbf{w_i^*} - \mathbf{D}\mathbf{a_i}\|_2^2 + \lambda_1 \|\mathbf{a}_i\|_0 \\ &= \min_{\mathbf{D},\mathbf{A}} \|\mathbf{W}^* - \mathbf{D}\mathbf{A}\|_2^2 + \lambda_1 \|\mathbf{A}\|_0\end{aligned} \tag{2.2}$$

Here, $\|\cdot\|_0$ is the $l^0$ norm, counting the nonzero entries of a vector, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ is the coefficient matrix of the sparse decomposition. In order to achieve sparse representations given a set of training vectors, we adapt a dictionary that leads to the best representation for each vector in this training set, under strict sparsity constraints.

### *2.3.2 Computational algorithm*

In DL-CHI, we used the K-SVD dictionary learning algorithm [49, 50] for sparse representation as an optimization problem, which can be efficiently solved via orthogonal matching pursuit (OMP) and singular value decomposition (SVD). The K-SVD approach is an iterative procedure that consists of two steps, and both steps in the algorithm are coherent with each other; working towards the minimization of the overall objective function.

First, we considered the sparse coding stage where we assumed that $\mathbf{D}$ was fixed and considered the optimization problem in (2.2) as a search for sparse representation with coefficients summarized in the matrix $\mathbf{A}$. The sparsity term of the constraint was relaxed so that the number of nonzero entries of each column $\mathbf{a}_i$ could be more than 1 and less than a number $T_0$. In doing so, the relaxed objective function becomes:

$$\min_{\mathbf{a}_i} \|\mathbf{w_i^*} - \mathbf{D}\mathbf{a_i}\|_2^2 \quad \text{s.t.} \quad \forall i, \|\mathbf{a}_i\|_0 \leq T_0, i = 1, 2, ..., N \tag{2.3}$$

In (2.3) $\mathbf{D}$ was first fixed such that we could focus on learning the coefficient matrix $\mathbf{A}$ using the orthogonal matching pursuit method, as long as it could supply a solution with a fixed and predetermined number of nonzero entries $T_0$. OMP is an iterative greedy algorithm that selects the column best correlated with the residual part of the signal, and represents the sub-optimal solution to the problem of sparse signal representation. The major advantage

of the OMP is its simplicity and fast implementation. The problem in (2.3) consists of $N$ distinct problems.

With a learned $\mathbf{A}$, we searched for the best dictionary $\mathbf{D}$. The search process is to update only one column of the dictionary, $\mathbf{d}_k$, at each time corresponding to $i^{th}$ row in $\mathbf{A}$, denoted as $\mathbf{a}_T^j$ (this is not the vector $\mathbf{a}_i$ which is the $i^{th}$ column in $\mathbf{A}$). The process of updating only one column of $\mathbf{D}$ at a time has a straightforward solution based on the singular value decomposition (SVD). The problem becomes looking only at the training vectors that uses only one column of the dictionary vector in its approximation, minimizing the approximation error $\mathbf{E}_k$. The matrix $\mathbf{E}_k = \mathbf{W}^* - \sum_{j \neq k}^{k} \mathbf{d}_j \mathbf{a}_T^j$ stands for the error for all the training samples when the $k$th basis is removed, and $\mathbf{a}_T^k$ is the $k$th row in $\mathbf{A}$. The SVD finds the closest rank-1 matrix (in Frobenius norm) that approximates $\mathbf{E}_k$. Hence, we re-wrote the penalty term in (2.3) as:

$$\sum_{i}^{N} \|\mathbf{w_i^*} - \mathbf{D a_i}\|_2^2 = ||\mathbf{W}^* - \mathbf{DA}||_F^2 \tag{2.4}$$

The notation $\|A\|_F$ stands for the Frobenius norm, defined as $\|A\|_F = \sqrt{\sum_{ij} A_{ij}^2}$. Then the penalty term in (2.2) can be rewritten as:

$$\begin{aligned} ||\mathbf{W}^* - \mathbf{DA}||_F^2 &= \left|\left|\mathbf{W}^* - \sum_{j=1}^{k} \mathbf{d}_j \mathbf{a}_T^j\right|\right|_F^2 \\ &= \left|\left|(\mathbf{W}^* - \sum_{j \neq k}^{k} \mathbf{d}_j \mathbf{a}_T^j) - \mathbf{d}_k \mathbf{a}_T^k\right|\right|_F^2 \\ &= \left|\left|\mathbf{E}_k - \mathbf{d}_k \mathbf{a}_T^k\right|\right|_F^2 \end{aligned} \tag{2.5}$$

Hence, we updated the $\left|\left|\mathbf{E}_k - \mathbf{d}_k \mathbf{a}_T^k\right|\right|_F^2$, assuming fixed coefficients $\mathbf{A}$ and error $\mathbf{E}_k$. The constraint is over the $j$th orthonormal basis $\mathbf{D}_j$. By decomposing the multiplication $\mathbf{DA}$ into the sum of $K$ rank 1 matrices, we can assume that the other $K-1$ terms were fixed, and the $k$th remains unknown. Then, the singular value decomposition find the closest $K-1$ terms that approximate $\mathbf{E}_k$, and this will effectively minimize the error in Eq. (2.5).

The above solution of vector $\mathbf{a}_T^k$ is very likely to be filled, because the sparsity constraint is not enforced. To enforce the sparsity constraint, we define $\omega_k$ as the group of indices pointing to examples $\mathbf{w}_i^*$ that use basis $\mathbf{d}_k$, and entries of $\mathbf{a}_T^k(i)$ that are non-zero. Thus, $\omega_k = \{i|1 \leq i \leq N, \mathbf{a}_T^k(i) \neq 0\}$. Then we compute $\mathbf{E}_k = \left\|\mathbf{W}^* - \sum_{j\neq k}^k \mathbf{d}_j \mathbf{a}_T^j\right\|_F^2$ by only choosing columns corresponding to $\omega_k$. We then apply the SVD decomposition $\mathbf{E}_k^R = \mathbf{U}\Lambda\mathbf{V}^T$. The solution for $d_k$ is the first column of $U$, and the updated coefficient vector is the first column of $\mathbf{U} \times \Lambda(1,1)$.

### *2.3.3 Summary of DL-CHI*

Putting all together, an overview of the DL-CHI method can be seen from Figure 2.3.2. A full description of the DL-CHI algorithm is also given in Algorithm 1. It can be seen in algorithm 1 that, we have to learn $\mathbf{W}, \mathbf{A}$, and $\mathbf{D}$. We split the algorithm into two phases for learning personalized CHI and dictionary learning. In the Phase I, we intend to solve $\mathbf{w}^*$ via CHI using the algorithm 3 described in Appendix A.2. In this phase we learn the model parameter vectors of all individuals, which leads to the construction of the matrix $\mathbf{W}^*$. In the Phase II, we use the K-SVD method to learn the dictionary by first computing the best representation matrix $\mathbf{A}$ via (2.3) using the matching pursuit algorithm, and then, searching for the best dictionary. With a learned dictionary, the representations of the individual's models could be identified and further used as the final individual models. Specifically, from the dictionary algorithm we can find the the low-dimensional structure of the model parameter matrix $\mathbf{W}^* \approx \mathbf{DA}$, where each column of $\mathbf{W}^*$ is a reconstructed model parameter vector of each individual to be linear combination of dictionary elements.

## *2.4 Numerical studies*

### *2.4.1 Real-world applications*

We implement the DL-CHI model on two real-world datasets that were collected in Alzheimer's disease (AD) and surgical site infection (SSI) research. Both diseases exhibit monotonic disease progression and significant patient heterogeneity. For the Alzheimer's disease data, we use the FDG-PET images of 162 subjects (Alzheimer's Disease: 74, Normal aging: 88)

**Algorithm 1** The DL-CHI algorithm

**Require:** $\mathbf{D}^{(0)} \in \mathbb{R}^{d\times k}$, $\mathbf{W}^* \in \mathbb{R}^{d\times n}$, and $\lambda \in \mathbb{R}$

**Ensure:** Find a dictionary $\mathbf{D} \in \mathbb{R}^{d\times k}$, and a corresponding coefficient matrix $\mathbf{A} \in \mathbb{R}^{k\times n}$ such that the representation error, $\mathbf{R} = \mathbf{W} - \mathbf{DA}$ is minimized and $\mathbf{A}$ fulfill sparseness criterion

**while** not converge **do**
  **Start iterations** t:= 1,2,... **do**
  **Update** $\mathbf{A}^{(t)}$
  **for** i=1,2,...,N,
    $\mathbf{a}_{(t)} = \min_{\mathbf{a}_i} \left|\left|\mathbf{w}_i^* - \mathbf{D}^{(\mathbf{t-1})}\mathbf{a_i}\right|\right|_2^2 \forall i, \|\mathbf{a}_i\|_0 \leq T_0$
  **end for**
  **Update** $\mathbf{D}^{(t)}$
  **for** $k = 1, \ldots, k$ Update the $k_{th}$ column of $\mathbf{D}^{(t)}$:
    Define $\omega_k = \{i | 1 \leq i \leq d, \mathbf{a}_T^k(i) \neq 0\}$
    Compute: $\mathbf{E}_k = \mathbf{W}^* - \sum_{t\neq k} \mathbf{d}_{(t)}\mathbf{a}_T^{(t)}$
    In $\mathbf{E}_k$, choose only columns corresponding to $\omega_k$
    Apply the SVD to the corresponding $\mathbf{E}_k$
    $\mathbf{E}_k = \mathbf{U}\Lambda\mathbf{V}^T$.
    The updated $d_k$ is the first column of $\mathbf{U}$.
    The updated $a_T^k$ is first column of $\mathbf{U} \times \Lambda(1,1)$.
  **end for**
  Reconstructed individual model parameter: $\mathbf{W}^* \approx \mathbf{DA}$

downloaded from the ADNI (`www.loni.usc.edu/ADNI`). For each subject, there are at least three time points and at most seven time points. The data has been preprocessed and the Automated Anatomical Labeling has been used to segment each image into 116 anatomical volumes of interest (AVOIs). We select 90 AVOIs that are in the cerebral cortex in our study. Each AVOI becomes a variable here. The measurement data of each region, according to

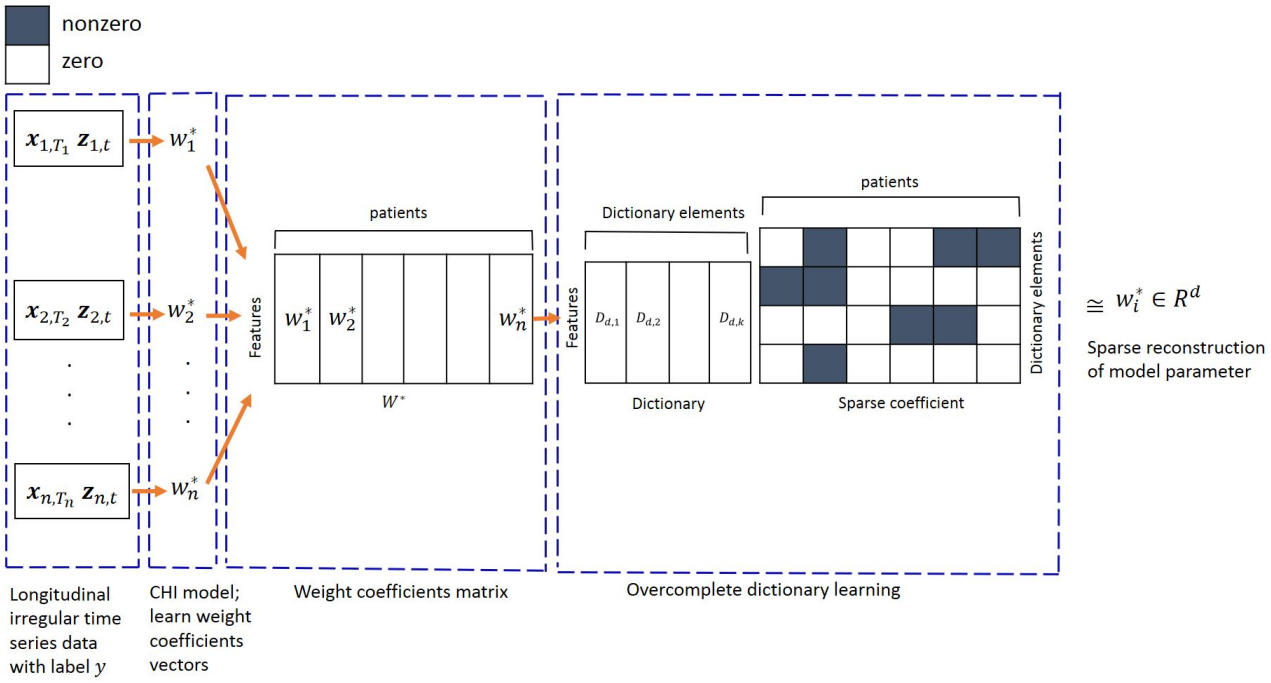



Figure 2.3.2: An algorithmic overview of the DL-CHI method

the mechanism of FDG-PET, is the regional average FDG binding counts, representing the degree of glucose metabolism. Extensive evidences in the literature have shown that the glucose metabolism will decline as a function of the aging, while the pathology of neurodegenerative diseases such as AD will further accelerate the declination, providing a perfect application example for implementing and testing the proposed DL-CHI method.

The SSI data exhibit similar characteristics as the AD data. There have been many models developed to monitor individuals who are subject to developing SSI [51, 52, 53], based on daily wound measurements, such as the temperature, granularity, distance of the wound, together with other non-wound related but important clinical signals such as heart rate, morning body temperature, and NG tube presence, etc. Figure 2.4.1 shows the longitudinal trend of a wound-related variable collected in our data, which clearly shows the monotonic degradation process of the SSI patients. The SSI data include longitudinal wound measure-

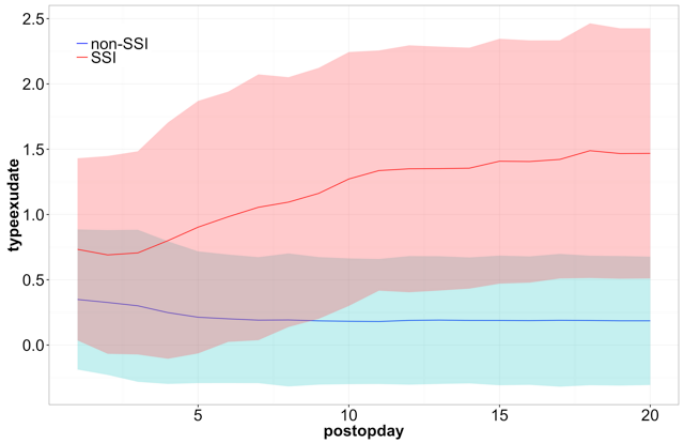



Figure 2.4.1: Example of the longitudinal data of wound assessment that could gradually separate the SSI group with the non-SSI group as the condition progresses over time [1]

ments from 857 patients, among which 169 are SSI patients and 539 are normal control. The data include wound measurement variables, for example, wound edge distance, temperature, include exudate amount, etc. Some other physiological variables such as heart rate are also provided in the data. Subjects were measured in time length ranging from 3 days to 21 days.

### *2.4.2 Parameter tuning and validation*

For each experiment, we randomly split the data into two equal parts, one for training and one for testing. For training, we used 10-fold cross validation to tune the parameters. As CHI is a complex data fusion mechanism that synthesizes monotonicity of the disease progression, label information, and statistical homogeneity, we use a comprehensive scheme to compare DL-CHI with CHI. Specifically, we compared the two models (1) when only monotonicity is used for model training (i.e., by setting $\beta = 0$ and optimizing for $\alpha$); (2) when only the label information is used for model training (i.e., by setting $\alpha = 0$ and optimizing for $\beta$); (3) when a full model is used (i.e., by optimizing for both $\alpha$ and $\beta$). In addition, we performed in each of the settings by randomly down-sampling the training data, i.e., only using a proportion of

the data randing from 15% to 75%, to train both models. A model that can maintain good performances with less training data in obviously more promising in healthcare applications while data collection is relatively more costly than other real-world applications.

### *2.4.3 Results*

Comparison between CHI and DL-CHI across a wide range of scenarios aforementioned are reported in Table 2.4.1. In general, it is observed that the DL-CHI model could significantly improve CHI model by accounting for the patient heterogeneity. This makes sense, since enforcing the constraint that the individual CHI model should be represented by a dictionary plays a role in the regularization of the model learning, as the dictionary basis vectors are numerical representations of patient heterogeneity. It is shown that in all of the three scenarios; using only monotonicity ($\beta = 0$), using only the label information $\alpha = 0$, or the full model DL-CHI model achieves satisfying results. Another observation is that enforcing monotonicity constraint alone leads to satisfactory performance for the DL-CHI model. As shown in Table 2.4.1, the DL-CHI method is also robust to small sample size. We investigate DL-CHI model's capability by selecting only 15% of the data as the training data, while the 10-folder cross validation was used to identify the optimal parameters in the model. The results show that our method achieves better prediction performance than the CHI model that uses the same ratio of the training data. Overall, the results show that the DL-CHI has a great potential for clinical applications to overcome the limitation of the CHI method in mitigating patient heterogeneity.

Table 2.4.2 shows the performance comparison of personalized DL-CHI method with the CHI model and and three dictionary methods; K-SVD, ILS-DLA and RLS-DLA. While for each model training, 10-fold cross validation is used on the training data and the AUC is evaluated on the testing data. Results in Table 2.4.2 show that the integration of dictionary learning with the CHI model improves the performance of the algorithm. The performance of RLS-DLA is in general considerable better than that of ILS-DLA and K-SVD. However, interestingly DL-CHI model performance demonstrates that it is superior to the

Table 2.4.1: AUC performance for ADNI and SSI data across different ratio of training and testing datasets obtained by 10-fold cross-validation

| Data | Ratio | CHI | DL-CHI |
|---|---|---|---|
| $\alpha = 0$, $\beta^*$ | | | |
| ADNI | 15% | 0.870 ± 0.024 | 0.887 ± 0.021 |
| | 20% | 0.883 ± 0.021 | 0.890 ± 0.016 |
| | 35% | 0.889 ± 0.014 | 0.936 ± 0.051 |
| | 50% | 0.890 ± 0.031 | 0.940 ± 0.047 |
| | 75% | 0.927 ± 0.012 | 0.959 ± 0.036 |
| SSI | 15% | 0.850 ± 0.055 | 0.867 ± 0.039 |
| | 20% | 0.861 ± 0.036 | 0.877 ± 0.020 |
| | 35% | 0.871 ± 0.012 | 0.886 ± 0.020 |
| | 50% | 0.862 ± 0.015 | 0.892 ± 0.041 |
| | 75% | 0.889 ± 0.024 | 0.914 ± 0.027 |
| $\alpha^*$, $\beta = 0$ | | | |
| ADNI | 15% | 0.780 ± 0.016 | 0.863 ± 0.034 |
| | 20% | 0.799 ± 0.054 | 0.873 ± 0.024 |
| | 35% | 0.804 ± 0.012 | 0.844 ± 0.034 |
| | 50% | 0.818 ± 0.019 | 0.869 ± 0.064 |
| | 75% | 0.855 ± 0.064 | 0.905 ± 0.024 |
| SSI | 15% | 0.829 ± 0.064 | 0.860 ± 0.023 |
| | 20% | 0.860 ± 0.021 | 0.879 ± 0.016 |
| | 35% | 0.870 ± 0.034 | 0.883 ± 0.034 |
| | 50% | 0.880 ± 0.042 | 0.892 ± 0.036 |
| | 75% | 0.883 ± 0.026 | 0.895 ± 0.016 |
| $\alpha^*$, $\beta^*$ | | | |
| ADNI | 15% | 0.865 ± 0.021 | 0.872 ± 0.025 |
| | 20% | 0.871 ± 0.023 | 0.881 ± 0.014 |
| | 35% | 0.874 ± 0.032 | 0.890 ± 0.026 |
| | 50% | 0.891 ± 0.021 | 0.910 ± 0.041 |
| | 75% | 0.901 ± 0.020 | 0.919 ± 0.036 |
| SSI | 15% | 0.741 ± 0.032 | 0.814 ± 0.041 |
| | 20% | 0.758 ± 0.034 | 0.820 ± 0.030 |
| | 35% | 0.770 ± 0.013 | 0.831 ± 0.036 |
| | 50% | 0.791 ± 0.026 | 0.887 ± 0.015 |
| | 75% | 0.806 ± 0.010 | 0.862 ± 0.036 |

Table 2.4.2: AUC performance comparison for ADNI and SSI data for CHI, DL-CHI, K-SVD, ILS-DLA and RLS-DLA models obtained by 10-fold cross-validation

| Data | ADNI | SSI |
|---|---|---|
| DL-CHI | 0.951 ± 0.025 | 0.902 ± 0.032 |
| CHI | 0.920 ± 0.021 | 0.880 ± 0.010 |
| RLS-DLA | 0.903 ± 0.030 | 0.873 ± 0.065 |
| K-SVD | 0.850 ± 0.043 | 0.803 ± 0.014 |
| ILS-DLA | 0.723 ± 0.012 | 0.653 ± 0.063 |

RLS-DLA despite its convergence as an online algorithm and its ability for reconstruction purposes.

### *2.4.4 Representation capacity of dictionary learning*

Figure 2.4.2 provides the results regarding the number of basis vectors needed for a sufficient representation of patient heterogeneity from AD. Apparently, the larger the dictionary size, the lower the representation error. On the other hand, we can also observe that the error of representation drops quickly with the increasing number of basis vectors in the dictionary. As the optimal dictionary size is not known in advance, hence we first obtained it through an initial dictionary $\mathbf{D}_0$ of large size $K$. The initial dictionary $\mathbf{D}_0 \in \mathbb{R}^{d\times k}$ is obtained by selecting $K$ samples randomly from input signals. The dictionary $\mathbf{D}_0$ helps minimizing the reconstruction error, and it is not yet optimal. For our experiment, we selected the number of basis based on the minimum error of representation given various dictionary sizes. To satisfy the overcompleteness we choose the size of $\mathbf{D}_0$ to be sufficiently larger than the dimension of an input signal.

## *2.5 Conclusion*

In this work, we presented a DL-CHI formulation to help build personalized contemporary health index (CHI) to monitor patient condition over time. Through applications on two real-world datasets of AD and SSI, the DL-CHI model is shown to better than the CHI model

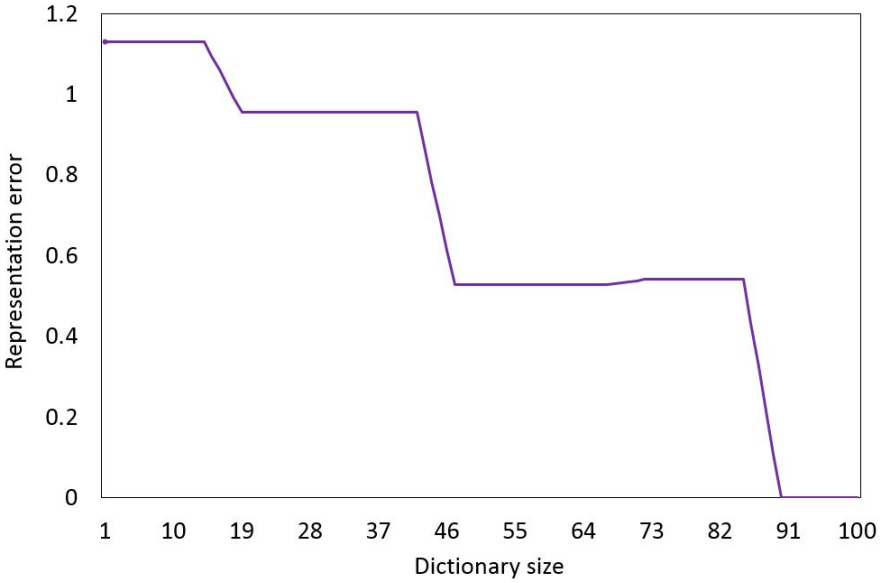



Figure 2.4.2: Representation error for different dictionary size.

in patient prediction and can achieve robust results with small sample sizes. In the future, we may further enhance the DL-CHI method in the following directions. First, note that, in the current DL-CHI formulation, the individual models have to be learned via the CHI formulation without information from the dictionary. Only with a learned dictionary, the representations of the individuals models are identified and further used as the final individual models. This is a possibility that a joint learning of both steps could further enhance the performance of DL-CHI by incorporating the dictionary into the CHI formulation. Second, the need of transfer learning when the supply of training data is limited is vital. One way to tackle this problem is by exploring the transfer learning through model-based transfer, where the prior knowledge from the generic recognizer enters through a modified regularization term in the CHI model. Last but not least, we can also consider an integration between data-based and model-based transfer learning. Where, by re-weighting the input source data we can minimize the discrepancy between the source and the target distributions, and then allowing CHI to be biased toward the parameters of another model.

# Chapter 3

# UQ-CHI: AN UNCERTAINTY QUANTIFICATION-BASED CONTEMPORANEOUS HEALTH INDEX FOR DEGENERATIVE DISEASE MONITORING

Developing a knowledge-driven contemporaneous health index (CHI) that precisely reflects the underlying patient condition over time holds considerable promise for informing a range of clinical decision-making challenges. This is particularly important for monitoring degenerative conditions such as Alzheimer's disease (AD), where patient condition decays over time. Detecting early symptoms, tracking progression signs, and continuous evaluation of severity, are all essential for disease management. Despite some recent model developments in the literature, the uncertainty quantification of health index models has been largely overlooked. To ensure continuity of the care, we should be more explicit about the level of confidence in model outputs. Ideally, decision-makers should be provided with recommendations that are robust in the face of substantial uncertainty regarding future outcomes. In this paper, we seek to fill this knowledge gap by developing an uncertainty quantification based contemporaneous longitudinal index, named UQ-CHI, with a particular focus on continuous patient monitoring of degenerative conditions. Our method combines convex optimization and Bayesian learning using the maximum entropy learning (MEL) framework, and also explicitly integrates uncertainty on labels. The methodology further provides closed-form solutions in some important decision making tasks, e.g., such as predicting the label of a new sample. Numerical studies demonstrate the effectiveness of the proposed UQ-CHI method in terms of prediction accuracy and monitoring efficacy, and unique advantages for application are discussed.

### *3.1 Introduction*

Effective monitoring of degenerative patient conditions represents a significant challenge in many clinical decision-making settings, which has led to the development of numerous mathematical and computational models [54, 55, 56, 57]. Developing a knowledge-driven contemporaneous health index (CHI) that precisely reflects underlying patient condition during the progression of conditions holds unique value. This includes facilitating a range of clinical decision-making opportunities [14, 15, 16], enhancing the continuity of care, and facilitating communications between clinicians, healthcare providers, and patients. It can also inform the development of many envisioned AI systems that implement adaptive interventions for better healthcare management, given a representation of the dynamic evolution of the patient's condition.

To ensure continuity of care it is important to be more explicit about our level of confidence in model outputs. Ideally, decision-makers should be provided with recommendations that are robust in the face of substantial uncertainty about future outcomes. However, computational models represent an abstraction of clinical observations, as such, they are often built on analytically tractable assumptions that overly simplify real-world problems. Also, most of these models are parameterized from sparse and typically imperfect data, subjecting them to a host of statistical errors. An approach that yields only a single prediction fail to adequately reflect the level of uncertainty, both in the empirical data and the estimated parameters [58]. As a result, the outcomes from such mathematical models may not show consistency with clinical observations. Uncertainty is an unavoidable feature that affects prediction capabilities in real-world domains, including healthcare [59, 60], manufacturing [61, 62], signal processing [63, 64], and many others. A certain amount of uncertainty is common in decision-making systems, for example when the experimental data are insufficient for robust model calibration. In such cases, a chance remains that model parameters will be unambiguously even in the existence of complex mathematical models. In clinical predictions, it is necessary to account for this uncertainty in an effective manner, because if

the model parameters are not well constrained, the resulting predictions may represent an unacceptable degree of posterior uncertainty. Furthermore, while most existing models in patient monitoring generate a single prediction without the reporting of confidence, uncertainty quantification could inform which samples we may not be ready to act based on the model. Therefore, uncertainty quantification is a much-needed capacity to develop reliable models for a clinically relevant prediction [65, 66, 67].

A number of patient monitoring index approaches have been developed in the literature. A standard formulation of these health indices involves the use of weighted sum models (e.g., regression models) that combine multiple static clinical measurements to predict disease condition. For example, many risk score models exist to predict AD using multi-modality data integration methods [68, 69, 70] integrating neuroimaging data [71, 72], genomics data [73], clinical data [74], etc. There are a few approaches that have quantified the decline of AD-related scores over time using a multi-task learning model [32, 31]. These existing efforts have been limited to combining static data rather than longitudinal data. Another challenge is that these data are usually sampled at irregular time points. Our problem's objective is fundamentally different from existing risk score models; we focus on developing the contemporaneous health index (CHI) that can use irregular multivariate longitudinal time-series data to quantify the severity of degenerative disease conditions that are required to fit the monotonic degradation process of the disease condition. For example, in our previous work [3] to address patient heterogeneity, we developed a dictionary learning-based contemporaneous health index for degenerative disease monitoring, called DLCHI. This index leveraged the knowledge of the monotonic disease progression process to fuse the data by integrating CHI with dictionary learning. The basic concept of DL-CHI involved the learning of individual models via the CHI formulation and then rebuilding the model parameters of each patient's models through supervised dictionary learning. However, both CHI and DL-CHI frameworks only produce one single set of predictions and thus ignore sampling uncertainty (i.e., it is common in healthcare that the label information is usually obtained

by subjective methods which are subject to uncertainty). Therefore, enabling CHI to also quantify uncertainty and incorporate this uncertainty in labels in its modeling, would result in broader applicability in real-world contexts. The main objective of this paper is to develop a framework that articulates the contemporaneous health index (CHI) developed in [1] and depicts the incorporation of uncertainty into CHI.

There is abundant literature describing how longitudinal measurements can be used in risk assessments [75, 76, 77]. Lee et al. developed a Bayesian functional Cox regression model [78] on the time-to-event data. Goldsmith et al. introduced a penalized functional regression model and inferential tools designed specifically for these emerging types of data [79]. However, clinical decision-making problems are sensitive to uncertainties; thus, it is critical to incorporate uncertainties into the modeling process. There are several methodologies to assess uncertainties in predictions of health risk models. One practical solution is using Bayesian inference as a principled technique to estimate model uncertainty. Here, uncertainty is a feature of the probability distribution of the output, which is induced by input parameter uncertainty. Characterizing uncertainties surrounding decision problems can be done by assigning prior distributions to each model input. Once prior distributions have been characterized, values drawn from these distributions can be propagated through the model. For example, a number of previous studies have used Monte Carlo simulation to propagate these prior distributions through the model [80, 81, 82]. However, Monte Carlo methods are based on random input configurations and the uncertainties are derived from the random selection of inputs. Therefore, the Monte Carlo estimate appears to be driven by aleatory uncertainty. Another example is where Rizopoulos [83] developed a Bayesian personalized prediction model called functional joint model (FJM). However, their model was incapable of handling incomplete label information.

In this paper, we develop a framework for uncertainty quantification based contemporaneous longitudinal index, named UQ-CHI, with a particular focus on continuous patient monitoring of degenerative conditions. Our method combines convex optimization and Bayesian learning

using the maximum entropy learning (MEL) framework, integrating uncertainty on labels as well. The basic idea of MEL is to identify parameter distributions of a statistical model that reflects maximum uncertainty, a principle that is conservative and robust [84, 85, 86]. This process has been investigated in a few machine learning models [87, 88, 89, 90] as well. For example, in [87], MEL was used to learn a distribution of the parameters in the support vector machine model rather than a single vector of the parameters. This distribution of the parameters could help us evaluate the uncertainty of the learned support vector machine model and translate into the uncertainty of predictions.

A few challenges must be addressed prior to adapting the MEL formulation and developing the UQ-CHI. The objective function of MEL, as its distinct feature, reflects the notion of maximum entropy: regardless of the specific model, the learning objective of MEL is to estimate the distribution of the model parameters that have the maximum entropy. If there is a prior distribution of the parameters, the Kullback-Leibler divergence can be used to extend this idea. In our case, the properties of the prior distribution should be studied to account for label uncertainties. In addition to the objective function, the MEL encodes information from the data into constraints; in the case of a classification model, for each sample, there would be a constraint that the expectation of the prediction over the distribution of the parameters matches the observed outcome on this sample. Here, we derive the constraints from the CHI model and integrate it with the MEL framework. Specifically, there are two steps in our method: training and prediction. During training, we consider a prior uncertainty over the labels to address uncertain or incomplete labels. Then we derive a solution to the optimization problem by using a specific prior formulation. Next, we develop a prediction method, with a rejection option method, for new samples with the obtained uncertainty quantification capacity. A distinct feature of our model is that it provides a closed-form solution for predicting the label of a new example. The whole pipeline of this UQ-CHI model is demonstrated in Figure 3.1.1.

Below, this paper is organized as follows. In Section 3.2, we will review related literature

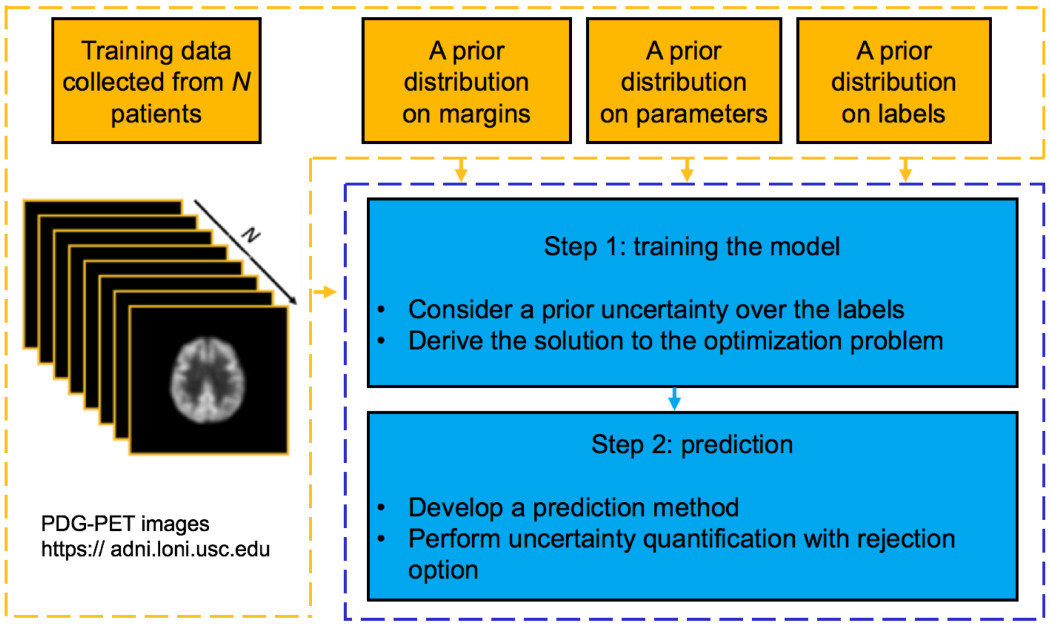



Figure 3.1.1: A conceptual overview of the UQ-CHI method

in modeling the contemporaneous health index for degenerative conditions and the MEL framework. In Section 3.3, the UQ-CHI framework will be presented. In Section 3.4, we will implement and evaluate the UQ-CHI using a simulated dataset. We then continue the numerical analysis with a real-world application on Alzheimer's disease dataset in Section 3.5. We will provide concluding remarks in Section 3.6. Note that, in this paper, we use lowercase letters, e.g., x, to represent scalars, boldface lowercase letters, e.g., $\mathbf{v}$, to represent vectors, and boldface uppercase letters, e.g., $\mathbf{W}$, to represent matrices.

## *3.2 Related works*

In this section, we briefly present the basic formulation of the contemporaneous health index (CHI) model, and its extension, the dictionary learning-based contemporaneous health index (DL-CHI), and then present the proposed UQ-CHI model.

### *3.2.1 The CHI model*

The CHI model exploits the monotonic pattern of disease over the course of progression to improve the data fusion of multivariate clinical measurements taken at irregular time points [1]. The CHI framework was inspired by the common characteristics of degenerative conditions (e.g., AD) that often cause irreversible degradation. For example, in AD, a number of biomarkers were developed to measure the degradation of the neural systems. These biomarkers included neuroimaging modalities such as PET and MRI scans [29, 30]. MRI scans show a decline in the brain volume over time associated with disease progression. The same phenomenon is observed on PET scans when there is a persistent shrinkage of metabolic activities. Such monotonic patterns indicate that once disease progression starts, it tends to increasingly deteriorate over time.

The task of CHI is to translate multivariate longitudinal and irregular clinical measurements into a contemporaneous health index $h_{n,t}$ to capture the changing conditions of the patient over the course of progression. Note, clinical measurements for each patient could be taken with different length of time and at different time locations. Targeting degenerative conditions, CHI is designed to be monotonic, i.e., $h_{n,t_1} \geq h_{n,t_2}$ if $t_1 \geq t_2$, where higher index represents a more severe conditions. CHI is a latent structure; hence, clinical variables associated with it should be measured over time to facilitate data for learning the index.

Let, $\mathbf{x}_{n,t} = [x_{n,1,t}, \ldots, x_{n,d,t}]^T \in \mathbb{R}^d$, denote a training set of $N$ patients. Each measurement $x_{n,i,t}$, is the value of the $i$th variable for the $n$th subject in a given time $t$, where $t \in \{1, \ldots, T_n\}$ is the time index. our goal is, given a training set, convert each measurement $\mathbf{x}_{n,t}$ into an health index $h_{n,t}$, which requires a mathematical model of $h_{n,t} = f(\mathbf{x}_{n,t})$. For simplicity, multivariable form of the hypothesis function $h_{n,t}$ was studies in [1], i.e., $h_{n,t} = \mathbf{x}_{n,t}.\mathbf{w}$, where $\mathbf{w} \in R^d$ is a vector of weight coefficients that combines the $d$ variables. The total number of positive and negative samples is shown by $N^+$ and $N^-$ respectively, i.e., $N^+ := |\{n|y_n = 1\}|$ and $N^- := |\{n|y_n = -1\}|$. The formulation of the CHI learning framework is shown in

below:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + \tag{3.1a}$$

$$\beta \sum_{n\in\{1,\dots,N\}} \max\left(0, 1 - y_n(\mathbf{x}_{n,T_n}^\top \mathbf{w} + b)\right) + \tag{3.1b}$$

$$\alpha \sum_{\substack{n\in\{1,,\dots,,N\} \\ t\in\{1,\dots,T_n-1\}}} \max\left(0, 1 - \mathbf{z}_{n,t}^\top \mathbf{w}\right) + \tag{3.1c}$$

$$\frac{\lambda}{2}\left(\frac{1}{N^+} \sum_{n\in\{N^+|y_n=1\}} \left((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^+)^T \mathbf{w}\right)^2\right) + \tag{3.1d}$$

$$\frac{\lambda}{2}\left(\frac{1}{N^-} \sum_{n\in\{N^-|y_n=-1\}} \left((\mathbf{x}_{n,T_n} - \bar{\mathbf{x}}_{T_n}^-)^T \mathbf{w}\right)^2\right) + \tag{3.1e}$$

$$\gamma\|\mathbf{w}\|_1. \tag{3.1f}$$

Items in (3.1) can be explained as follows:

- The first term (3.1a) and the second term (3.1b) are derived from a general formulation of support vector machine (SVM). These two terms are used to enhance the discriminatory power of CHI by utilizing the label information. Here, $y_n \in \{1, -1\}$ is the label of the $n$th sample that indicates if the $n$th subject has the disease or not.

- To accommodate the monotonic pattern of disease progression, and to enforce the monotonicity of the learned health index, the term (3.1c) is invented, i.e., $h_{n,t_1} \geq h_{n,t_2}$ if $t_1 \geq t_2$. Here, $\mathbf{z}_{n,t}$ is the difference of two successive data vectors $\mathbf{z}_{n,t} := \mathbf{x}_{n,t+1} - \mathbf{x}_{n,t}$.

- To encourage the homogeneity of CHI within the group that has the same health status terms (3.1d) and (3.1e) are invented. Here, $\bar{\mathbf{x}}_{T_n}^+$ and $\bar{\mathbf{x}}_{T_n}^-$ represent the center of data vectors at time $T_n$ for all positive and negative samples, respectively, that are,

$$\bar{\mathbf{x}}_{T_n}^+ := \frac{1}{N^+} \sum_{n\in\{n|y_n=1\}} \mathbf{x}_{n,T_n}$$

$$\bar{\mathbf{x}}_{T_n}^- := \frac{1}{N^-} \sum_{n\in\{n|y_n=-1\}} \mathbf{x}_{n,T_n}.$$

- To encourage sparsity of the features, $L_1$-norm penalty is used as shown in the last term (3.1f).

The CHI formulation can be solved by using the block coordinate descent algorithm that is illustrated in [1]. Note, the CHI formulation generalizes many existing models, such as SVM, sparse SVM, LASSO, etc.

### *3.2.2 The DL-CHI model*

CHI formulation seeks to model the average of a population, and thus, ignores important among patient heterogeneity. However, patients who suffer from AD have very heterogeneous progression patterns [18, 19, 20]. Building a personalized model on an individual basis could be used to capture heterogeneity. However, such models require a significant amount of labeled training samples, which is often not feasible in clinical settings. Towards this goal, the DL-CHI approach was further developed in [3] by integrating CHI with dictionary learning [21, 18]. Dictionary learning algorithms reconstruct the input signals as an approximated signal via a sparse linear combination of a few dictionary elements or basis [22] (each column of the dictionary represents a basis vector). Dictionary learning algorithms can reveal hidden structures in the data (much like principal component analysis) by spanning the space of a personalized model and capturing patient heterogeneity. They play a role during regularization of the model learning, in a way that each dictionary basis vector can be viewed as the numerical representations of patient heterogeneity. Thus, DL algorithms can improve overall classification performance. Translating this knowledge into DL-CHI, the basic idea is first to learn individual models through the CHI formulation, and then, reconstruct the model parameters of the individual learned models via supervised dictionary learning. As such, each model is represented as a sparse linear combination of the basis vectors. Numerous experiments in both simulated and real-world data have shown the effectiveness of DL-CHI in creating personalized CHI models.

Despite accounting for patient heterogeneity, DL-CHI ignores sampling uncertainty, which can limit its applicability in real-world applications. This provides the motivation to enable

CHI to conduct uncertainty quantification.

### *3.2.3 The MEL formulation*

Maximum entropy can be considered as a special case of a standard Bayesian approach to support vector machines. Unlike standard Bayesian framework, (MEL) framework has a straightforward probabilistic interpretation of the learning scheme and the prediction rule. As mentioned in Section 3.1, MEL formulation has a distinct objective function that aims to learn the parameter distributions of a model that encodes maximum uncertainty (i.e., evaluated by the entropy concept). It also imposes constraints that encode information from the data. For example, in the case of a classification model, for each sample, there would be a constraint that the expectation of the prediction over the distribution of the parameters should match the observed outcome on this sample. To further illustrate some details, a common application of the MEL is the maximum entropy discrimination (MED) method that focuses on the application of MEL on classification models.

Let's consider a binary classification problem, where the response variable $y$ takes values from $\{+1, -1\}$. Let $\mathbf{x}_n = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be an input feature vector and $\mathcal{D}(\mathbf{x}_n|\mathbf{w})$ be a discriminant function parameterized by $\mathbf{w}$, and $\gamma$ e.g., $\mathcal{D}(\mathbf{x}_n|\mathbf{w}) = \mathbf{w}^T\mathbf{x}_n$. The training set is defined by $D = \{\mathbf{x}_n, y_n\}_{n=1}^N$ and the hinge loss is defined as $h(x) = \max(0, y_i\mathcal{D}(\mathbf{x}_n|\mathbf{w}))$. The classification margin is defined as $y_n\mathcal{D}(\mathbf{x}_n, \mathbf{w})$, and it is large and positive when the label $y_n$ agrees with the prediction. Traditional learning machines such as the max-margin methods learn the optimal parameter setting $\mathbf{w}, \gamma$ by the empirical loss and the regularization penalty as shown below:

$$\begin{aligned} &\min_{(\mathbf{w},\gamma_n)} R(\mathbf{w}) + \sum_n L(\gamma_n) \\ &\text{s.t.} \quad y_n\mathcal{D}(\mathbf{x}_n \mid \mathbf{w}) - \gamma_n \geq \mathbf{0}, \quad \forall n \end{aligned} \tag{3.2}$$

Where $L()$ is the loss function which is a non-increasing and convex function of the margin, and $R(\mathbf{w})$ is the regularization penalty. However, MED considers a more general problem of finding a distribution $p(\mathbf{w}, \gamma)$ over $\mathbf{w}$ and classification margin parameters $\gamma$. This could be

done by minimizing its relative entropy with respect to some prior target distribution $p_0(\mathbf{w}, \gamma)$ under certain margin constraints. Specifically, suppose that a prior distribution, denoted as $p_0(\mathbf{w}, \gamma)$, is available, then MED learns a distribution $p(\mathbf{w}, \gamma)$ by solving a regularized risk minimization problem. When the prior distribution is not a uniform distribution, this can be generalized as minimizing the relative entropy (or Kullback-Leibler divergence) and the regularization penalty as follows (penalizing larger distances from priors):

$$\min_{p(\mathbf{w},\gamma)} KL\big(p\left(\mathbf{w},\gamma\right)||p_0\left(\mathbf{w},\gamma\right)\big) + CR\big(p(\mathbf{w},\gamma)\big). \tag{3.3}$$

Here, $C$ is a constant and $R(p\big(\mathbf{w},\gamma)\big) = \sum_n h\big(y_n E_{p(\mathbf{w},\gamma)}[\mathcal{D}\left(\mathbf{x}_n|\mathbf{w}\right) - \gamma_n]\big)$ is the hinge-loss that captures the large-margin principle underlying the MED prediction rule:

$$\hat{y} = sign\big(E_{p(\mathbf{w},\gamma)}[\mathcal{D}\left(\mathbf{x}_n|\mathbf{w}\right) - \gamma_n])\big). \tag{3.4}$$

And the KL divergence is defined as follows:

$$KL\big(p\left(\mathbf{w},\gamma\right)||p_0\left(\mathbf{w},\gamma\right)\big) = \int p\left(\mathbf{w},\gamma\right)\log\frac{p\left(\mathbf{w},\gamma\right)}{p_0\left(\mathbf{w},\gamma\right)}. \tag{3.5}$$

Here in (3.3), the classification margin quantities are included; $\gamma_n$ as slack variables in the optimization, which represents the minimum margin that $y_n\mathcal{D}(\mathbf{x}_n|\mathbf{w})$ must satisfy. MED considers an expectation form of the traditional approaches and casts Eq. (3.2) as an integration. The classification constraints will also be applied in an expected form. As a result, MED no longer finds a fixed set of the parameters, rather it estimates a distribution of parameter values, and it uses a convex combination of discriminant functions rather than one single discriminant function to derive model averaging for decisions. In particular, MED formulation estimates parameter distributions that are as close as possible with the prior distribution over all parameters regarding KL-divergence subject to various moment constraints. This analogy extends to cases where the distributions reflect unlabeled samples, missing values, or other probabilistic entities that are introduced when designing the discriminant function. Correspondingly, MED is an effective approach to learn a discriminative classifier as well as consider uncertainties over model parameters; thus, it combines

generative and discriminative learning [91, 90]. This generalization facilitates a number of extensions of the basic approach, including uncertainty quantification described in this paper. The present work introduces a novel generalization of CHI formulation by integrating the MED to perform the task of uncertainty quantification.

## *3.3 The proposed work: the UQ-CHI model*

The overall goal of UQ-CHI is to learn a distribution $p(\mathbf{w})$ over the parameters of CHI model $\mathbf{w}$. An additional goal is to achieve modeling success even if only partial labels are given, and when the labels might also contain uncertainty. Therefore, the first step in constructing the UQ-CHI is to create the constraint structure. To design the UQ-CHI, we incorporate some features from the original formulation of the CHI via Eq. (3.1) as follows: First, we utilize the label information by defining the discriminant function $\mathcal{D}\left(\mathbf{x}_{n,Tn}|\mathbf{w}\right) = \mathbf{w}^T\mathbf{x}_{n,Tn}$ which corresponds to (3.1b). We, then incorporate the distinct feature of the CHI formulation, the monotonicity regularization function $\mathcal{M}\left(\mathbf{z}_{n,t}|\mathbf{w}\right) = \mathbf{w}^T\mathbf{z}_{n,t}$ that corresponds to Eq. (3.1c). Note that, here, we do not incorporate the additional terms in Eq. (3.1d) and Eq. (3.1e) as they require full knowledge of labels of the samples. In addition, we don't include the sparsity regularization term (3.1f), because our focus is to learn $p(\mathbf{w})$ rather than the parameter vector $\mathbf{w}$. Also, our model can induce sparsity, e.g., if we impose a Laplace prior distribution for the parameters as to what is done in Bayesian Lasso model [92].

In the following subsections, we introduce the design the prior distributions, the constraints, and explain how to derive computational algorithms and closed-form solutions for training and prediction.

### *3.3.1 Design of constraints and prior distributions*

As aforementioned, there are two types of constraints that we can extract from the CHI formulation into the development of UQ-CHI. One corresponds to the discriminant function $\mathcal{D}\left(\mathbf{x}_{n,Tn}|\mathbf{w}\right) = \mathbf{w}^T\mathbf{x}_{n,Tn}$ used in CHI, to generate prediction on samples, while the other one corresponds to the monotonicity regularization function $\mathcal{M}\left(\mathbf{z}_{n,t}|\mathbf{w}\right) = \mathbf{w}^T\mathbf{z}_{n,t}$. Based on the CHI formulation, it is supposed that the model should lead to $y_n\mathcal{D}\left(\mathbf{x}_{n,Tn}|\mathbf{w}\right) = 1$ and

$\mathcal{M}(\mathbf{z}_{n,t}|\mathbf{w}) \geq 0$. As this perfect model may not exist, a set of margin variables $\gamma = [\gamma_1, \ldots, \gamma_n]$ are introduced. We consider an expectation form of the previous approach and cast Eq. (3.1) as an integration. Hence, the classification constraints are applied in an expected sense. This will lead to the following formulation for the constraints:

$$\int p(\mathbf{w}, \gamma) [p(y_n)\mathcal{D}(\mathbf{x}_{n,Tn} \mid \mathbf{w}) - \gamma_n] d\mathbf{w} d\gamma + \tag{3.6a}$$

$$\int p(\mathbf{w}, \gamma) [\mathcal{M}(\mathbf{z}_{n,t}|\mathbf{w}) - \gamma_n] d\mathbf{w} d\gamma \geq 0. \tag{3.6b}$$

Here, the term (3.6a) is the discriminant function and the term (3.6b) is the monotonicity regularization function. And, $p(y_n)$ is the distribution of $y_n$, and $p(\mathbf{w}, \gamma)$ is the distribution of $\mathbf{w}, \gamma$. With the prior distribution, we can derive the prediction rule: $\hat{y} = sign(E_{p(\mathbf{w})}[\mathcal{D}(\mathbf{x}_n|\mathbf{w}]))$.

Now we move on to the design of the prior distribution $p_0(\mathbf{w}, \gamma, y)$. It is natural to decompose the joint prior distribution as a product of three distributions:

$$p_0(\mathbf{w}, \gamma, y) = p_0(\mathbf{w}) \prod_1^N p_0(\gamma_n) \prod_1^N p_0(y_n). \tag{3.7}$$

In what follows we discuss each of the three prior distributions. Specifically, it is reasonable to assume that a level of uncertainty can be designed to each example in defining $p_0(y_n)$. A simple solution is to set $p_0(y_n) = 1$ whenever $y_n$ is observed and $p_0(y_n) = 0.5$ otherwise. To define $p_0(\mathbf{w})$, we choose $p_0(\mathbf{w})$ to be a Gaussian distribution with mean vector as $\mathbf{0}$ and covariance matrix as an identity matrix $\mathbf{I}$. To define the prior over the margin variables, we assume that it could be factorized $p(\gamma) = \prod_n p_0(\gamma_n)$. Further, following the idea proposed in [87], we can set $p_0(\gamma_n) = ce^{-c(1-\gamma_n)}$ and $\gamma_n \leq 1$. Here, $1 - \frac{1}{c}$ is actually the mean of the prior distribution of $\gamma_n$, so the idea of this distribution is to incur a penalty only for margins smaller than $1 - \frac{1}{c}$, while for margins larger than this quantity are not penalized. More details about the design of prior distributions will be given in Section 3.3.4.

### *3.3.2 The computational algorithm for UQ-CHI*

The full formulation of the proposed UQ-CHI model is shown below:

$$\min_{p(\mathbf{w},\gamma)} KL\big(p\left(\mathbf{w},\gamma\right)||p_0\left(\mathbf{w},\gamma\right)\big) \tag{3.8a}$$

$$\text{s.t.} \int p\left(\mathbf{w},\gamma\right)\left[p(y_n)\mathcal{D}(\mathbf{x}_{n,Tn} \mid \mathbf{w}) - \gamma_n\right]d\mathbf{w}d\gamma + \tag{3.8b}$$

$$\int p\left(\mathbf{w},\gamma\right)\left[\mathcal{M}\left(\mathbf{z}_{n,t}|\mathbf{w}\right) - \gamma_n\right]d\mathbf{w}d\gamma \geq 0. \tag{3.8c}$$

Essentially, solving optimization formulation Eq. (3.8) is to find a solution by calculating the relative entropy projection from the overall prior distribution $p_0(\mathbf{w},\gamma,y)$ to the admissible set of distributions $p$ that are consistent with the constraints. In what follows, we develop the computational algorithm to solve this formulation Eq. (3.8) and further derive the method for the prediction on samples.

*Step 1: Training the model*

In the training step, we consider a joint distribution of $\mathbf{w}$, and the margin vector of $\gamma = [\gamma_1,\ldots,\gamma_n]$ while fixing $p(y_n)$. In this step, we first explain the solution to the MED optimization problem subject to the terms in (3.3).

**Lemma 3.3.1.** *Let the loss function be a non-increasing and convex function of the margin, and let the Lagrangian of the optimization problem defined as $\mathcal{L}$ and $\lambda = [\lambda_1,\ldots,\lambda_n]$ be a set of non-negative Lagrange multipliers. Given the prior distribution $p_0(\mathbf{w})$ and the model distribution $p(\mathbf{w})$, and the discriminant function $\mathcal{D}\left(\mathbf{x}_n|\mathbf{w}\right)$ in order to minimize the relative entropy in terms of the KL-divergence ($KL(p(\mathbf{w}||p_0(\mathbf{w}))$ subjected to set of defined constraints, the MED optimization problem* (3.3) *can be written as:*

$$\begin{aligned} &\max_{\lambda} J(\lambda) = -logZ(\lambda) \\ &\quad s.t. \quad \lambda_i \geq 0 \quad for \quad i = 1,\ldots,N \end{aligned} \tag{3.9}$$

*Here, $Z(\lambda)$ is the normalization constant defined as:*

$$Z(\lambda) = \int p_0(\mathbf{w}) \exp\left(\sum_n \lambda_n y_n \mathcal{D}\left(\mathbf{x}_n|\mathbf{w}\right)\right) d\mathbf{w}, \tag{3.10}$$

The proof of Lemma 3.3.1 can be found in B.1. Now, the model training problem is revealed to be another optimization problem, that is learning optimal $\lambda^*$ by solving the dual objective function $J$ under positivity constraint. Based on the results from Lemma 3.3.1, after adding dual variables for the constraint in Eq. (3.8), the Lagrangian of the optimization problem can be written as:

$$\begin{aligned}\mathcal{L} = &\int p\left(\mathbf{w},\gamma\right)\log\frac{p\left(\mathbf{w},\gamma\right)}{p_0\left(\mathbf{w},\gamma\right)}d\mathbf{w}d\gamma- \\ &\left(\sum_{n\in\{1,\ldots,N\}}\int p\left(\mathbf{w},\gamma\right)\lambda_n[p(y_n)\mathcal{D}(\mathbf{x}_{n,Tn}\mid\mathbf{w})-\gamma_n]d\mathbf{w}d\gamma+\right. \\ &\left.\sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}}\int p\left(\mathbf{w},\gamma\right)\lambda_n\left[\mathcal{M}\left(\mathbf{z}_{n,t}|\mathbf{w}\right)-\gamma_n\right]d\mathbf{w}d\gamma\right).\end{aligned} \tag{3.11}$$

In order to find a solution, we require:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial p\left(\mathbf{w},\gamma\right)} &= \log\frac{p\left(\mathbf{w},\gamma\right)}{p_0\left(\mathbf{w},\gamma\right)}+1- \\ &\left(\sum_{n\in\{1,\ldots,N\}}\lambda_n[p(y_n)\mathcal{D}(\mathbf{x}_{n,Tn}\mid\mathbf{w})-\gamma_n]+\right. \\ &\left.\sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}}\lambda_n\left[\mathcal{M}\left(\mathbf{z}_{n,t}|\mathbf{w}\right)-\gamma_n\right]\right) \\ &= \mathbf{0},\end{aligned} \tag{3.12}$$

Which results in the following theorem.

**Theorem 3.3.2.** *The solution to the UQ-CHI problem has the following general form:*

$$\begin{aligned}p\left(\mathbf{w},\gamma\right)^* = &\frac{1}{Z(\lambda)}p_0\left(\mathbf{w},\gamma\right) \\ &\exp\left(\sum_{n\in\{1,\ldots,N\}}\lambda_n[p(y_n)\mathcal{D}(\mathbf{x}_{n,Tn}\mid\mathbf{w})-\gamma_n]+\right. \\ &\left.\sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}}\lambda_n\left[\mathcal{M}\left(\mathbf{z}_{n,t}|\mathbf{w}\right)-\gamma_n\right]\right).\end{aligned} \tag{3.13}$$

Thus, finding the solution to (3.8) depends on being able to evaluate the normalization constant $Z(\lambda)$.

**Lemma 3.3.3.** *Let $Z(\lambda)$ be the normalization constant defined in Eq.* (3.10)*. Based on the finding in* (3.13)*, $Z(\lambda)$ can be reformulated as follows:*

$$Z(\lambda) = Z_{\mathbf{w}}(\lambda) + Z_{\gamma}(\lambda) \tag{3.14a}$$

$$= \exp\left(\frac{1}{2}\Big(\sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} \sum_{\substack{n\in\{1,\dots,N\}\\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\right)^T \tag{3.14b}$$

$$\Big(\sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} \sum_{\substack{n\in\{1,\dots,N\}\\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Big)\right) + \tag{3.14c}$$

$$\prod_{n\in\{1,\dots,N\}} \frac{1}{1-\lambda_n/c} \exp(-\lambda_n). \tag{3.14d}$$

*Where, $Z_{\mathbf{w}}(\lambda)$ is defined in* (3.14b) *and* (3.14c) *$Z_{\gamma}(\lambda)$ is defined in* (3.14d)*.*

The proof of Lemma 3.3.3 can be found in the B.2. Given the reformulated normalization constant $Z(\lambda)$ in (3.14), the maximum of the jointly concave function objective function $J(\lambda)$ showing in Eq. (3.9) can be found through a constrained non-linear optimization. As a result, by substituting Eq. (3.14) in Eq. (3.9) we get:

$$\begin{aligned} J(\lambda) = \sum_{n\in\{1,\dots,N\}} \Big(\lambda_n + log(1-\lambda_n/c)\Big) - \\ \frac{1}{2}\Big(\sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} \sum_{\substack{n\in\{1,\dots,N\}\\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Big)^T \\ \Big(\sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} \sum_{\substack{n\in\{1,\dots,N\}\\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Big). \end{aligned} \tag{3.15}$$

Here, $\lambda \geq \mathbf{0}$. Thus, we have the following dual optimization problem:

$$\begin{aligned}\max_{\lambda} \sum_{n\in\{1,\ldots,N\}} \Big(\lambda_n + log(1-\lambda_n/c)\Big) - \\ \frac{1}{2}\Big(\sum_{n\in\{1,\ldots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} \sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Big)^T \\ \Big(\sum_{n\in\{1,\ldots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} \sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Big) \\ \text{s.t.} \quad \lambda \geq \mathbf{0}\end{aligned} \tag{3.16}$$

The Lagrange multiplier $\lambda$, is recovered by solving the convex optimization problem Eq. (3.16). Note that since the prior factorizes across $\mathbf{w}, \gamma$, UQ-CHI solution also factorized as well, i.e., $p(\mathbf{w}, \gamma) = p(\mathbf{w})p(\gamma)$.

**Corollary 3.3.4.** *From results in Theorem 3.3.2 the marginal distribution $p(\mathbf{w})$ can be found as follows:*

$$\begin{aligned}p\left(\mathbf{w}\right) = \frac{1}{Z_{\mathbf{w}}(\lambda)} p_0\left(\mathbf{w}\right) \exp\Big(\sum_{n\in\{1,\ldots,N\}} \lambda_n p_n(y_n)\mathbf{w}^T\mathbf{x}_{n,Tn} + \\ \sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}} \lambda_n \mathbf{w}^T\mathbf{z}_{n,t}\Big).\end{aligned} \tag{3.17}$$

Where, $Z_{\mathbf{w}}(\lambda)$ can be obtained from Eq. (3.14b) and (3.14c).

*Step 2: Prediction*

After obtaining the marginal distribution $p(\mathbf{w})$ in (3.17), the following lemma is used to predict the label of a new example $\mathbf{x}_{new}$. Referring to the solution of the UQ-CHI problem in (3.13), we can easily modify the regularization approach for predicting a new label from a new input sample $\mathbf{x}_{new}$ that is shown by $\hat{y} = sign\ \mathcal{D}\left(\mathbf{x}|\hat{\mathbf{w}}\right)$. In what follows, we generate the predictive label for the upcoming new labels.

**Lemma 3.3.5.** *Given the marginal distribution $p\left(\mathbf{w}\right)$ in (3.17) and the convex combination of discriminant functions $\int p(\mathbf{w})\mathcal{D}\left(\mathbf{x}|\mathbf{w}\right) d\mathbf{w}$, and let $\lambda^*$ be the optimal Lagrangian multiplier*

*obtained from the optimization problem* (3.16)*, and given* $Z_{\mathbf{w}}(\lambda)$ *obtained from* (3.14d)*, then the predictive label for the new* ($\mathbf{x}_{new}$) *can be generated as:*

$$\hat{y} = sign\left(\left(\sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} + \sum_{\substack{n\in\{1,\dots,N\}\\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\right)^T \mathbf{x}_{new}\right). \tag{3.18}$$

The proof of Lemma 3.3.5 is shown in Appendix B.3.

*Summary of the algorithms*

A full description of the training and prediction of UQ-CHI model is given in Algorithm 2.

**Algorithm 2** The UQ algorithm

**Require:** $\mathbf{x}_n \in \mathbb{R}^d$, $\mathbf{y}_n$, $p_0(\mathbf{w})$, and $(\epsilon)$

**Ensure:**

**while** not converge **do**
  **Start iterations** t:= 1,2,... **do**
  **Find** $p(\mathbf{w})$
  **for** $n = 1, 2, \dots, N$,
  $\min_{p(\mathbf{w})} H(p(\mathbf{w}))$
    s.t. $\int p(\mathbf{w})[y_n \mathcal{D}(\mathbf{x}_n, \mathbf{w}) - \gamma_n]d\mathbf{w} \geq \mathbf{0}, \quad \forall n$
  **Rejection option**
  **reject the sample if**: $p(\mathbf{w}|x_i) < \epsilon$

  **end for**

### *3.3.3 UQ-CHI with rejection option*

The performance of a predictive model is typically evaluated based on its accuracy, on a scheme of classifying all samples, regardless of the degree of confidence associated with the classification of the samples. However, accuracy is just one measure of model performance.

In many healthcare applications, it is preferred to make predictions when the confidence assigned to the classification is relatively high, rather than classify all samples when confidence is low. In this case, a sample can be omitted if it doesn't fit into any of the classes. In pattern recognition, this problem is often solved by estimating the class conditional probabilities and rejecting the samples that have the lowest class posterior probabilities (i.e., samples that are the most unreliable). As UQ-CHI enables uncertainty quantification, here, we create a rejection option in prediction to show the utility of uncertainty quantification in practice. The basic idea of rejection option is that the predictive model does not generate a prediction if uncertainty is higher than a given probability threshold. In other words, a sample that is most likely to be misclassified is rejected as described below:

$$p(\mathbf{w}|x_i) < \kappa \quad i = 1, \ldots, N, \tag{3.19}$$

Here, $\kappa$ is the rejection rate. The samples $x_i$ are rejected for which the maximum posterior probability $p(\mathbf{w}|x_i)$ is below a probability threshold. And a sample is accepted when:

$$p(\mathbf{w}|x_i) \geq \kappa \quad i = 1, \ldots, N \tag{3.20}$$

Thus, we define a classification with rejection option as $\hat{y}^{Rej}$, where, if a sample is rejected $\hat{y}_i^{Rej} = 0$, denotes rejection, else, the classification is $\hat{y}_i$, where, $\hat{y}_i$ corresponds to the classification of the $i$th sample defined in Eq. (3.18).

### *3.3.4 Tractability of UQ-CHI related to design of prior distribution*

Recall that by applying the MED to our optimization problem we no longer learn the model parameter, and instead, we specify the probability distributions. These distributions give rise to penalty functions for the model and the margins via KL-divergence. In detail, the model distribution will give rise to a divergence term $KL(p(\mathbf{w})||p_0(\mathbf{w}))$, and the margin distribution will give rise to the divergence term $KL(p(\gamma)||p_0(\gamma))$ which corresponds to the regularization penalty and the loss function respectively. The trade-off between classification

| Algorithm name | UQ-CHI | | | | CHI |
|---|---|---|---|---|---|
| Label ratio | Training ratio | Rejection rate | | | |
| | | Low = 20 | Medium = 40 | High = 60 | |
| Low = 10 | 30 | 0.69 | 0.74 | 0.81 | 0.61 |
| | 50 | 0.73 | 0.76 | 0.83 | 0.62 |
| | 70 | 0.75 | 0.77 | 0.85 | 0.65 |
| Medium = 20 | 30 | 0.66 | 0.72 | 0.73 | 0.55 |
| | 50 | 0.69 | 0.73 | 0.74 | 0.60 |
| | 70 | 0.71 | 0.75 | 0.78 | 0.64 |
| High = 50 | 30 | 0.64 | 0.69 | 0.72 | 0.53 |
| | 50 | 0.67 | 0.71 | 0.73 | 0.56 |
| | 70 | 0.70 | 0.73 | 0.75 | 0.60 |

Table 3.3.1: Corresponding testing accuracies for different rejection options for the simulated dataset

loss and regularization now are on a common probabilistic scale, since both terms are based on probability distributions and KL-divergence. Hence, there is a relationship between defining a prior distribution over margins and parameters and defining the objective function and the penalty term in the original function. Recall that, $\gamma_n$ are the classification margins as slack variables in the optimization which represent the minimum margin that $y_n\mathcal{D}(X_n; w)$ must satisfy. Hence, the choice of the margin distribution corresponds to the use of the slack variables in the formulation of the UQ-CHI. For example, in our case we set $p_0\left(\gamma_n\right) = ce^{-c(1-\gamma_n)}$ and $\gamma_n \leq 1$. If we mathematically expand the normalization function in (3.10), we get the two terms $Z_{\mathbf{w}}(\lambda)$ and $Z_{\gamma}(\lambda)$ as shown in (3.14), and given the choice of margin priors in Section 3.3.1 we get:

$$logZ_{\gamma_n}(\lambda_n) = log \int_{\gamma_n=-\infty}^{1} c \exp\Big(-c\big(1-\gamma_n\big)\Big) \exp\Big(-\lambda_n\gamma_n\Big) d\gamma_n. \tag{3.21}$$

From (3.21) we can see that a penalty occurs when the margins are smaller than $E[\gamma_n] = 1-\frac{1}{c}$, and any margins larger than this would not be penalized. The margin distribution becomes

peaked when $\gamma_n = 1$ that is when $c \to \infty$, and this is equivalent to having fixed margins. If the margin values are held fixed the discriminant function might not be able to separate the training examples with such pre-specified margin values. Because of non-separable datasets this will generate an empty convex hull for the solution space. Thus, we need to revisit the setting of the margin values, and the loss function upon them. The parameter $c$ will play an almost identical role as the regularization parameter which upper bounds the Lagrange multipliers. Note, if the objective function $J()$ grow without a bound, it may generate a search space for parameters that are no longer a convex hull. This compromises the uniqueness and solvability of the problem. Therefore, the selection of a prior forms a concave function $J()$ for a unique optimum in the Lagrange multiplier space.

## *3.4 Numerical studies*

In this section, we design our simulation studies to evaluate the efficacy of UQ-CHI in terms of prediction and uncertainty quantification, when compared to the CHI model, under a variety of practical scenarios.

### *3.4.1 Simulated dataset*

We simulate data following the procedure described below. The synthetic dataset is generated with two classes with partial labels. We conduct several experiments with the simulated data to investigate the performance of our method across different settings. Without loss of generality, we assume that there are two groups, normal vs. diseased with a proportion of 60% of class normal and 20% of complete labels. For all the experiments, we set the number of features $d = 90$, For each class, we simulate 50 subjects, where we assumed that $x_{n,t}^k \sim N\big(u, \sigma_k^2\big)$ for $k \in \{1, \ldots, d\}$.

### *3.4.2 Incomplete labels and length of longitudinal data*

UQ-CHI can also model data with partial labels by assigning a prior distribution of the labels and obtaining posterior distributions after model training. In our experiment, we consider a low, medium and high level of label availability, i.e., 10%, 20% and 50% of unlabeled

examples. Also, we evaluate our methodology's robustness in the presence of down-sampling of the training data, i.e., only using a percentage of the data (for example, ranging from 30%, 50% and 70%), to train both UQ-CHI and CHI models. A model that can predict well with less longitudinal data holds great value in clinical applications.

### *3.4.3 Uncertainty quantification with rejection option*

As mentioned in 3.3.3, UQ-CHI has the unique capability of only classifying all examples that demonstrate a sufficiently high posterior probability. In this case, the classifier rejects to predict on a sample if it cannot be predicted reliably. The key parameter is the threshold ($\kappa$) that will be used in the rejection option. Here, the rejection rate is defined as the probability that the classifier rejects the example,

$$p(reject) = \int p\left(x\right) dx = p(p(\mathbf{w}|x_i) < \kappa) \tag{3.22}$$

We can define two types of error when classifying with the rejection option. The error, $\varepsilon = p(\widehat{y} \neq y\ )$, which is the probability of making an incorrect classification, The conditional error $\varepsilon^{cond} = p(\widehat{y} \neq y|accept)$, which is the probability of making an incorrect classification, given the classifier has accepted the example. There is a general relationship between the error and rejection rate: the error rate decreases monotonically with increasing rejection rates, thus implying that the classifier is more reliable. Therefore, incorporating a rejection option removes uncertain samples and improves the prediction accuracy of classifiers.

The rejection option relies upon the threshold ($\kappa$) parameter. In our experiments, we use several levels of the threshold ($\kappa$) to create a range of rejection options from loose to strict, and further calculate the resulting accuracies on the predictions on the accepted samples. Specifically, we examined the size of the rejection region from 20%, 40%, to 60%.

### *3.4.4 Parameter tuning and validation*

In our experiments, we randomly split the data into the two parts, one for training and one for testing sets. For the training dataset, we use 10-fold cross-validation to tune the

parameters. The average accuracies from the split of the testing dataset are reported in the result section. In Section 3.3.4 we specify the conditions under which the computation remain tractable. It has been pointed out that, based on the choice of the margin distribution described in 3.3.4, $\gamma_n$ is bounded by the parameter $c$. Recall that $c$ is a parameter in the prior for the margins. Therefore, the parameter $c$ will play an important role. For this reason we conduct experiments with the parameter $c$ chosen from $(1.5, 3, 5, 10, 20, 100)$ to see the impact of various choices of $c$ on the testing accuracy.

### *3.4.5 Discussion*

We discuss the tractability of the model given the simulated data for various choices of the parameter $c$ in Table 3.4.1. We simulated different selection of the parameter $c$ to evaluate its impact on the testing accuracy. The accuracy of a classifier is defined as the probability of making a correct decision. If we discover that increasing this parameter imposes little effect on model performance, we would then ignore the higher values for reasons already explained in Section 3.3.4. We found that model accuracy decreases with increasing values of parameter $c$. As shown in Table 3.4.1, additional potential terms of the parameter $c$ have minimal effects as the margin distribution may have become at its peak ($\gamma_n$)) which is equivalent to having fixed margins. Note that to fully evaluate the impact of the parameter $c$ we simulated the data with a proportion of 60% of class normal and 20% complete labels. Here, we found that after increasing the values for the parameter $c$ beyond 5, the performance of the model doesn't change significantly. This indicates that the margin distribution may have reach its peak, and hence it is equal to a fixed value. Higher values of this parameter generate relatively similar performances. Consequently, lower values of $c$ preserve flexibility to estimate a distribution over parameters instead of using fixed margins.

Next, we examined how incomplete label information could affect the performance of UQ-CHI with regards to the testing accuracy given different sampling ratios in Table 3.4.2. A model which can be built with less training data is more promising in healthcare applications because data collection is relatively more expensive compared to other real-world applica-

tions. Therefore, we tested our model by randomly down-sampling the training data, i.e., only using a proportion of the data with different sampling ratios (30%, 50% and 70%). The results showed that the UQ-CH performed reasonably with a testing accuracy of 74% in the extreme case where a small proportion of training samples (30%) was used and 50% of incomplete label information was available (Table 3.4.2).

Error rates on the testing set for different rejection options are reported in Table 3.5.1. As depicted in Eq. (3.22), the rejection rate is defined as the probability that the classifier rejects the example. Hence, the classifier becomes more reliable as the rejection rate increases and as a result the probability of making incorrect classification decreases. Comparisons of varying rejection rates for the UQ-CHI confirms this expectation. For a high rejection rate of 60%, the testing error reduces to 0.19, whereas a lower rejection rate of 20% resulted in a testing error of 0.31. Our methodology was also compared to the CHI framework (Table 3.5.1). Recall that CHI is not strictly a supervised learning problem. In [28], both simulation studies and real-world applications demonstrated that without label information, it was still possible to train the CHI method. Here, we found that the UQ-CHI can outperform CHI by incorporating the rejection option. UQ-CHI can obtain a testing error in a range of 0.19% to 0.25% for a given rejection rate of 60%.

Table 3.4.1: Model average testing accuracies (%) for simulated dataset

| Parameter c | Testing accuracy |
|---|---|
| 1.5 | 81.2 |
| 3 | 80.2 |
| 5 | 79.8 |
| 10 | 77.2 |
| 20 | 77.3 |
| 100 | 76.1 |

Table 3.4.2: The average testing accuracies and standard deviations for the simulated dataset for various sample and label ratio

| Sample ratio | Label ratio (%) | | |
|---|---|---|---|
| | Low = 10 | Medium = 20 | High = 50 |
| 30 | 0.85 ± 0.033 | 0.80 ± 0.032 | 0.74 ± 0.033 |
| 50 | 0.86 ± 0.060 | 0.83 ± 0.053 | 0.76 ± 0.027 |
| 70 | 0.88 ± 0.074 | 0.85 ± 0.041 | 0.78 ± 0.037 |

### *3.5 Real-world application on Alzheimer's disease*

We tested the predictive utility of UQ-CHI using Alzheimer's disease data which exhibited monotonic disease progression. We use the FDG-PET images of 162 patients (Alzheimer's Disease: 74, Normal aging: 88) downloaded from ADNI (`www.loni.usc.edu/ADNI`). The data represents samples taken at irregular time points where each patient has between 3 and 7 time points. The data is preprocessed, and the Automated Anatomical Labeling (AAL) is used to segment each image into 116 anatomical volumes of interest (AVOIs). For this study, 90 AVOIs that are located in the cerebral cortex are selected (each AVOI becomes a variable here). According to the mechanism of FDG-PET, the measurement data of each region are the local average FDG binding counts, which represents the degree of glucose metabolism. The glucose metabolism declines as the function of age, and the progression of many neurodegenerative diseases such as AD further accelerates this declination. Thus, ADNI dataset provides an optimal application example to test the proposed method. Whereas the ADNI dataset consists of fully labeled examples, we artificially induce a variety of uncertainties to the label information.

The results for tuning the parameter $c$ for the ADNI dataset are reported in Table 3.5.2. This shows decreasing accuracy with increasing values of parameter $c$. The performance of UQ-CHI across different uncertainty levels as well as different sampling ratios was also evaluated (Table 3.5.3). The proposed method demonstrated excellent capability to quantify

Table 3.5.1: Corresponding testing errors for different rejection options for the simulated dataset

| Algorithm name | UQ-CHI testing error | | | | CHI testing error |
|---|---|---|---|---|---|
| Label ratio (%) | Sampling ratio (%) | Rejection rate (%) | | | |
| | | Low = 20 | Medium = 40 | High = 60 | |
| Low = 10 | 30 | 0.31 | 0.26 | 0.19 | 0.39 |
| | 50 | 0.27 | 0.24 | 0.17 | 0.38 |
| | 70 | 0.25 | 0.23 | 0.15 | 0.35 |
| Medium = 20 | 30 | 0.34 | 0.28 | 0.27 | 0.45 |
| | 50 | 0.31 | 0.27 | 0.26 | 0.40 |
| | 70 | 0.29 | 0.25 | 0.22 | 0.36 |
| High = 50 | 30 | 0.36 | 0.31 | 0.28 | 0.47 |
| | 50 | 0.33 | 0.29 | 0.27 | 0.44 |
| | 70 | 0.30 | 0.26 | 0.25 | 0.40 |

uncertainties in the real-world dataset. As shown in Table 3.5.3, UQ-CHI is even capable of handling data that has 50% of incomplete labels; in this case demonstrating an accuracy in the range of $70\% - 77\%$ for the ADNI dataset.

By contrast, we show that by only using a small proportion of the training samples (as low as 30% of the data) that it is still possible to maintain reasonable performance in a range of $70\% - 82\%$. This suggests that UQ-CHI for the real-world dataset can be trained with less training data. The rejection options against the testing error, as well as these values against the sampling ratios, are shown in Tables 3.5.4. Comparisons of different rejection rates for the UQ-CHI confirmed the capability of the rejection ratio for the real-world applications.

## *3.6 Conclusion*

In this paper, we developed the UQ-CHI methodology to enable uncertainty quantification for continuous patient monitoring. This probabilistic generalization can facilitate further extensions of the basic CHI model for decision-making purposes. For example, in many degenerative disease conditions such as AD, it is essential to triage patients to determine the

Table 3.5.2: Model average testing accuracies (%) for ADNI dataset

| Parameter c | Testing accuracy |
|---|---|
| 1.5 | 78.8 |
| 3 | 77.9 |
| 5 | 77.3 |
| 10 | 75.3 |
| 20 | 72.0 |
| 100 | 68.9 |

Table 3.5.3: The average testing accuracies and standard deviations for ADNI dataset for various sample and label ratio

| Sample ratio (%) | Label ratio | | |
|---|---|---|---|
| | Low = 10 | Medium = 20 | High = 50 |
| 30 | 0.82 ± 0.022 | 0.79 ± 0.052 | 0.70 ± 0.032 |
| 50 | 0.84 ± 0.014 | 0.82 ± 0.005 | 0.74 ± 0.049 |
| 70 | 0.87 ± 0.040 | 0.83 ± 0.032 | 0.76 ± 0.043 |

Table 3.5.4: Corresponding testing errors for different rejection options for the ADNI dataset

| Algorithm name | UQ-CHI testing errors | | | | CHI testing errors |
|---|---|---|---|---|---|
| Label ratio (%) | Sampling ratio (%) | Rejection rate (%) | | | |
| | | Low = 20 | Medium = 40 | High = 60 | |
| Low = 10 | 30 | 0.29 | 0.24 | 0.17 | 0.26 |
| | 50 | 0.25 | 0.22 | 0.16 | 0.34 |
| | 70 | 0.23 | 0.21 | 0.13 | 0.20 |
| Medium = 20 | 30 | 0.33 | 0.29 | 0.28 | 0.42 |
| | 50 | 0.30 | 0.28 | 0.25 | 0.38 |
| | 70 | 0.29 | 0.25 | 0.24 | 0.36 |
| High = 50 | 30 | 0.44 | 0.30 | 0.29 | 0.45 |
| | 50 | 0.31 | 0.29 | 0.26 | 0.42 |
| | 70 | 0.29 | 0.28 | 0.26 | 0.38 |

prioritization of resource allocations and patient care. The UQ-CHI framework can support optimal decision making that considers imperfect and continuous delivery of knowledge. In the future, it may be advantageous to extend this method to other diseases that may show different degradation characteristics in the context of degenerative diseases. Another extension of this methodology is to adapt for a non-linear index and further explore the feasibility of varying discriminant functions.

# Chapter 4

# DETECT DEPRESSION FROM COMMUNICATION: HOW COMPUTER VISION, SIGNAL PROCESSING, AND SENTIMENT ANALYSIS JOIN FORCES

Depression is a common illness worldwide. Traditional procedures have generated controversy and criticism such as accuracy and agreement consistency of depression diagnosis and assessment among clinicians. More objective biomarkers are needed for better treatment evaluation and monitoring. Depression will leave recognizable markers in patient's vocal acoustic, linguistic, and facial patterns, all of which have demonstrated increasing promise on evaluating and predicting patient's mental condition in a more objective way. We applied a multi-modality prediction model to combine the audio, video, and text modalities, to identify the biomakers that are predictive of depression with consideration of gender differences. We identified promising biomarkers from successive search on feature extraction analysis for each modality. We found that gender disparity in vocal and facial expressions play an important role in detecting depression. Audio, video and text biomarkers provided the possibility of detecting depression in addition to traditional clinical assessments. Biomarkers detected for gender-dependent analysis were not identical, indicating that gender affect the depression manifestations.

## *4.1 Introduction*

Depression has been recognized as a significant health concern worldwide. According to the World Health Organization (WHO), more than 300 million people suffer from depression in their daily lives [1]. Psychiatric diagnoses have traditionally been made based on clinical

[1] http://www.who.int/mediacentre/factsheets/fs369/en/

criteria. For example, the Diagnostic and Statistical Manual of Mental Disorders (DSM) is a standard diagnostic tool published by the American Psychiatric Association APA, and is used by healthcare professionals as an authoritative guide to the diagnosis of mental disorders [93]. DSM contains descriptions, symptoms, and many other criteria for diagnosing a mental disorder. Alternatively, clinicians can also monitor the severity of depression using the Patient Health Questionnaire depression scale (PHQ-8) questionnaire, a self-administered tool consisting of 8 symptom questions [94]. However, such traditional procedures have generated controversy and criticism such as accuracy and agreement consistency among clinicians due to their subjective nature [95]. Therefore, determining an accurate and feasible screening tool in the general population remains important.

An accurate diagnosis that is made in a timely manner has the best opportunity for a positive health outcome since the subsequent decision-making will be tailored to a correct understanding of the patient's health problem [96]. It is desirable to develop more biomarkers that can be automatically extracted from objective measurements. As audio, video, and text modalities are very useful data for characterizing human interaction and communications, they have been used in many applications such as speech processing [97], safety and security applications [98] and human-machine interaction [99]. Also, since depression is a complex disease that manifests in all of these modalities, many research efforts have been directed toward using these datasets for depression prediction and evaluation. Studies show that depression will leave recognizable markers in a patient's vocal acoustic, linguistic and facial patterns, all of which have demonstrated increasing promise in the objective evaluation and prediction of a patient's mental state [5, 6, 7]. Central to the success of these models, extracting useful quantitative biomarkers from the audio, video, and text data have resulted in the discovery of many interesting biomarkers.

Audio biomarkers are powerful tools to assist in detecting depression severity and monitoring over the course of a depressive disorder. For example, Alpert et al. suggested that intra-personal pause duration and speaking rate are closely related to change in depression severity

over time [100]. Several studies have found distinguishable audio biomarkers such as pitch [101, 102], decreased speech intensity [103], loudness [104], and energy [105], that are useful for detecting depression severity. There have been few studies documenting the relationship between audio biomarkers and depression severity using clinical subjective ratings [7, 106, 107]. Overall, they found that audio biomarkers discriminate between individuals with and without depression. On the other hand, previous studies have shown that word and phrase-related biomarkers can influence the classification performances [108, 109]. For example, in [110], they showed that selecting a few phrases can improve depression prediction. Similarly, facial-based video biomarker sets extracted from dynamic analysis of facial expressions were also identified to be predictive for depression [111, 112, 113]. Video biomarkers can be extracted via head pose movement analysis [114], automatic facial action coding system [115], etc. To date, there has been no systematic study that investigates biomarkers with respect to all of these modalities together, nor has any study attempted to further investigate their clinical interpretations and interactions with gender.

Depression is a complex mental disorder that could not be solely captured from one single modality. Modalities that integrate acoustic, textual and visual biomarkers to analyze psychological distress have shown promising performances [116, 117]. In these mechanisms, each acquisition framework is denoted as a modality and is associated with one dataset. In a recent study [118], researchers have investigated the performance of each of the modalities (audio, linguistic, and facial) for the task of depression severity evaluation by a multi-modal mechanism. Further, authors improved the results by using a confidence-based fusion mechanism to combine all three modalities. Experiments on the recently released AVEC 2017 [117] depression dataset have verified the promising performance of the proposed model in [118]. In this work, we extended [118], and identified unique biomarkers from audio, linguistic, and facial biomarkers that were predictive of depression. Compared with [118], in this paper, we identified a much more comprehensive array of biomarkers in all three modalities and further studied their clinical interpretations. For example, we captured prosodic characteristics of

the speaker and quality of voice in both time and frequency domains. We learned discriminative audio biomarkers such as low-pitched voice and multiple pauses during a speech that were predictive to depression. For video biomarkers, because of the unavailability of raw video recordings, we extracted facial biomarkers as a complementary cue that was measured simultaneously with other cues (e.g. speech, and text). Hence, eye, eyebrow, mouth and head movement biomarkers were used for recognizing depression. In the text analysis, we captured valence ratings of the text using the tool of AFINN sentiment analysis. In addition, we identified depression-related words, a total number of words and sentences, which were nonverbal manifestations of depression.

The combination of these modalities and their interactions with gender information were also investigated. A recent study showed that gender plays an important role in depression severity assessment [119]. Similarly, in our study, we found that biomarkers detected for gender-dependent analysis were not identical, indicating that gender can affect manifestations of depression. This result is consistent with previous findings on gender differences which show that depression in women may be more likely to be detected than depression in men [120]. We found that this might be related to the theory that women are more likely to amplify their mood. For example, women are likely tending to employ positive emotions including expressing joy or laughter [121]. Research has found that facial movement dynamics are discriminative for females and males. For example, men's movement may be more asymmetrical [122] and women's more animated [123]. Hence, there is a need to investigate a subset of biomarkers that contribute significantly to the target class (depressed) and are unique to women and men. Furthermore, existing works demonstrated that including gender information is effective in improving the prediction performance [124, 125, 126, 127]. This suggests that it is reasonable to assess the performance of the multi-modality fusion model based on gender-bias. In the present work, a gender-based classification system was implemented by building two separate classifiers, training two gender-specific models for male and female separately. In addition, we investigated significant biomarkers that had domi-

nant importance for prediction in audio, video and text modalities using biomarker selection methods. Finally, we investigated the clinical implications of these observations and their relationships with respect to depression detection.

This paper is structured in the following manner. Section 4.2 describes the process of extracting biomarkers from three modalities. The results of biomarkers that are predictive to depression along with their performances are shown in Section 4.3. Discussions and conclusion are drawn in Section 4.4.

## *4.2 Methods*

### *4.2.1 Data collection*

The database is part of a larger corpus, the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ), that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety and depression. The dataset includes interviews, conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room, and the participants include both distressed and non-distressed individuals. Participants were drawn from two distinct populations living in the greater Los Angeles metropolitan area and from the general public. Participants were coded for depression based on PHQ-8.

Of the 140 participants, 76 were male, 63 were female, and 1 did not report his/her gender. The mean age of this group of participants was 39.34 (SD = 12.52). The collected data include audio and video recordings and extensive questionnaire responses. During the interview, the participant was recorded by a camera and a close-talking microphone. Data were transcribed and annotated for a variety of verbal and non-verbal features combining both participant and interviewer. However, the focus of this paper is on the participant speech, and the transcription part of Ellie was not considered. Information on how to obtain shared data can be found at this location: [2]. Data is freely available for research purposes.

[2] `http://dcapswoz.ict.usc.edu`

### *4.2.2 Study population*

The Distress Analysis Interview Corpus database consists of the word-level human transcriptions including auxiliary speech behaviors, raw audio recordings and facial landmarks along with the gender of participants provided in two partitions training set ($n = 107$), and development set ($n = 36$). Socioeconomic background was not provided by the dataset. The multi-modality fusion model was employed on the training set, and we validated the model with the given extracted biomarkers on the development set. For the generalization purpose, we used 10-fold cross-validation in all the experiments. The average depression severity on the training and development set is $M = 6.67$ ($SD = 5.75$) out of a maximum score of 24. The level of depression is labeled with a single value per recording using a standardized self-assessed subjective depression questionnaire, the PHQ-8 score, which was recorded prior to every human-agent interaction. The distribution of the depression severity binary labels (PHQ8 Scores $\geq 10$) on the training and development set for females and males can be seen in Figure 4.2.1.

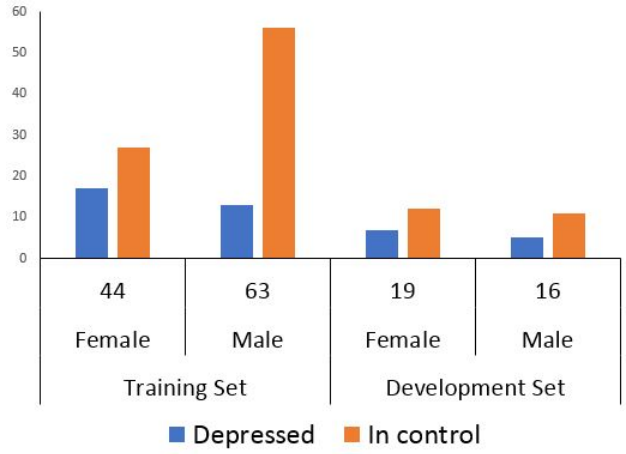



Figure 4.2.1: Distribution of the depression severity on the training and development set

### *4.2.3 Biomarkers*

To detect biomarkers that are predictive of depression, we adopted a separate processing pipeline for each modality, followed by a multi-modality methodology to predict the result using the proposed method in [118]. The settings of individual prediction models and the multi-modal fusion approach are discussed in the next section.

*Overview of the multi-modality fusion model*

In this work, we used a multi-modal fusion approach to combine information across different modalities. Here, we used modality loosely to refer to different biomarkers extracted from audio, video and text datasets. The audio recordings include soundtracks with significant communicative cues that are symptomatic of depression. The 2D facial landmarks are indicative of changes in facial expression, which could be further analyzed for depression. In addition, people's emotional status may change from sentence to sentence, and from word to word. The fusion of these multiple modalities can provide complementary information and increase the accuracy of the overall decision-making process. However, the benefit of multi-modal fusion comes with a certain cost and complexity in the analysis process. This is due to the fact that different modalities are captured in different formats and at different rates. Also, different modalities usually have varying confidence levels. Therefore, building a robust multi-modality framework that is capable of dealing with complex datasets is essential.

Existing multi-modality fusion techniques can be categorized as a biomarker-level fusion and a decision-level fusion [128]. A biomarker-level fusion learns shared or similarity-regularized biomarkers across different modalities and then performs classification based on a multi-modal representation vector. On the other hand, a decision-level fusion learns an input-specific classifier for each modality and finds a decision rule to aggregate decisions made based on a single modality. In the present work, we found that each modality contained highly varying data volume and sample dimensionality, therefore, we chose separate biomarker processing pipelines for each modality, followed by a decision-level fusion module to predict

the final result.

For each modality, a random forest model was built to convert the predictive information of these biomarkers into scores, and further combined them into a confidence-based fusion method to make a final prediction on the PHQ-8 scores. Random Forest is an ensemble of unpruned classification created by using bootstrap samples of the training data and random biomarker selection in tree induction. Usually, the prediction of random forest is made by aggregating (majority vote or averaging) the predictions of the ensemble [129]. However, the dataset is coming from three different modalities, e.g., audio, video, and text, hence compromising them by aggregating the predictions from all the trees is not recommended. Therefore, for each modality we calculated the standard deviation of predictions from all the trees, defined as the modality-wise confidence score.

Let us assume a set of $M$ randomized decision trees within a random forest $\{\varphi_{L,\theta_m}|m = 1, \ldots, M\}$ all learned on the same data $L$ (e.g., audio, video, or text). In order to improve the classification accuracy over decision trees, the individual trees in a random forest need to be different, which can be achieved by introducing randomness in the generation of the trees. The randomness is influenced by the seed $\theta_m$ for each model. Ensemble methods work by combining the predictions of these models into a new ensemble model, denoted by $\overline{\psi}_{L,\theta_1,\ldots,\theta_m}$. The randomized models are combined into an ensemble by computing the standard deviation $(S)$ of the predictions to form the final prediction as:

$$S = \sqrt{\frac{\sum_m^M \left(\varphi_{L,\theta_m} - \overline{\psi}_{L,\theta_1,\ldots,\theta_n(x)}\right)^2}{M-1}}. \tag{4.1}$$

Here, $\overline{\psi}_{L,\theta_1,\ldots,\theta_n(x)}$ is the average prediction which is defined as:

$$\overline{\psi}_{L,\theta_1,\ldots,\theta_n(x)} = \frac{1}{M}\sum_m^M \varphi_{L,\theta_m}. \tag{4.2}$$

After conducting several different strategies, the winner-take-all strategy, i.e., picking the

single-modality prediction with the highest confidence score as the final result, seemed to be the most effective and reliable one in our setting. The overview of the multi-modal fusion framework is shown in Figure 4.2.2. We identified many biomarkers that were predictive to depression with the help of signal processing on the audio files, computer vision on 2D facial coordinates and sentiment analysis on individual's transcript files. We will next introduce how we extracted biomarkers based on the given data for each modality.

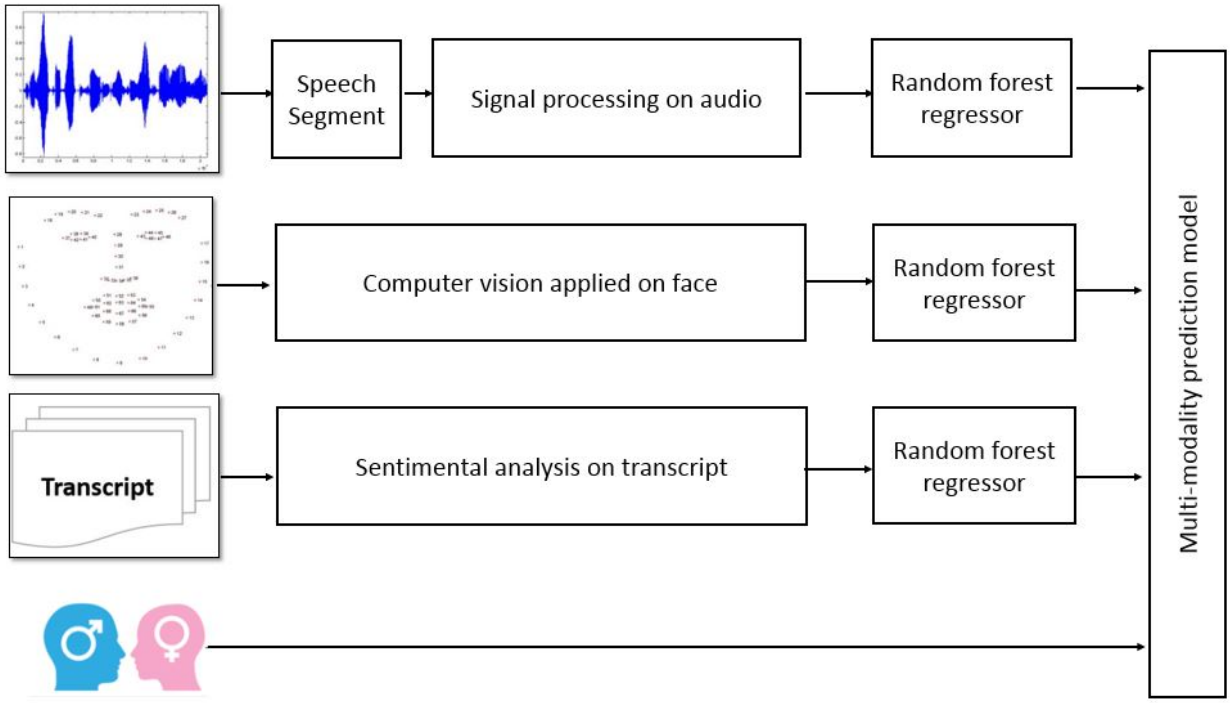



Figure 4.2.2: Fusing audio, facial and textural biomarkers with the help of a multi-modality fusion model

*Audio biomarkers extracted by signal processing*

Audio biomarkers discriminate between individuals with and without depression. The first step in the speech recognition system is to extract biomarkers, i.e., to identify the components of the audio signal that have relations in a clinical context, and to distinguish other signals which carry information like background noise, pause, etc. In most audio processing and

analysis techniques, it is common to divide the input audio signal into short-term frames before biomarker extraction. In particular, the audio signal is broken into non-overlapping, short frames and a set of biomarkers are extracted for each frame. An audio signal is changing very rapidly, hence we assume that a signal doesn't change statistically (i.e. statistically stationary) in short-term frames. We, therefore, broke each audio signal into 20-40 ms frames. The result of this procedure was a sequence of biomarker vectors per audio signal.

A discriminative biomarker can be learned from either the time domain or the frequency domain. For example, analyzing audio signals with respect to time provides information with regards to the value of the signal at any given instance. Hence, time domain is useful when amplitude or energy of a signal needs to be examined. On the other hand, time domain doesn't convey information with respect to the rate at which the signal is varying. Therefore, representing the signal in a domain where the frequency of a signal is described is needed. Frequency-based biomarkers (spectral biomarkers) are obtained by converting a time-based signal into a frequency domain by using Fourier transform. For example, one of the most dominant methods used to extract frequency-based biomarkers is calculating Mel-Frequency Cepstral Coefficients (MFCC). MFCC biomarkers are derived from a cepstral representation of an audio clip (a Fourier transform of the power spectrum of a voice signal). MFCC of a signal extracts biomarker from the speech which approximates the human auditory system's response, while at the same time deemphasizing all other information. Another example of the frequency-based technique is Perceptual Linear Prediction (PLP). PLP biomarkers are obtained based on the concept of the psychophysics of hearing and they are also used to derive an estimate of the human auditory spectrum. Differences between PLP and MFCC lie in the intensity-to-loudness conversion of a sound.

In our setting, we extracted a set of validated and tested biomarker extraction techniques that aim to capture prosodic characteristics of the speaker as well as the quality of voice in both the time and frequency domains. Descriptions of audio biomarkers in both time and frequency domains are shown in Table4.2.1 and 4.2.2. A total of 35 audio biomarkers were

used.

| **Audio biomarkers** | **Description** | **No. of biomarkers** |
|---|---|---|
| Modulation of amplitude | It is used to find the amplitude of two signals that are multiplied by the superimposed signals. | 1 |
| Envelope | It represents the varying level of an audio signal over time. | 1 |
| Autocorrelation | It shows the repeating patterns between observations as a function of the time lag between them. | 1 |
| Onset detector | It is used to detect,a sudden change in the energy or any changes in the statistical properties of a signal. | 1 |
| Entropy of energy | It is a measure of abrupt changes in the energy level of an audio signal. | 1 |
| Tonal power ratio | It is obtained by taking the ratio of the tonal power of the spectrum components to the overall power. | 1 |
| RMS power | Root mean square (RMS) approximates the volume of an audio frame. | 1 |
| ZCR | Zero Crossing Rate (ZCR) is the number of times the signal changes sign in a given period of time. | 1 |

Table 4.2.1: Description of audio biomarkers used in a time domain

*Visual biomarkers extracted by computer vision*

Several biomarkers were extracted from the set of the 2D coordinates of the 68 points [130] on the face provided by the AVEC'17, despite the unavailability of video recordings. Hence, a preprocessing on the 2D facial landmarks was done to obtain head and distance biomarkers. Overall, we obtained 133 video biomarkers using both head motion and distance with complementary statistical measures.

1. *Head biomarkers*: Initially, we drew the 2D polygons of facial landmarks using patch function in MATLAB. Patch function creates several polygons using elements of 2D coordinates for each vertex, which connects the vertices to specify them. Such polygons, which encode facial landmark regions as well as temporal movements, have resulted in different video frames. Landmark motion history using patch function is shown

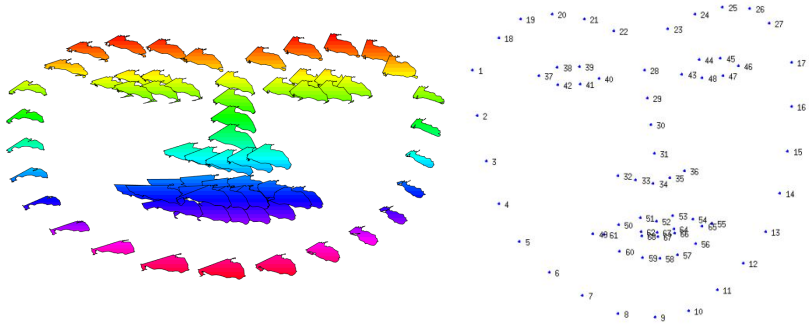
| **Audio biomarkers** | **Description** | **No. of biomarkers** |
|---|---|---|
| PLP | It is a technique to minimize the differences between speakers. | 9 |
| MFCC | It is a representation of the short-term power spectrum of an audio signal. | 12 |
| Spectral decrease | It computes the steepness of the decrease of the spectral envelope. | 1 |
| Spectral rolloff | It can be treated as a spectral shape descriptor of an audio signal. | 1 |
| Spectral flux | It is a measure of spectral change between two successive frames. | 1 |
| Spectral centroid | It is a measure to characterize the center mass of the spectrum. | 1 |
| Spectral slope | It is the gradient of the linear regression of a spectrum. | 1 |
| Spectral autocorrelation | It is a function that measures the regular harmonic spacing in the spectrum of the speech signal. | 1 |

Table 4.2.2: Description of audio biomarkers used in a frequency domain

in Figure 4.2.3 a. blueThe points, which represent head motions and have minimal involvement in facial movement, were chosen. Points 2, 4, 14, and 16 are among these landmarks, as shown in Figure 4.2.3 b, [131]. In addition, given that the regions between eyes and mouth are the most stable regions due to their minimal involvement in facial expression, we followed the approach in [132], and calculated the mean shape of 46 stable points without regard to gender. For each of these facial points, statistical measures such as mean, and median were calculated, resulting in 41 biomarkers.

2. *Distance biomarkers*: Both head motion and facial expression can be an indicator of a person's behavior, and may convey sets of essential information regarding the emotional state of a person. We split the facial landmarks into three groups of different regions: the left eye and left eyebrow, the right eye and right eyebrow, and the mouth. The pairwise Euclidean distance between coordinates of the landmarks was calculated, as

well as the angles (in radians) between the points, resulting in 92 biomarkers. In detail, for the left and right eye, the distance between points {37, 40} and {43, 46} were measured as horizontal distance, and the distance between {38, 42} and {44, 48} were measured as vertical distance respectively. For the mouth, the distance between points {52, 58} were measured as vertical distance, an average of the distance between two pairs of points {55, 49}, {65, 61} were measured as horizontal distance. Usually, eyebrow movements occur simultaneously. Hence we took the average distance between two pairs, using points {22, 23} and {27, 18}, as horizontal measures, and {31, 25}, {31, 20} as vertical distances. We calculated the difference between the coordinates of the landmarks, and those from the mean shape, and finally calculated the Euclidean distance (L2-norm) between the points for each group.



(a) Landmark motion history using patch function

(b) Facial landmark numbering

Figure 4.2.3: Facial landmark motion and facial landmark numbering

*Text biomarkers extracted by sentimental analysis*

A set of text biomarkers were calculated based on the individual's transcript file. The original transcript file included translated communication content between each participant and the animated virtual interviewer (named as 'Ellie' who is controlled by a human interviewer), as well as the duration (start time and stop time) for each round of communication. The focus of this paper deals solely with participant speech. The transcription part of Ellie

was not considered. The first part of text biomarkers was about basic statistics of words or sentences from the transcription file. This included the number of sentences over the duration, the number of words,the ratio of the number of instances of laugher over the number of words. The use of these biomarkers is supported by the literature, indicating that slowed and reduced speech, elongated speech pauses, and short answers are nonverbal manifestations of depression [133]. The second category of text biomarkers concerned depression-related words, namely, the ratio of depression-related words over the total number of words over the duration. Words relating to depression were identified from a dictionary of more than 200 words, which can be downloaded from the Depression Vocabulary Word List [3]. In addition, we introduced a new set of text sentiment biomarkers, obtained using the tool of AFINN sentiment analysis [134], that would represent the valence of the current text by comparing it to an existing word list with known sentiment labels. The outcome of AFINN sentiment analysis is an integer between minus five (negative) and plus five (positive), where a negative or positive number corresponds, in turn, to a negative or positive sentiment. These ratings are calculated according to the psychological reaction of a person to a specific word, where each word is categorized for valence with numerical values. In addition, the mean, median, min, max, and standard deviation of the sentiment analysis outcomes (as a time series) were measured. A total of 8 biomarkers were extracted.

## *4.3 Experimental results*

### *4.3.1 Identifying biomarkers that are predictive to depression*

To discover promising biomarkers, we studied the details of the multi-modality fusion model proposed in [118] and compared the results to single modalities. Baseline scripts provided by the AVEC have been made available in the data repositories, where depression severity was computed using random forest regressor. We used the training and the validation set to pick the optimal number of trees. The model deviance was calculated based on the mean squared error between the training and the validation set. Figure 4.3.1 shows the model deviance on

[3] `https://myvocabulary.com/word-list/depression-vocabulary/`

both training set (blue) and validation set (yellow), as a function of the number of trees for audio dataset. Each random forest model aggregates the outcomes of 50 individual trees. According to our experiments, if only using audio biomarkers, then the performance of the random forest model gets saturated with around 25 trees. After blending video and text modalities, the random forest model demands a larger learning capacity and grows to 50 trees, until it meets another performance plateau.

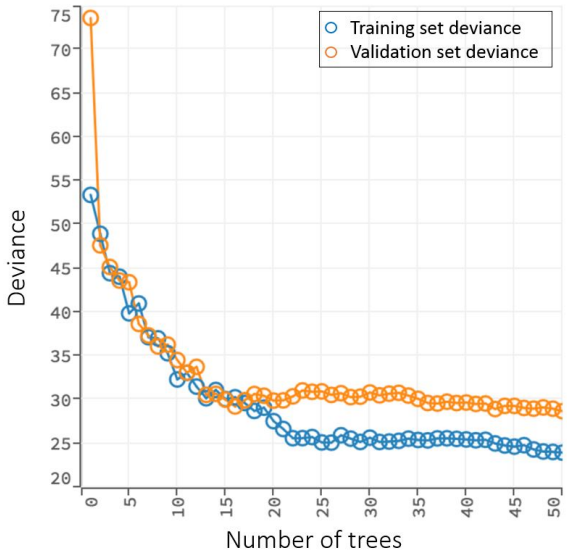



Figure 4.3.1: Scoring history over number of trees in prediction with audio biomarkers

We employed single modalities as well as the multi-modality fusion model proposed in [118] on the training set and validated the model using the development set. The depression severity baseline was computed using random forest regressor [117]. Baseline performance was provided by the AVEC to show the state-of-the-art prediction performance AVEC can achieve on this dataset. In the baseline, the fusion of audio and video modalities was performed by averaging the regression outputs of the unimodal random forest regressors. In addition, the root mean square error (RMSE) and mean absolute error (MAE) averaged over all observations were used to evaluate the models, while the best model was selected

based on RMSE. Table 4.3.1 reports the performance of single modalities, proposed multi-modality fusion model and the baseline for development and training set. The results show that fusing the audio, video and text biomarkers is powerful in detecting depression with RMSE of 5.12 and MAE 4.12, attained on development set compared with single modalities. In addition, the multi-modality fusion model had relatively better performance compared with the baseline results of RMSE 6.62 and MAE 5.52.

The performance of the multi-modality fusion model indicates promising values of capturing depression by integration of different biomarkers. The marginal differences between the multi-modality mechanism and single modalities might be due to many reasons. For example, the RSME and MAE of audio and text biomarkers are 6.00, 5.25 and 5.95 and 5.21 respectively. However, the performance for both RMSE and MAE of video for development set are 6.67 and 5.64 respectively. This indicates that fusing the video with other modalities might not have been very helpful. Moreover, the current dataset includes biomarkers and gender information only, and doesn't include other measures.

### *4.3.2 Investigating the gender-information on biomarkers*

Existing works demonstrated that including gender information is effective in improving the prediction performance [124, 125]. Thus, we investigated the effect of gender by comparing the performance of the model both with and without inputting subject gender. The multi-modality mechanism further improved the performance of depression severity prediction significantly with RMSE of 4.78 and MAE 4.05 attained on the development set with gender included as a variable shown in Table 4.3.1. Besides the biomarkers extracted from the three modalities, the gender of subjects was found to be highly correlated with the depression severity prediction. We proposed to include gender information, by training two gender-specific models for male and female separately.

Detailed results of gender-specific models can be found in Table 4.3.2. The multi-modality fusion model could attain RMSE of 5.09 and MAE 4.35 and RMSE of 4.95 and MAE 4.32 for females and males respectively. Although there are marginal differences in the multi-

| Biomarkers used | 'development' | | 'train' | |
|---|---|---|---|---|
| | RMSE | MAE | RMSE | MAE |
| The baseline provided by the AVEC organizer | | | | |
| Visual only | 7.13 | 5.88 | 5.42 | 5.29 |
| Audio only | 6.74 | 5.36 | 5.89 | 4.78 |
| Audio & Video | 6.62 | 5.52 | 6.01 | 5.09 |
| The model that doesn't include gender variable | | | | |
| Visual only | 6.67 | 5.64 | 6.13 | 5.08 |
| Audio only | 6.00 | 5.25 | 5.62 | 4.89 |
| Text only | 5.95 | 5.21 | 5.68 | 5.17 |
| Multi-modality fusion model | 5.12 | 4.12 | 4.25 | 4.54 |
| The model that includes the gender variable | | | | |
| Visual only | 5.65 | 4.87 | 4.99 | 4.46 |
| Audio only | 5.89 | 5.18 | 5.66 | 5.06 |
| Text only | 5.86 | 4.88 | 5.67 | 4.96 |
| Multi-modality fusion model | 4.78 | 4.05 | 4.35 | 3.69 |

Table 4.3.1: Performance comparison among single modalities and multi-modality fusion model

modality methods used between females and males, the performance of gender-specific models demonstrates variation in prediction. The performance of gender-specific models shows that individual audio and text modalities (except video) have a marginally better performance for males than females. This might be due to the gender-based differences in audio and text biomarkers. For example, the vocal characteristics of females and males are different, and detecting speech patterns might be affected by variation in pitch, loudness, and the rhythm of speech.

| Biomarkers used | 'development' | | 'train' | |
|---|---|---|---|---|
| | RMSE | MAE | RMSE | MAE |
| female | | | | |
| Visual only | 6.35 | 5.01 | 5.69 | 4.87 |
| Audio only | 6.45 | 5.38 | 5.03 | 4.66 |
| Text only | 6.41 | 5.68 | 5.06 | 4.65 |
| Multi-modality fusion model | 5.09 | 4.35 | 4.89 | 4.23 |
| male | | | | |
| Visual only | 6.44 | 5.18 | 5.92 | 4.89 |
| Audio only | 5.77 | 5.26 | 4.78 | 4.02 |
| Text only | 5.76 | 4.79 | 4.91 | 4.36 |
| Multi-modality fusion model | 4.95 | 4.32 | 4.91 | 4.02 |

Table 4.3.2: Performance comparison among single modalities and multi-modality by training gender-specific models

### *4.3.3 Investigating dominant biomarkers and their significance*

We extracted significant biomarkers from each modality to investigate gender differences. Selecting a subset of biomarkers that significantly contribute to the target class (depressed) can reduce the length of the process. Vocal, linguistic and facial significant biomarkers were attained from a successive search on biomarkers during the extraction analysis. Significant biomarkers were extracted by computing estimates of predictor importance based on permutation. Below, we provide brief details of how it works. The predictor importance measures how influential a candidate predictor variable is at the predicting response variable. The influence of a predictor increases with the value of this measure, hence larger values indicate predictors that have a greater influence on the prediction. For example, if a predictor is influential in prediction, then permuting its value should affect the error of the model. If a predictor is not influential, then permuting its values should have little to no effect on the model error.

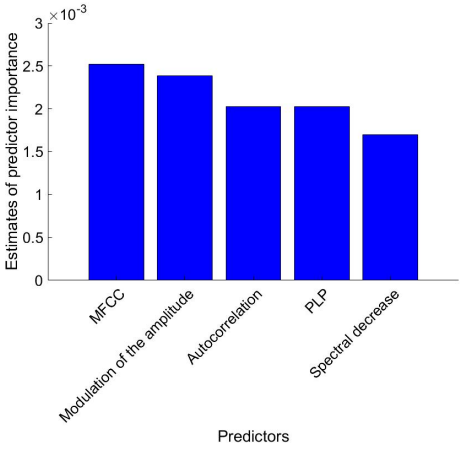
Figure 4.3.3 shows the top 5 significant biomarkers from each individual modality for females and males. Furthermore, we found the significant $p$-value for each top biomarker in relation to the target class by using the multiple linear regression shown in Table 4.3.3. The resulted significant biomarkers behave differently for females and males in each modality. In addition, a crossover interaction was observed between gender. For example, MFCC ranked the highest in the list for females, while PLP and spectral decrease were ranked the lowest. However, spectral slope and PLP were both ranked high for males. As a result, in-depth clinical interpretation is needed to decipher the clinical implications of these observations, and their relationship with respect to depression detection.

## *4.4 Discussion*

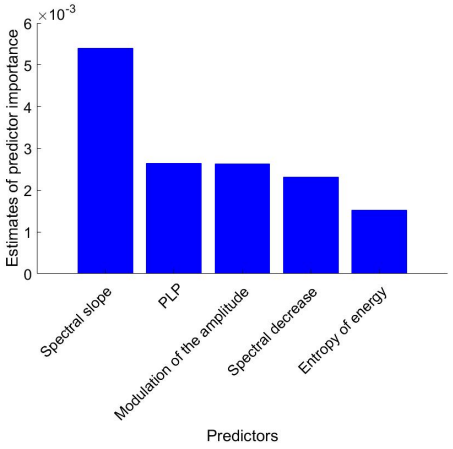
In this paper, we studied the details of the multi-modality fusion model proposed in [118] by integrating vocal, facial and textual modalities to discover promising biomarkers that were predictive of depression. The impact of the biomarker selection method was exploited to highlight each biomarker's communicative power for depression detection, and crossover interactions were observed in gender-specific predictions. Developing biomarkers that can be automatically extracted from objective measurements demonstrates the increasing promise of this methodology in the accurate diagnosis of a patient's mental condition. Therefore, to interpret the observations, and to identify disparities between gender, we investigated their clinical interpretations with respect to depression detection.

### *4.4.1 Audio biomarkers and their relationship with depression detection*

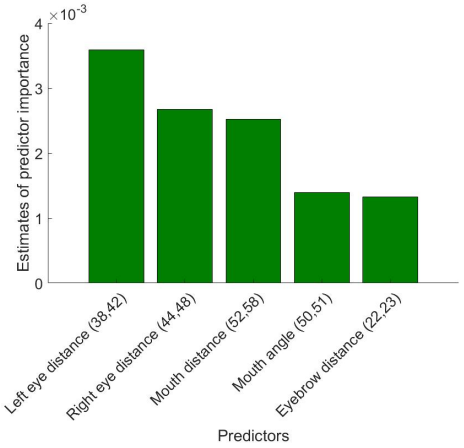
We found that non-linguistic speech patterns provided significant communicative power through variations in pitch, loudness, interruption, anger, and laughter. Hence, they are helpful in providing cues that aid in interpretation of audio signals. For example, one of the most important perceptual elements of sound is pitch, as shown in Figure 4.3.3 a, and b. Studies show that when someone speaks in a slow, soft voice, they may be signaling sadness or depression [135]. Pitch is determined from frequencies that are clear, and can be detected via modulation of amplitude [136, 137], autocorrelation signals [101], MFCC



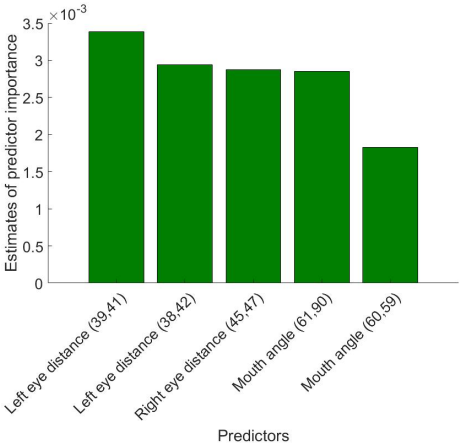

(a) Audio biomarkers for females



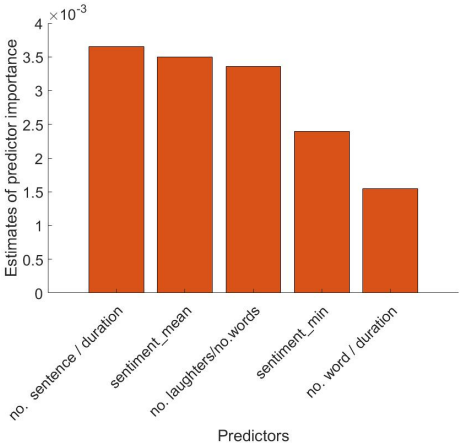

(b) Audio biomarkers for males



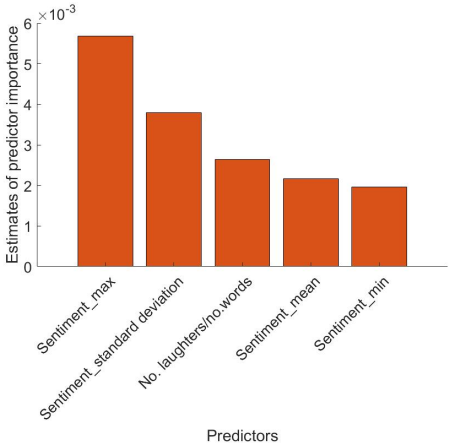

(c) Video biomarkers for females




(d) Video biomarkers for males




(e) Text biomarkers for females




(f) Text biomarkers for males

Figure 4.3.2: Top 5 most important biomarkers that have dominant importances for the prediction in audio, video and text modalities

| **Selected biomarkers** | **Females** | ***p*-value** | **Males** | ***p*-value** |
|---|---|---|---|---|
| Audio biomarkers | MFCC | 0.0012 | Spectral slope | <0.0001 |
| | Modulation of amplitude | 0.0008 | PLP | <0.0001 |
| | Autocorrelation | <0.0001 | Modulation of amplitude | 0.022 |
| | PLP | <0.0001 | Spectral decrease | 0.012 |
| | Spectral decrease | 0.002 | Entropy of energy | 0.011 |
| | Spectral slope | 0.0080 | MFCC | 0.1140 |
| | Entropy of energy | 0.2100 | Autocorrelation | 0.0210 |
| Video biomarkers | Left eye distance (38,42) | <0.0001 | Left eye distance (39,41) | <0.0001 |
| | Right eye distance (44,48) | <0.0001 | Left eye distance (38,42) | 0.0028 |
| | Mouth distance (52,58) | 0.0014 | Right eye distance (45,47) | 0.0012 |
| | Mouth angle (50,51) | 0.001 | Mouth angle (61,90) | <0.0001 |
| | Eyebrow distance (22,23) | <0.0001 | Mouth angle (60,59) | 0.0311 |
| | Left eye distance (39,41) | <0.0001 | Left eye distance (38,42) | <0.0001 |
| | Left eye distance (38,42) | <0.0001 | Right eye distance (44,48) | <0.0001 |
| | Right eye distance (45,47) | 0.0063 | Mouth distance (52,58) | 0.6510 |
| | Mouth angle (61,90) | <0.0001 | Mouth angle (50,51) | 0.0052 |
| | Mouth angle (60,59) | <0.0001 | Eyebrow distance (22,23) | 0.1000 |
| Text biomarkers | No.sentence/duration | <0.0001 | Sentiment_max | 0.003 |
| | Sentiment_mean | <0.0001 | Sentiment_standard deviation | <0.0001 |
| | No.laughters/no.words | 0.032 | No.laughters/no.words | 0.126 |
| | Sentiment_min | 0.887 | Sentiment_mean | 0.366 |
| | No. word/duration | 0.311 | Sentiment_min | 0.369 |
| | Sentiment_standard deviation | 0.0036 | No.sentence/duration 0.322 | |

Table 4.3.3: *p*-value of the selected top 5 significant biomarkers for females and males

[138, 139], and energy [140]. As we could see from the result of audio biomarkers in Figure 4.3.3 a, MFCC, autocorrelation, and modulation of amplitude signals were among the most expressive biomarkers for females. The significance of these biomarkers is shown in Table 4.3.3. This might be due to the fact that the pitch of the male voice falls under low frequencies whereas the female voice is higher pitched.

Another important indication of depression is anger, which is characterized by an increase in pitch. MK Biaggio et.al, studied the effect of anger over 112 university students on depression. They found that, among depressed subjects, there was a more intense experience of hostility and less control over anger [141]. Angry voices typically have large proportion of high-frequency energy in the spectrum, and hence can be detected via frequency-based biomarkers like MFCC [142, 143], and PLP [144]. Interestingly, the results showed that these biomarkers appear for both females and males, as shown in Figure 4.3.3 a, and b.

Another voice characteristic the scientific community agrees aids in detecting depression is laughter. The influence of depression on the reduction of laughter frequency in speech has been shown in [145]. It has been found that laughter is able to improve negative consequences of stressful events and depressive symptoms. This characteristic of sound can be found through changes in spectral/cepstral biomarkers [146]. As shown in Table 4.3.3, the spectral/cepstral biomarkers were significant for both females and males. However, we found cross-relations across speech patterns. Any changes in the intonation or the rhythm of speech could reflect the spectral/cepstral biomarkers and could be considered as either laughter or anger. This shows, despite the great value of sound to determine an emotional state of the speaker, further investigation is required for drawing any conclusion. The complete list of vocal characteristics and their relationship with audio biomarkers is shown in Table 4.4.1.

### *4.4.2 Video biomarkers and their relationship with depression detection*

We found that biomarkers related to eye, eyebrow, and mouth movements were predictive to depression detection, as shown in Figure 4.3.3 c, and d. We also investigated the influence of gender-dependent classification on the selected video biomarkers. For example, we found that eye movement holds a discriminative power for detecting depression. The results show that the vertical L2 distance of the left eye [(38,42) and (39,41)], and the vertical L2 distance of the right eye [(44,48) and (45,47)] were significant biomarkers among depressed subjects for both female and male participants, as shown in Table 4.3.3. This might be an indication

| **Audio biomarkers** | **↓ Loudness** | **↓ Pitch** | **↑ Silence** | **↑ Interruption** | **↑ Pauses** | **↑ Anger** | **↓ Laughter** |
|---|---|---|---|---|---|---|---|
| Modulation of amplitude | ✓[147, 148] | ✓[136, 137] | | | | | |
| Envelope | | ✓[149] | ✓[150] | | | | |
| Autocorrelation | ✓[151] | ✓[101] | | | | | |
| Onset detector | | ✓[152] | ✓[153] | | | | |
| Entropy of energy | ✓[154] | ✓[140] | | | | | |
| Zero crossing | | ✓[155] | ✓[150] | ✓[156] | | | |
| PLP | ✓[157] | | | | ✓[158] | ✓[144] | ✓[146] |
| MFCC | | ✓[138, 139] | ✓[159] | | [160, 161] | ✓[142, 143] | ✓[146] |
| Spectral decrease | ✓[162] | ✓[149] | ✓[163] | | | | ✓[146] |
| Spectral rolloff | ✓[162] | | | | | | ✓[146] |
| Spectral flux | ✓[162] | | | | | | ✓[146] |
| Spectral centroid | ✓[162] | | | | | | ✓[146] |
| Spectral slope | ✓[162] | | | | | | ✓[146] |
| Spectral autocorrelation | ✓[162] | | | | | | ✓[146] |

Table 4.4.1: Audio biomarkers relationship with non-linguistic speech patterns. ↑ and ↓ show that, higher and lower level of these biomarkers, indicates higher risk of depression respectively.

of fatigue, which is a common symptom of depression.

Another biomarker is the mouth movement, which is accompanied by a deep breath. The mouth movement sometimes happens when the lips are closed, or pushed forward, and twisted to one side. Usually, rapid mouth movement is an indication of self-consciousness in social situations. However, from Table 4.3.3, we could see this was not a significant biomarker for male participants. Further, we observed that the angle between the points in the mouth are significant (both females and males). This is because non-speech mouth movements have shown indications of depression symptoms [164].

Another significant biomarker was the rapid eyebrow raising which was observed in females only. Studies show that when the frequency of an audio is pronounced it is usually associated with the eyebrow movement too. This is due to the fact that audio and facial signals are integrated. The fact that speech and facial biomarkers interact suggests that both are controlled by the same control system, and need further investigation.

### *4.4.3 Text biomarkers and their relationship with depression detection*

We found that the ratio of depression-related words over the duration was a significant text biomarker for detecting depression in females, as shown in Figure 4.3.3. In contrast, this biomarker did not show any significance for men, as shown in Table 4.3.3. Interestingly, the ratio of the number of instances of laughter over the number of words was a significant biomarker for females, and not for males. Research has confirmed that emotional responses in women to report feeling are stronger. This is because women likely tend to exhibit positive emotions, including the expression of joy or laughter [121].

However, the remaining results on gender differences show that there is no gender dependence in detecting depression using text biomarkers. One reason is that the number of depression-related words detected were identified from a common dictionary for both groups. A key factor known to differ between males and females is the use of affiliative language versus assertive language. Affiliative language positively engages the other person by expressing

agreement, versus the assertive language that includes directive statement. Women tend to use affiliative language more, and men tend to use assertive language more [165]. Perhaps further research is needed in a more discriminative setting in order to properly construct the dictionary and shed light on this issue.

### *4.4.4 Limitations and next steps*

One of the challenges that became apparent, was the availability and limitations of the baseline dataset. For example, the video dataset had some restrictions, due to its unavailability and restricted format. The results showed that fusing the video with other modalities was not very helpful. The availability of raw video recordings could help in extracting more relevant biomarkers. Hence, it could improve the predictability of the multi-modality mechanism. Another limitation of the dataset was the unavailability of some measures that could improve the detectability of depression. For example, studies showed that physiological signals such as heart rate are strongly correlated with depression detection [166, 167], and are complementary to audiovisual data. Thus, access to these measures could help the detection task. On the other hand, Wizard of Oz interviews were transcribed from a composite video, combining both participant and interviewer. Later automated interviews were transcribed from the audio stream of the participant only without including the interviewer questions. However, full comprehension of the text is not possible. For example, several transcribed files included unrecognizable words marked by "x", suggesting that prediction also could be challenged with the existing dataset.

Another limitation of this study is the virtual interviewer, Ellie [4]. She functions through a computer program that uses different algorithms to determine her questions and motions and analyzes the emotional cues. The computer program reads a patient's current facial expression and the variation in expressions throughout the session and taking any flat expression symptomatic of depression. Despite her ability to collect signs of mental health problems, she cannot deliver solutions and doesn't have the humanity to aid in therapy work. For example,

[4] `http://viterbi.usc.edu/news/news/2013/a-virtual-therapist.htm`

unlike a human therapist, Ellie cannot respond to questions. Moreover, some people find it easier to open up to a human therapist when speaking face-to-face. Furthermore, Ellie's response is limited to changes in facial expression and recordings of the conversation, however, human behavior can be difficult to understand, and may rely on changes in psychological parameters (e.g., blood pressure, weight, cardiac conditions etc.).

We also faced the challenge presented by the limitations of the current model. The most essential endeavor to be taken in using the multi-modality fusion model is, to learn about relationships between modalities. Thus, prior knowledge of modality-specific information and interactions among them, and other mutual properties could be very useful. The current multi-modality fusion model ignored the interactions among modalities, such as audio-visual relations. Some studies have shown that combining audio and visual biomarkers can improve emotion recognition [3], which is a closely related topic to depression evaluation [95]. For example, Ekman investigated the systematic relations of speech and eyebrow movement [168]. He found that when the frequency of a sound is pronounced, it is usually associated with the eyebrow movement. This is due to the fact that audio and facial signals are integrated. Further, we found that there are cross-relations in speech patterns too. For example, we found that both anger and laughter appear in audio signals, when frequencies are amplified. Further investigation is needed to distinguish these two signals. In addition, we recognized the importance of biomarkers in a multi-modality noisy data. With the addition of more relevant biomarkers, overall performance could be largely improved. For example, the silence detection could be used to check how often the participants keep quiet, which may act as a depression indicator.

One possible way to improve the multi-modality fusion model is to build a personalized model by constructing an individual model and then aggregating them to generate prediction outcomes [3]. Towards this goal, the model could be extended to a more general social signal processing mechanism. This could help us in characterizing social interactions by assessing audio-visual recordings in real-world conditions. Another solution is to transcribe the audio

dataset. Hence, learning complex emotions that interact with one another, such as anger and laughter, may help optimize recognition accuracy.

# Chapter 5

# UNCERTAINTY QUANTIFICATION FOR DEEP CONTEXT-AWARE MOBILE ACTIVITY RECOGNITION AND UNKNOWN CONTEXT DISCOVERY

Activity recognition in wearable computing faces two key challenges: i) the incorporation of the dependency of activities on user contexts is very important; ii) unknown contexts and activities may occur from time to time, requiring flexibility and adaptability of the algorithm. We develop a context-aware mixture of deep models termed the $\alpha$-$\beta$ network to enhance human activity recognition performance (accuracy and F score) by 10% through identifying high-level contexts in a data-driven way to guide model development. Furthermore, we equip the $\alpha$-$\beta$ network with the capability of uncertainty quantification based on the maximum entropy learning, for unknown context discovery, and demonstrate improvements on public benchmarks.

## *5.1 Introduction*

Context-aware systems are used to design innovative user interfaces and are often used as a part of wearable computing to improve activity recognition. Context can be defined as "any information that can be used to characterize the situation" [169] or to improve recognition [170]. Context-aware systems have previously been used in many applications, including activity recognition [171], online, personalized and adaptive activity classification [172], and healthcare applications [173, 174]. The definition of contexts heavily relies on domain knowledge, such as a user's tasks (e.g., spontaneous activity, engaged tasks) or a user's social environment (e.g., co-location of others, group dynamics), etc. However, in practice, pre-defined contexts may not always be available, or definitions of contexts

may change in different environments. Additionally, new unknown contexts may emerge over time. For these reasons, there is a general data insufficiency and lack of contextual information to develop accurate context-aware activity recognition systems that could adapt to these unknown contexts. To overcome these challenges, we show that better activity recognition performance can be achieved if we enhance activity recognition with context awareness capability. Our method can adapt to data due to its data-driven nature, and can effectively discover unknown contexts with our UQ method. This work allows for models trained in laboratory settings to extend to natural environments for monitoring behaviors and performances of users.

While existing works are not adequate to address the challenges such as a lack of context information in data while training the model, or the emergence of unknown contexts when the model is applied [175, 176], our work employs a more powerful data-driven approach by learning unknown contexts and the distribution of each user's specific activity likelihood within each context. We develop the $\alpha$-$\beta$ framework, where the $\alpha$ network is the context detector to learn a distribution over contexts as a mixture of weights, and the $\beta$ network models activity recognition for context-specific data. For example, given the sensor reading data from a user, the $\alpha$ network detects the context by generating a distribution over different contexts; then, each context has a dedicated $\beta$ network that outputs a distribution over different activities. Further, to make our model robust with new unknown contexts, we equip the $\alpha$-$\beta$ network with UQ based on the maximum entropy learning (MEL) principal to detect unknown context. MEL identifies the distribution of the parameters of a statistical model that bears the maximum uncertainty, rather than one single best model, as a principle to achieve robustness in prediction and modeling. The prediction model could refuse to predict on given data if the uncertainty for making a prediction on this data is higher than a threshold.

## *5.2 Methods*

Contextual information can help group and model similar activities together, resulting in an improvement of activity recognition performance. Our proposed method eliminates the need for contextual data collection by unsupervised context detection through our proposed $\alpha$-$\beta$ network. A further complexity is that contextual data can be different from person to person, and it can change over time for the same person. Therefore, these systems should be able to identify unknown contexts. In section 5.2.1, we will introduce our $\alpha$-$\beta$ network which enables unsupervised context detection, and in section 5.2.2, we will discuss how UQ improves the $\alpha$-$\beta$ network to discover unknown contexts from data.

### *5.2.1 Context-awareness processing*

Studies show that context-awareness plays an important role in improving the performance of activity recognition systems [177, 178]. Activity data that are captured by different sensors normally have a heterogeneous set of input features, which make feature extraction challenging. This feature extraction can be automated using deep learning, simplifying the task of defining contexts for the data. In human activity recognition, the input data is multichannel and multimodal, which can pose serious challenges to conventional activity recognition frameworks, especially in the feature extraction phase. Convolutional neural networks (CNNs) can automate the feature extraction process to extract task-specific features with state-of-the-art results in clustering and classification [179, 180]. In this work we develop a mixture of CNNs, the $\alpha$-$\beta$ network, where each mixture component is dedicated to one specific context. There are two types of networks: $\alpha$ and $\beta$. Given the sensor data, the $\alpha$ network detects context by generating a probability distribution over all known contexts. Each context has a dedicated $\beta$ network that outputs a probability distribution over different activities. Our activity recognition problem features a latent context variable and can be formulated as:

$$\begin{aligned}\log p(ACTIVITY|\mathbf{X},\theta) &= \sum_{i=1}^{N} \log p(activity_i|\mathbf{x}_i,\theta) \\ &= \sum_{i=1}^{N} \log \sum_{c=1}^{N_c} p(activity_i|c_i=c,\mathbf{x}_i,\theta)p(c_i=c|\mathbf{x}_i,\theta),\end{aligned} \tag{5.1}$$

where $\theta$ denotes the mixture component parameters, $N$ denotes the number of data samples, and $N_c$ denotes the number of expected clusters (contexts) to which each data point may belong. Our objective is to maximize Eq. (5.1) with respect to $\theta$. As shown by [181], the log-likelihood has a lower bound:

$$\begin{aligned}&\log p(ACTIVITY|\mathbf{X},\theta) \geq \\ &\sum_{n=1}^{N}\sum_{c=1}^{N_c} q(c_i=c) \log \frac{p(activity_i|c_i=c,\mathbf{x}_i,\theta).p(c_i=c)}{q(c_i=c)},\end{aligned}$$

where $q(\cdot)$ is the distribution over different contexts and $p(\cdot|context,x,\theta)$ is the distribution over activities given the context and the input. $q(\cdot)$ is modeled using the context-detecting $\alpha$ network, and each $p(\cdot|context,x,\theta)$ is modeled using a $\beta$ network (each context has its own $\beta$ network). Following EM algorithm, the lower bound in Eq. (5.2) can be maximized. Specifically, the loss, the negative of the lower bound, is minimized by EM. In the E-step, $q(\cdot)$ is optimized which translates to optimizing $\alpha$ network while freezing $\beta$ networks. In the M-step, $\theta$ (model parameters) need to be optimized which translates to optimizing $\beta$ networks while freezing $\alpha$ network. The EM training alternates iterations of training either $\alpha$ network or $\beta$ networks while keeping the other(s) frozen. It should be noted that no labeled contextual data is used in the training process for the $\alpha$-$\beta$ network.

### *5.2.2 Unknown context discovery*

The $\alpha$ network enables context detection; however, in practice, contexts may change over time, or may not always be pre-defined. It is possible to improve context-aware systems by detecting the uncertainty of possible unknown contexts as a result of potential distribution mismatch between known and unknown contexts. To identify unknown contexts, we combine the feature extraction power of deep learning with the learning power of MEL to define a

probabilistic mechanism for unknown context discovery.

UQ has been critical for robust learning under different contexts (known or unknown) in mobile activity recognition [182], healthcare [4, 60], signal processing [63], and manufacturing [62]. Existing models with Gaussian process [182] or Bayesian approximation methods [183] either rely on the assumption that variables in a system can be characterized by explicit probabilistic relationships (e.g., Bayesian models) or rely on generating one best model in the learning algorithm.

In the literature, measuring the predictive uncertainty for deep neural networks is a challenging problem. Studies show that this could be done by modeling predictive uncertainty by parameterizing a prior distribution over predictive distributions [184], by utilizing probabilities from softmax distributions and detecting out-of-distribution examples [185], or by obtaining the class conditional Gaussian distributions by introducing confidence scores [186]. For example, Malinin & Gales proposed a framework called Prior Networks (PNs) for modeling predictive uncertainty which explicitly modeled distributional uncertainty by parameterizing a prior distribution over predictive distributions [184]. Unlike PNs that attempted to quantify the uncertainty through modeling predictive uncertainty, Hendrycks & Gimpel proposed a framework that utilized probabilities from softmax distributions and detected out-of-distribution examples, by introducing confidence scores that were proposed based on density estimators [187]. This model was further improved by processing the input and output of DNNs in [185]. On the other hand, Lee et al. proposed a method for detecting abnormal test samples by including both out-of-distribution and adversarial samples, to obtain the class conditional Gaussian distributions by introducing confidence scores based on the Mahalanobis distance [186].

Our method is different from these works as we aim to learn a distribution of deep models rather than one single best one. While many of these works focus on revising general deep models with a probabilistic evaluation of their model or prediction, here we have a different aim: to make the proposed $\alpha$-$\beta$ network adaptive to changing contexts hidden in data. To

achieve this goal, our approach is to relax our expectation of identifying one single optimal model of the $\alpha$-$\beta$ network; rather, we consider solving for a full distribution over multiple models. The intuition is that many different models might generate relatively similar performance, so it would be better to estimate a distribution over parameters $p(\mathbf{w})$, from the output layer of the $\alpha$ networks that detects context. This intuition aligns with the basic principal of MEL [188, 189]. Therefore, we equip the $\alpha$-$\beta$ network with UQ capacity based on the MEL principal. By identifying the distribution of the parameters of a statistical model that bears the maximum uncertainty through MEL, our model is able to detect any occurrence of an unknown context.

**Uncertainty quantification via minimizing relative entropy** To learn the distribution of the $\alpha$-$\beta$ network parameters that encode maximum uncertainty, we employ the MEL formulation. There are two steps. First, we create constraints that encode information from the data. For example, for each sample we derive a constraint such that the expected prediction on this sample over all the possible model parameters matches the observed outcome on this sample. Second, on the top of this constraint structure, the learning objective of MEL is to learn the distribution of the model parameters with the maximal entropy. Thus, unlike traditional machine learning methods that estimate a single optimal setting of the parameter, MEL considers a more general problem of these methods by solving for a full distribution over multiple $p(\mathbf{w})$ values (see Appendix C.1).

The distribution of parameters forms a quantitative evaluation of model uncertainty, which could be further used in subsequent decision making by using probability laws to track the uncertainty propagation process. In this paper, we create a rejection option, a flexibility enabled by the UQ capacity. The rejection option allows for the prediction model to refuse to generate a prediction if the uncertainty is higher than a given threshold. This is typically solved by estimating the class conditional probabilities and rejecting the samples that have lower posterior probability of class. As we could obtain the distribution of parameters by

solving MEL, we can therefore derive the formula to calculate prediction uncertainty in each given sample by following the uncertainty propagation process from model to samples.

## *5.3 Experiments*

### *5.3.1 Dataset*

We used the UCI OPPORTUNITY dataset [190] for context-aware human activity recognition. The dataset contained 18 different activities performed in five different contexts and sensed by 72 different sensors. Each of the 18 activities had one of the five contexts, but not all contexts contained every activity. Therefore, the UCI OPPORTUNITY dataset provided a realistic capture of a situation where not all of the human activities occurred with an equal likelihood in all contexts. The UCI OPPORTUNITY dataset has seven levels of hierarchical labels. Higher level labels described details such as subject posture, while lower level labels described the hand movements or interactions with other subjects. In this study, we chose a higher level label (e.g., cleaning time) as the context and a lower level (e.g., opening a door) label as the activity.

### *5.3.2 Implementation*

In our work, 19 different preprocessed sensors were fed into the network. Time series data were divided into non-overlapping segments of 1 second (30 samples). In the OPPORTUNITY dataset, we used all the body-worn sensors which included seven inertial measurement units (IMUs) and twelve 3D acceleration sensors. Five IMUs were on the upper body while two were on user's shoes. Accelerometers were on the upper body, hip, and leg, which translated to 133 columns in the raw dataset.

We used a five-fold cross-validation to evaluate our models with testing accuracy and micro F score as performance metrics. In our experiments, we used 3 convolutional layers followed by 3 fully connected layers for both $\alpha$ and $\beta$ networks. Note that these two networks' architectures were different in terms of the number of neurons in the output layer. Networks were trained using stochastic gradient descent with an initial learning rate of 0.001 and a


(a) Without pre-training

(b) With pre-training

Figure 5.3.1: Average of $\alpha$ network output in testing without (a) and with (b) pre-training. The model collapses into a single network in the former. This shows the effectiveness of pre-training in making sure that the $\alpha$ network finds subgroups in the data and hence can take advantage of context-aware recognition.

momentum of 0.9, which provided the best results in cross-validation.

**Pre-training** Initialization is an essential step for both the optimization and for the training of the neural networks. Without proper initialization, the model can collapse into selecting only one specific $\beta$ network while eliminating the contribution of others. Pre-training is an important stage in the $\alpha$-$\beta$ network training. The model could approach the base model of a single neural network classifier without proper pre-training because of the large gradients at the beginning of the training stage; large gradients cause the selector to saturate and select only one model. This phenomena can be seen in Figure 5.3.1 where no pre-training results in selection of only one network. The effect of pre-training can also be seen in the network performance. Without proper initialization, the $\alpha$-$\beta$ network with four contexts achieves an accuracy of 0.87 and an F-score of 0.87. This is 4% less than the performance of an identical $\alpha$-$\beta$ network with proper initialization. This is roughly equal to the performance achieved by a single $\beta$ network (refer to Table 5.3.1). Therefore, the full capabilities of the $\alpha$-$\beta$ network can only be achieved using proper pre-training, which proves useful in finding subgroups of data as well as in yielding a better performance.

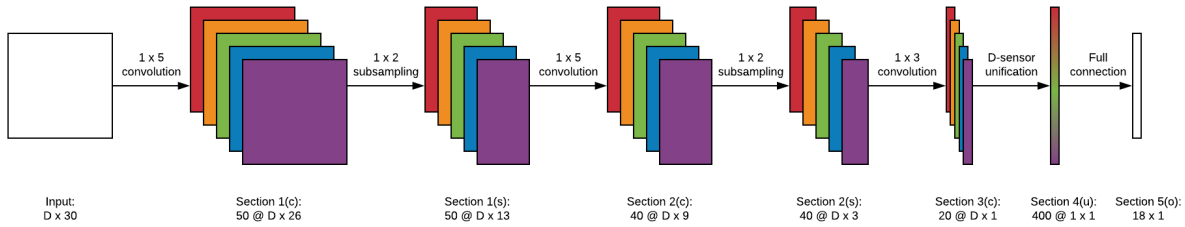



Figure 5.3.2: CNN based deep activity recognition neural network. This network, designed for the OPPORTUNITY dataset, is used as the baseline for our model.

Pre-training the $\alpha$ network required a relatively stable clustering of data. Because of the unavailability of context information, we first clustered the activity data and then trained the network. We used the idea presented in [180] to cluster the activity data. The details of the CNN used in this work can be seen in Figure 5.3.2. In detail, a base network was trained with sensor readings as input and activities as output. Next, the CNN segment of the network was used to embed the sensor readings into the features which have proven to be descriptive of the input data [179, 191]. Subsequently, we used K-means to cluster the input into a fixed number of clusters. Finally, we trained our $\alpha$ network to learn the mapping between sensor readings of activity to clusters. This phenomena can be seen in Figure 5.3.1. In this case, the model approaches the base model of a single neural network classifier. The collapse happens because of large gradients at the beginning of the training stage; large gradients cause the selector to saturate and select only one model.

### *5.3.3 Results: $\alpha$-$\beta$ Network vs. Baseline*

Table 5.3.1 compares the testing accuracy of the $\alpha$-$\beta$ network and the baseline (a single network that is equivalent to a single $\beta$ network) for predicting labels in the UCI OPPORTUNITY dataset. The testing accuracies are averaged among all their corresponding bootstraps (bootstrapped 5 times). As can be seen in Table 5.3.1, the mixture of the context-specific neural networks improved on the accuracy of the baseline from 86% to 96% when using nine

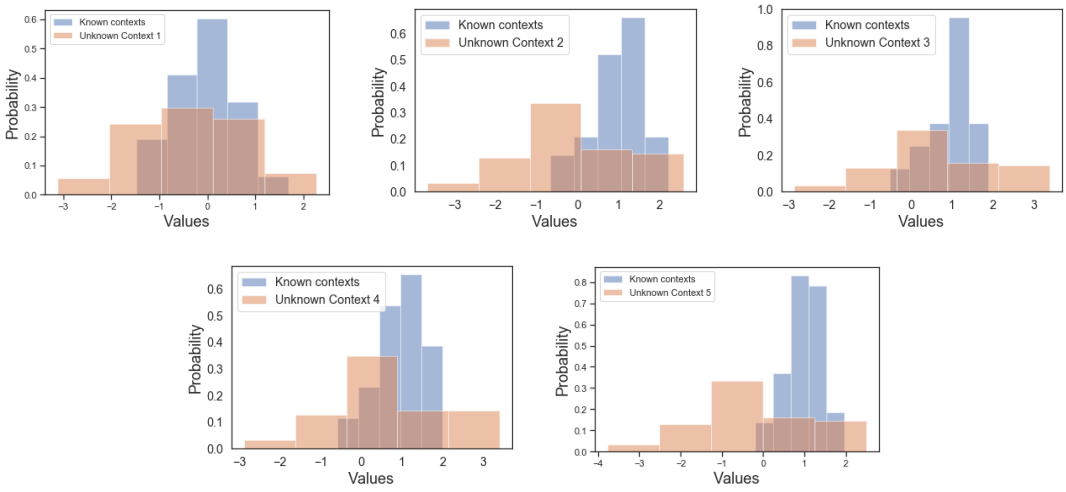



Figure 5.3.3: Predicted probability distributions when a context is removed v.s the aggregate of all known contexts within each rotation.

Table 5.3.1: Accuracy and F score for the $\alpha$-$\beta$ network and baseline. 1 cluster is a single $\beta$, and additional models are $\alpha$ and $c * \beta$ where $c$ is the number of clusters.

| Model clusters | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Accuracy | 0.86 | 0.89 | 0.89 | 0.91 | 0.92 |
| F-score | 0.86 | 0.90 | 0.90 | 0.91 | 0.91 |
| | | | | | |
| | 6 | 7 | 8 | 9 | 10 |
| Accuracy | 0.91 | 0.96 | 0.96 | 0.96 | 0.95 |
| F-score | 0.92 | 0.96 | 0.96 | 0.96 | 0.95 |

contexts. Thus, our context-aware $\alpha$-$\beta$ network was able to find subgroups in the data in an unsupervised manner with different numbers of contexts as they all yield higher performance when compared to the baseline. This fact was reflected in the accuracy boost due to the subgroup modeling by different $\beta$ networks.

Table 5.3.2: UQ results for unknown context discovery where contexts are 1 = Relaxing, 2 = Coffee time, 3 = Early morning, 4 = Clean up, 5 = Sandwich time

| Performance measure | Context number that was removed | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Sensitivity | 0.63 | 0.71 | 0.75 | 0.62 | 0.73 |
| Specificity | 0.72 | 0.72 | 0.76 | 0.79 | 0.82 |
| Testing accuracy | 0.67 | 0.71 | 0.75 | 0.69 | 0.77 |
| F-score | 0.67 | 0.72 | 0.77 | 0.70 | 0.78 |

In the UCI OPPORTUNITY dataset, we had five contexts: relaxing, coffee time, early morning, clean-up, and meal time. To test our unknown context discovery, we adopted a *rotating* strategy. In this strategy, we removed one label and its corresponding data from the training dataset at each rotation, and trained an $\alpha$-$\beta$ network only on the remaining known contexts rather than using the full context information. In detail, after training the $\alpha$-$\beta$ network on a rotating base, one context was assumed to be unknown and was treated as a hold-out to be used for unknown context discovery assessment. Our UQ approach compared the uncertainty with a threshold to see whether a given sample should be detected as belonging to an unknown context or not. We selected the threshold based on a cross-validation set, and then we presented the resulting F-score, accuracy, sensitivity, and specificity at the optimal threshold. We defined a classification with rejection, where if a sample was rejected (if the uncertainty was higher than a specific threshold), the prediction model refused to generate a prediction by setting the predicted context to zero.

Figure 5.3.3 presents a main result in this experiment from the UCI OPPORTUNITY dataset. Here, the distribution of predicted probabilities of the removed context (the red histogram) is shown with the distribution for all the other contexts combined (the blue histogram). It can be seen that our UQ algorithm is able to detect the uncertainty for the unknown context, as the unknown context usually leads to smaller probabilities in compari-

son with the distribution of the known contexts. UQ results of unknown context discovery for the UCI OPPORTUNITY dataset when each context is removed are reported in Table 5.3.2. This shows that the UQ-framework reaches average sensitivity of 0.69 on correctly identifying an unknown context and an average specificity of 0.76 while correctly identifying a known context with an average accuracy of 0.72.

## *5.4 Discussion*

In this paper, we developed a novel $\alpha$-$\beta$ network together with its UQ formulation in tackling a range of realistic situations where context information was unknown but critical for enhanced situation awareness and human activity recognition. Experiments on a real-world dataset showed that this combination of deep learning and uncertainty quantification led to superior performances by its efficacy and efficiency in extracting context information, recognizing unknown contexts, and using UQ in prediction.

This work can be used as a foundation for a more comprehensive analysis of contextual discovery in a variety of modeling efforts. Multiple levels of contextual information can be gathered and learned from the system, and understanding those in a truly unsupervised setting can enhance a number of recognition tasks and create a flexible and more realistic ontology for how to define context in human activity recognition tasks, rather than relying on high-level but general descriptions of context or too restrictive pre-defined contexts. In addition, other sources of uncertainty in prediction which result from data of new distributions or noisy data from old distributions should be considered.

# Chapter 6

# CONCLUSION AND FUTURE RESEARCH

This dissertation proposes novel statistical, machine learning, and optimization models that seek to contribute to the design of efficient disease progression monitoring.

## *6.1 For disease trajectory modeling*

This research informs the development of a personalized contemporary health index (CHI) and quantifies model predictive uncertainty when monitoring patient condition over time. Our frameworks specifically leverage the monotonic progression patterns of the target degenerative disease conditions such as the AD and SSI and quantify these patterns with systematic optimization formulations. Strong numerical performances on two real-world applications suggest the promising capability of the proposed models.

## *6.2 For biomarker engineering*

We proposed effective biomarker engineering pipelines to enable possible extensions to the CHI for building trajectory models from complex data (video, audio, text, and mobile sensor reading data). First, a multi-modality fusion model was applied to the audio, video, and text modalities to identify the biomarkers that are predictive of depression after accounting for gender differences. Second, we focused on context-aware systems as a part of wearable computing to improve activity recognition and to detect useful digital biomarkers that are predictive or indicative of the patient's conditions.

## *6.3 Future research: CHI extension to other healthcare applications*

Our work points to a number of promising research directions. This includes extensions of the proposed methods to many other diseases that may have degradation characteristics

that differ from degenerative diseases. This may ultimately lead to the development of control system engineering that can implement adaptive interventions for better healthcare management that accounts for dynamic conditions over time. For example, the alarming rise in opioid misuse has resulted in a public health crisis in the U.S. characterized by dramatic increases in drug overdose deaths. Despite recent efforts to tackle the issue, the rates of opioid misuse and non-fatal and fatal overdose remain staggeringly high. Opioid use disorder (OUD) develops due to a complex range of factors, and currently, the understanding of these interrelationships is limited. Hence, a global health index could be designed to uncover how OUD is developed and to characterize the trajectory throughout progression.

### *6.4 Future research: disease monitoring with multi-modal sources of data*

Advanced analytical techniques can be used to represent the progression of degenerative disease, however, they usually ignore the multi-factorial aspects of the disease. For example, building trajectory models solely based on measures from clinical data often only capture one part of the SSI progression, whereas there are many other measures (text, images, etc.) that can be used. We recommend that future studies better accommodate the need for multi-modality prediction models combining different sources of data using domain knowledge of the disease.

# BIBLIOGRAPHY

[1] Yijun Huang, Qiang Meng, Heather Evans, William Lober, Yu Cheng, Xiaoning Qian, Ji Liu, and Shuai Huang. Chi: A contemporaneous health index for degenerative disease monitoring using longitudinal measurements. *Journal of biomedical informatics*, 73:115–124, 2017.

[2] S Swaroop Vedula and Gregory D Hager. Surgical data science: the new knowledge domain. *Innov Surg Sci*, 2(3):109–121, 2017.

[3] Aven Samareh and Shuai Huang. Dl-chi: a dictionary learning-based contemporaneous health index for degenerative disease monitoring. *EURASIP Journal on Advances in Signal Processing*, 2018(1):17, 2018.

[4] Aven Samareh and Shuai Huang. Uq-chi: An uncertainty quantification-based contemporaneous health index for degenerative disease monitoring. *arXiv preprint arXiv:1902.08246*, 2019.

[5] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[6] John K Darby, Nina Simmons, and Philip A Berger. Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders*, 17(2):75–85, 1984.

[7] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[8] Aven Samareh, Yan Jin, Zhangyang Wang, Xiangyu Chang, and Shuai Huang. Detect depression from communication: how computer vision, signal processing, and sentiment analysis join forces. *IISE Transactions on Healthcare Systems Engineering*, 8(3):196–208, 2018.

[9] Aven Samareh, Yan Jin, Zhangyang Wang, Xiangyu Chang, and Shuai Huang. Predicting depression severity by multi-modal feature engineering and fusion. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] Jong-Tae Park, Jae-Wook Nah, Su-Wook Kim, Sung-Man Chun, Song Wang, and Su-Ho Seo. Context-aware handover with power efficiency for u-healthcare service in wlan.

In *2009 International Conference on New Trends in Information and Service Science*, pages 1279–1283. IEEE, 2009.

[11] JinSeok Ko, Manar Mohaisen, JaeYeol Rheem, and Jeong-Won Kim. A context awareness model for u-healthcare based on artificial neural network.

[12] A. Beck and L. Tetruashvili. Safer care for the acutely ill patient: learning from serious incidents. *National Patient Safety Agency*, 2013.

[13] Robert P Gaynes, David H Culver, Teresa C Horan, Jonathan R Edwards, Chesley Richards, James S Tolson, and National Nosocomial Infections Surveillance System. Surgical site infection (ssi) rates in the united states, 1992–1998: the national nosocomial infections surveillance system basic ssi risk index. *Clinical Infectious Diseases*, 33(Supplement_2):S69–S77, 2001.

[14] Bonnie Spring, Marientina Gotsis, Ana Paiva, and Donna Spruijt-Metz. Healthy apps: mobile devices for continuous monitoring and intervention. *IEEE pulse*, 4(6):34–40, 2013.

[15] Daniel E Rivera. Optimized behavioral interventions: What does system identification and control engineering have to offer? *IFAC Proceedings Volumes*, 45(16):882–893, 2012.

[16] Sunil Deshpande, Daniel E Rivera, Jarred W Younger, and Naresh N Nandola. A control systems engineering approach for adaptive behavioral interventions: illustration with a fibromyalgia intervention. *Translational behavioral medicine*, 4(3):275–289, 2014.

[17] Gloria Zen, Lorenzo Porzi, Enver Sangineto, Elisa Ricci, and Nicu Sebe. Learning personalized models for facial expression analysis and gesture recognition. *IEEE Transactions on Multimedia*, 18(4):775–788, 2016.

[18] Jeffrey L Cummings. Cognitive and behavioral heterogeneity in alzheimer's disease: seeking the neurobiological basis. *Neurobiology of aging*, 21(6):845–861, 2000.

[19] Marshal F Folstein. Heterogeneity in alzheimer's disease. *Neurobiology of aging*, 10(5):434–435, 1989.

[20] Robert P Friedland, Elisabeth Koss, James V Haxby, Cheryl L Grady, Jay Luxenberg, Mark B Schapiro, and Jeffrey Kaye. Alzheimer disease: clinical and biological heterogeneity. *Annals of internal medicine*, 109(4):298–311, 1988.

[21] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996.

[22] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust

face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

[23] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

[24] Meng Yang, Lei Zhang, Jian Yang, and David Zhang. Metaface learning for sparse representation based face recognition. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 1601–1604. IEEE, 2010.

[25] Qiong Xu, Hengyong Yu, Xuanqin Mou, Lei Zhang, Jiang Hsieh, and Ge Wang. Low-dose x-ray ct reconstruction via dictionary learning. *IEEE Transactions on Medical Imaging*, 31(9):1682–1697, 2012.

[26] Yang Chen, Xindao Yin, Luyao Shi, Huazhong Shu, Limin Luo, Jean-Louis Coatrieux, and Christine Toumoulin. Improving abdomen tumor low-dose ct images using a fast dictionary learning based processing. *Physics in medicine and biology*, 58(16):5803, 2013.

[27] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3501–3508. IEEE, 2010.

[28] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM, 2007.

[29] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005.

[30] Jeffrey R Petrella, R Edward Coleman, and P Murali Doraiswamy. Neuroimaging and early diagnosis of alzheimer disease: a look to the future. *Radiology*, 226(2):315–336, 2003.

[31] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103. ACM, 2012.

[32] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Modeling disease progression via multi-task learning. *NeuroImage*, 78:233–248, 2013.

[33] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE Transactions on image processing*, 17(1):53–69, 2008.

[34] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, 2013.

[35] Michael Elad, Mario AT Figueiredo, and Yi Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6):972–982, 2010.

[36] Bin Cheng, Jianchao Yang, Shuicheng Yan, Yun Fu, and Thomas S Huang. Learning with l1-graph for image analysis. *IEEE transactions on image processing*, 19(4):858–866, 2010.

[37] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang, and Shuicheng Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[38] J. A. Bagnell and David M. Bradley. Differentiable sparse coding. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 113–120. 2009.

[39] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R. Bach. Supervised dictionary learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1033–1040. 2009.

[40] Soheil Bahrampour, Nasser M Nasrabadi, Asok Ray, and William Kenneth Jenkins. Multimodal task-driven dictionary learning for image classification. *IEEE Transactions on Image Processing*, 25(1):24–38, 2016.

[41] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.

[42] Kjersti Engan, Sven Ole Aase, and John Håkon Husøy. Multi-frame compression: Theory and design. *Signal Processing*, 80(10):2121–2140, 2000.

[43] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

[44] Kjersti Engan, Karl Skretting, and John Håkon Husøy. Family of iterative ls-based dictionary learning algorithms, ils-dla, for sparse signal representation. *Digital Signal Processing*, 17(1):32–49, 2007.

[45] Julien Mairal, Guillermo Sapiro, and Michael Elad. Learning multiscale sparse rep-

resentations for image and video restoration. *Multiscale Modeling & Simulation*, 7(1):214–241, 2008.

[46] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

[47] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[48] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[49] Zhuolin Jiang, Zhe Lin, and Larry S Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1697–1704. IEEE, 2011.

[50] Michal Aharon, Michael Elad, and Alfred Bruckstein. *rmk*-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[51] Joseph T Dipiro, Robert G Martindale, Alan Bakst, Paul F Vacani, Phillip Watson, and Martin T Miller. Infection in surgical patients: effects on mortality, hospitalization, and postdischarge care. *American journal of health-system pharmacy*, 55(8):777–781, 1998.

[52] Elise H Lawson, Bruce Lee Hall, and Clifford Y Ko. Risk factors for superficial vs deep/organ-space surgical site infections: implications for quality improvement initiatives. *JAMA surgery*, 148(9):849–858, 2013.

[53] Lauren Saunders, Marion Perennec-Olivier, Pascal Jarno, François L'Hériteau, Anne-Gaëlle Venier, Loïc Simon, Marine Giard, Jean-Michel Thiolet, Jean-François Viel, et al. Improving prediction of surgical site infection risk with multilevel modeling. *PloS one*, 9(5):e95295, 2014.

[54] AL Brownell, BG Jenkins, and O Isacson. Dopamine imaging markers and predictive mathematical models for progressive degeneration in parkinson's disease. *Biomedicine & pharmacotherapy*, 53(3):131–140, 1999.

[55] James P Gratwicke and Thomas Foltynie. Early nucleus basalis of meynert degeneration predicts cognitive decline in parkinson's disease. *Brain*, 141(1):7–10, 2017.

[56] Daniel A Llano, Saurabh Bundela, Raksha A Mudar, Viswanath Devanarayan, Alzheimer's Disease Neuroimaging Initiative (ADNI, et al. A multivariate predictive modeling approach reveals a novel csf peptide signature for both alzheimer's

disease state classification and for predicting future disease progression. *PloS one*, 12(8):e0182098, 2017.

[57] Chih-Chuan Chen and Sheng-Tun Li. Credit rating with a monotonicity-constrained support vector machine model. *Expert Systems with Applications*, 41(16):7235–7247, 2014.

[58] Moritz Allmaras, Wolfgang Bangerth, Jean Marie Linhart, Javier Polanco, Fang Wang, Kainan Wang, Jennifer Webster, and Sarah Zedler. Estimating parameters in physical models through bayesian inversion: a complete example. *SIAM Review*, 55(1):149–167, 2013.

[59] F Owen Hoffman and Jana S Hammonds. Propagation of uncertainty in risk assessments: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk analysis*, 14(5):707–712, 1994.

[60] N Meghdadi, Hanieh Niroomand-Oscuii, M Soltani, F Ghalichi, and M Pourgolmohammad. Brain tumor growth simulation: model validation through uncertainty quantification. *International Journal of System Assurance Engineering and Management*, 8(3):655–662, 2017.

[61] Francesco Montomoli, Mauro Carnevale, Antonio D'Ammaro, Michela Massini, and Simone Salvadori. *Uncertainty quantification in computational fluid dynamics and aircraft engines*. Springer, 2015.

[62] Saideep Nannapaneni and Sankaran Mahadevan. Uncertainty quantification in performance evaluation of manufacturing processes. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 996–1005. IEEE, 2014.

[63] Edwin Reynders, Kristof Maes, Geert Lombaert, and Guido De Roeck. Uncertainty quantification in operational modal analysis with stochastic subspace identification: validation and applications. *Mechanical Systems and Signal Processing*, 66:13–30, 2016.

[64] Amir Nobari, Huajiang Ouyang, and Paul Bannister. Uncertainty quantification of squeal instability via surrogate modelling. *Mechanical Systems and Signal Processing*, 60:887–908, 2015.

[65] Joe Collis, Anthony J Connor, Marcin Paczkowski, Pavitra Kannan, Joe Pitt-Francis, Helen M Byrne, and Matthew E Hubbard. Bayesian calibration, validation and uncertainty quantification for predictive modelling of tumour growth: a tutorial. *Bulletin of mathematical biology*, 79(4):939–974, 2017.

[66] Giovanni Biglino, Claudio Capelli, Jan Bruse, Giorgia M Bosi, Andrew M Taylor, and Silvia Schievano. Computational modelling for congenital heart disease: how far are we from clinical translation? *Heart*, 103(2):98–103, 2017.

[67] Silvia Bozzi, Umberto Morbiducci, Diego Gallo, Raffaele Ponzini, Giovanna Rizzo,

Cristina Bignardi, and Giuseppe Passoni. Uncertainty propagation of phase contrast-mri derived inlet boundary conditions in computational hemodynamics models of thoracic aorta. *Computer methods in biomechanics and biomedical engineering*, 20(10):1104–1112, 2017.

[68] Kaibo Liu, Nagi Z Gebraeel, and Jianjun Shi. A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis. *IEEE Transactions on Automation Science and Engineering*, 10(3):652–664, 2013.

[69] Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632, 2012.

[70] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.

[71] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, Enchi Liu, et al. The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 2013.

[72] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Jesse Cedarbaum, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, et al. 2014 update of the alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & dementia*, 11(6):e1–e120, 2015.

[73] Alessandro Biffi, Christopher D Anderson, Rahul S Desikan, Mert Sabuncu, Lynelle Cortellini, Nick Schmansky, David Salat, and Jonathan Rosand. Genetic variation and neuroimaging measures in alzheimer disease. *Archives of neurology*, 67(6):677–685, 2010.

[74] Christiane Reitz, Ming-Xin Tang, Nicole Schupf, Jennifer J Manly, Richard Mayeux, and José A Luchsinger. A summary risk score for the prediction of alzheimer disease in elderly persons. *Archives of neurology*, 67(7):835–841, 2010.

[75] Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, pages 571–591, 2003.

[76] Gareth M James. Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):411–432, 2002.

[77] Philip T Reiss and R Todd Ogden. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996, 2007.

[78] Dehan Kong, Joseph G Ibrahim, Eunjee Lee, and Hongtu Zhu. Flcrm: Functional linear cox regression model. *Biometrics*, 74(1):109–117, 2018.

[79] Jeff Goldsmith, Ciprian M Crainiceanu, Brian Caffo, and Daniel Reich. Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(3):453–469, 2012.

[80] Kimberly M Thompson, David E Burmaster, and Edmund AC Crouch3. Monte carlo techniques for quantitative uncertainty analysis in public health risk assessments. *Risk Analysis*, 12(1):53–63, 1992.

[81] Karl Claxton, Peter J Neumann, Sally Araki, and Milton C Weinstein. Bayesian value-of-information analysis: an application to a policy model of alzheimer's disease. *International Journal of Technology Assessment in Health Care*, 17(1):38–55, 2001.

[82] Parvin Eslami Shahrbabki, Bahador Hajimohammadi, Shahram Shoeibi, Mehdi Elmi, Arash Yousefzadeh, Gea Oliveri Conti, Margherita Ferrante, Maryam Amirahmadi, Yadolah Fakhri, and Amin Mousavi Khaneghah. Probabilistic non-carcinogenic and carcinogenic risk assessments (monte carlo simulation method) of the measured acrylamide content in tah-dig using quechers extraction and uhplc-ms/ms. *Food and chemical toxicology*, 118:361–370, 2018.

[83] Dimitris Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011.

[84] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[85] Alan Julian Izenman. Modern multivariate statistical techniques. *Regression, classification and manifold learning*, 2008.

[86] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.

[87] Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In *Advances in neural information processing systems*, pages 470–476, 2000.

[88] Shiliang Sun and Guoqing Chao. Multi-view maximum entropy discrimination. In *IJCAI*, pages 1706–1712, 2013.

[89] Guoqing Chao and Shiliang Sun. Semi-supervised multi-view maximum entropy discrimination with expectation laplacian regularization. *Information Fusion*, 45:296–306, 2019.

[90] Changming Zhu and Zhe Wang. Semi-supervised soft margin consistency based multi-view maximum entropy discrimination. *Applied Computing and Informatics*, 2018.

[91] Shiliang Sun, Yuhan Liu, and Liang Mao. Multi-view learning for visual violence recognition with maximum entropy discrimination and deep features. *Information Fusion*, 2018.

[92] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[93] Alex J Mitchell, Amol Vaze, and Sanjay Rao. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, 374(9690):609–619, 2009.

[94] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1):163–173, 2009.

[95] Jonghwa Kim. Bimodal emotion recognition using speech and physiological changes. In *In Robust Speech Recognition and Understanding*. Citeseer, 2007.

[96] Eric S Holmboe and Steven J Durning. Assessing clinical reasoning: moving from in vitro to in vivo. *Diagnosis*, 1(1):111–117, 2014.

[97] Bertrand Rivet, Wenwu Wang, Syed Mohsen Naqvi, and Jonathon A Chambers. Audiovisual speech source separation: An overview of key methodologies. *IEEE Signal Processing Magazine*, 31(3):125–134, 2014.

[98] Matthew J Beal, Nebojsa Jojic, and Hagai Attias. A graphical model for audiovisual object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[99] Shankar T Shivappa, Mohan Manubhai Trivedi, and Bhaskar D Rao. Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 98(10):1692–1715, 2010.

[100] Murray Alpert, Enrique R Pouget, and Raul R Silva. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69, 2001.

[101] Lawrence Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE transactions on acoustics, speech, and signal processing*, 25(1):24–33, 1977.

[102] Dr Rajni and Nripendra Narayan Das. Emotion recognition from audio signal.

[103] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4(2):142–150, 2013.

[104] HH Stassen et al. Speaking behavior and voice sound characteristics in depressive patients during recovery. *Journal of psychiatric research*, 27(3):289–307, 1993.

[105] Ling He, Margaret Lech, and Nicholas Allen. On the importance of glottal flow spectral energy for the recognition of emotions in speech. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[106] Barbara A Bettes. Maternal depression and motherese: Temporal and intonational features. *Child development*, pages 1089–1096, 1988.

[107] Michael Cannizzaro, Brian Harel, Nicole Reilly, Phillip Chappell, and Peter J Snyder. Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56(1):30–35, 2004.

[108] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, and Gordon Parker. Detecting depression: a comparison between spontaneous and read speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7547–7551. IEEE, 2013.

[109] Johanna D Moore, Leimin Tian, Catherine Lai, and A Gelbukh. Word-level emotion recognition using high-level features. In *CICLing (2)*, pages 17–31, 2014.

[110] Stanley Newman and Vera G Mather. Analysis of spoken language of patients with affective disorders. *American journal of psychiatry*, 94(4):913–942, 1938.

[111] Sharifa Alghowinem, Roland Goecke, Julien Epps, Michael Wagner, and Jeffrey F Cohn. Cross-cultural depression recognition from vocal biomarkers. In *INTERSPEECH*, pages 1943–1947, 2016.

[112] Gary E Schwartz, Paul L Fair, Patricia Salt, Michael R Mandel, and Gerald L Klerman. Facial muscle patterning to affective imagery in depressed and nondepressed subjects. *Science*, 192(4238):489–491, 1976.

[113] Bowen Cheng, Zhangyang Wang, Zhaobin Zhang, Zhu Li, Ding Liu, Jianchao Yang, Shuai Huang, and Thomas S Huang. Robust emotion recognition from low quality and low bit rate video: A deep learning approach. *arXiv preprint arXiv:1709.03126*, 2017.

[114] Kuan Ee Brian Ooi, Lu-Shih Alex Low, Margaret Lech, and Nicholas Allen. Prediction of clinical depression in adolescents using facial image analaysis. In *WIAMIS 2011: 12th International Workshop on Image Analysis for Multimedia Interactive Services, Delft, The Netherlands, April 13-15, 2011*. TU Delft; EWI; MM; PRB, 2011.

[115] Jihun Hamm, Christian G Kohler, Ruben C Gur, and Ragini Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, 200(2):237–256, 2011.

[116] Sayan Ghosh, Moitreya Chatterjee, and Louis-Philippe Morency. A multimodal

context-based approach for distress assessment. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 240–246. ACM, 2014.

[117] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017 – real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2017.

[118] Aven Samareh, Yan Jin, Zhangyang Wang, Xiangyu Chang, and Shuai Huang. Predicting depression severity by multi-modal feature engineering and fusion. *arXiv preprint arXiv:1711.11155*, 2017.

[119] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender. *Journal on Multimodal User Interfaces*, 9(1):17–29, 2015.

[120] Susan Nolen-Hoeksema. Sex differences in unipolar depression: evidence and theory. *Psychological bulletin*, 101(2):259, 1987.

[121] Stephanie A Shields. Thinking about gender, thinking about theory: Gender and emotional experience. *Gender and emotion: Social psychological perspectives*, pages 3–23, 2000.

[122] Richard D Alford. Sex differences in lateral facial facility: The effects of habitual emotional concealment. *Neuropsychologia*, 21(5):567–570, 1983.

[123] Judith A Hall. *Nonverbal sex differences: Accuracy of communication and expressive style*. Johns Hopkins University Press, 1990.

[124] Zhaocheng Huang, Brian Stasak, Ting Dang, Kalani Wataraka Gamage, Phu Le, Vidhyasaharan Sethu, and Julien Epps. Staircase regression in oa rvm, data selection and gender dependency in avec 2016. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 19–26. ACM, 2016.

[125] Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, Panagiotis Simos, Kostas Marias, et al. Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 27–34. ACM, 2016.

[126] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 147–152. IEEE, 2013.

[127] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parkerx, and Michael

Breakspear. Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 283–288. IEEE, 2013.

[128] Tom Zahavy, Alessandro Magnani, Abhinandan Krishnan, and Shie Mannor. Is a picture worth a thousand words? a deep multi-modal fusion architecture for product classification in e-commerce. *arXiv preprint arXiv:1611.09534*, 2016.

[129] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[130] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[131] Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, Panagiotis Simos, Kostas Marias, et al. Depression assessment by fusing high and low level features from audio, video, and text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 27–34. ACM, 2016.

[132] Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 89–96. ACM, 2016.

[133] Heiner Ellgring. *Non-verbal communication in depression*. Cambridge University Press, 2007.

[134] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[135] Aron W Siegman and Stephen Boyle. Voices of fear and anxiety and sadness and depression: the effects of speech rate and loudness on fear and anxiety and sadness and depression. *Journal of Abnormal Psychology*, 102(3):430, 1993.

[136] John W Morley and Mark J Rowe. Perceived pitch of vibrotactile stimuli: effects of vibration amplitude, and implications for vibration frequency coding. *The Journal of physiology*, 431(1):403–416, 1990.

[137] CJ Darwin, Valter Ciocca, and GJ Sandell. Effects of frequency and amplitude modulation on the pitch of a complex tone with a mistuned harmonic. *The Journal of the Acoustical Society of America*, 95(5):2631–2636, 1994.

[138] A Michael Noll. Cepstrum pitch determination. *The journal of the acoustical society of America*, 41(2):293–309, 1967.

[139] Mohan Sondhi. New methods of pitch extraction. *IEEE Transactions on audio and electroacoustics*, 16(2):262–266, 1968.

[140] A Seither-Preisler, K Krumbholz, R Patterson, S Seither, and B Lütkenhöner. Interaction between the neuromagnetic responses to sound energy onset and pitch onset suggests common generators. *European Journal of Neuroscience*, 19(11):3073–3080, 2004.

[141] Mary Kay Biaggio and William H Godwin. Relation of depression to anger and hostility constructs. *Psychological Reports*, 61(1):87–90, 1987.

[142] Tobias Grossmann. The development of emotion perception in face and voice during infancy. *Restorative neurology and neuroscience*, 28(2):219–236, 2010.

[143] Takashi Saito. On the use of f0 features in automatic segmentation for speech synthesis. In *ICSLP*, 1998.

[144] Felix Burkhardt, Tim Polzehl, Joachim Stegmann, Florian Metze, and Richard Huber. Detecting real life anger. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4761–4764. IEEE, 2009.

[145] Laura Fonzi, Gabriella Matteucci, and Giuseppe Bersani. Laughter and depression: hypothesis of pathogenic and therapeutic correlation. *Rivista di psichiatria*, 45(1):1–6, 2010.

[146] Khiet P Truong and David A Van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[147] Jay J Bauer, Jay Mittal, Charles R Larson, and Timothy C Hain. Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude. *The Journal of the Acoustical Society of America*, 119(4):2363–2371, 2006.

[148] Patricia Gramming, Johan Sundberg, Sten Ternström, Rolf Leanderson, and William H Perkins. Relationship between changes in voice pitch and loudness. *Journal of Voice*, 2(2):118–126, 1988.

[149] Axel Röbel and Xavier Rodet. Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *International Conference on Digital Audio Effects*, pages 30–35, 2005.

[150] Iris Bohnet and Bruno S Frey. Social distance and other-regarding behavior in dictator games: Comment. *The American Economic Review*, 89(1):335–339, 1999.

[151] S Sato, T Kitamura, and Y Ando. Loudness of sharply (2068 db/octave) filtered noises in relation to the factors extracted from the autocorrelation function. *Journal of Sound and Vibration*, 250(1):47–52, 2002.

[152] André Holzapfel, Yannis Stylianou, Ali C Gedik, and Barış Bozkurt. Three dimensions of pitched instrument onset detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1517–1527, 2010.

[153] Paul Brossier, Juan Pablo Bello, and Mark D Plumbley. Real-time temporal segmentation of note objects in music signals. In *Proceedings of ICMC 2004, the 30th Annual International Computer Music Conference*, 2004.

[154] Harvey Fletcher and JC Steinberg. The dependence of the loudness of a complex sound upon the energy in the various frequency regions of the sound. *Physical Review*, 24(3):306, 1924.

[155] N Geckinli and Davras Yavuz. Algorithm for pitch extraction using zero-crossing interval sequence. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(6):559–564, 1977.

[156] Nurgun Erdol, Claude Castelluccia, and Ali Zilouchian. Recovery of missing speech packets using the short-time energy and zero-crossing measurements. *IEEE Transactions on Speech and Audio Processing*, 1(3):295–303, 1993.

[157] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[158] Petr Pollak, Pavel Sovka, and Jan Uhlir. Cepstral speech/pause detectors. In *Proc. IEEE Workshop on Nonlinear Signal and Image Processing*, pages 388–391, 1995.

[159] B Atal and L Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(3):201–212, 1976.

[160] Rafa l Samborski, David Sierra, et al. Hybrid wavelet-fourier-hmm speaker recognition. *International Journal of Hybrid Information Technology*, 4(4):25–42, 2011.

[161] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[162] HP Combrinck and EC Botha. On the mel-scaled cepstrum. *department of Electrical and Electronic Engineering, University of Pretoria*, 1996.

[163] Vaishnevi S Varadarajan and John HL Hansen. Analysis of lombard effect under different types and levels of noise with application to in-set speaker id systems. In *Ninth International Conference on Spoken Language Processing*, 2006.

[164] Michael Heller and Véronique Haynal. Depression and suicide faces. *What the face reveals'(Oxford, 1997)*, pages 398–407, 1997.

[165] Campbell Leaper and Tara E Smith. A meta-analytic review of gender variations in children's language use: talkativeness, affiliative speech, and assertive speech. *Developmental psychology*, 40(6):993, 2004.

[166] Yen-Ting Chen, I-Chung Hung, Min-Wei Huang, Chun-Ju Hou, and Kuo-Sheng Cheng. Physiological signal analysis for patients with depression. In *Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on*, volume 2, pages 805–808. IEEE, 2011.

[167] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8. ACM, 2015.

[168] Christian Cavé, Isabelle Guaïtella, and Serge Santi. Eyebrow movements and voice variations in dialogue situations: an experimental investigation. In *Seventh International Conference on Spoken Language Processing*, 2002.

[169] Omer Berat Sezer, Erdogan Dogdu, and Ahmet Murat Ozbayoglu. Context-aware computing, learning, and big data in internet of things: a survey. *IEEE Internet of Things Journal*, 5(1):1–27, 2018.

[170] Anind K Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.

[171] Daniele Riboni and Claudio Bettini. Cosar: hybrid reasoning for context-aware activity recognition. *Personal and Ubiquitous Computing*, 15(3):271–289, 2011.

[172] Jie Xu, Linqi Song, James Y Xu, Gregory J Pottie, and Mihaela Van Der Schaar. Personalized active learning for activity classification using wireless wearable sensors. *IEEE journal of selected topics in signal processing*, 10(5):865–876, 2016.

[173] Javier Andreu-Perez, Daniel R Leff, Henry MD Ip, and Guang-Zhong Yang. From wearable sensors to smart implants—toward pervasive and personalized healthcare. *IEEE Transactions on Biomedical Engineering*, 62(12):2750–2762, 2015.

[174] Brennan MR Spiegel, Marc Kaneshiro, Marcia M Russell, Anne Lin, Anish Patel, Vartan C Tashjian, Vincent Zegarski, Digvijay Singh, Samuel E Cohen, Mark W Reid, et al. Validation of an acoustic gastrointestinal surveillance biosensor for postoperative ileus. *Journal of Gastrointestinal Surgery*, 18(10):1795–1803, 2014.

[175] James Y Xu, Hua-I Chang, Chieh Chien, William J Kaiser, and Gregory J Pottie. Context-driven, prescription-based personal activity classification: methodology, ar-

chitecture, and end-to-end implementation. *IEEE journal of biomedical and health informatics*, 18(3):1015–1025, 2014.

[176] Odongo Steven Eyobu and Dong Han. Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network. *Sensors*, 18(9):2892, 2018.

[177] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. Activity recognition from accelerometer data. In *Aaai*, volume 5, pages 1541–1546, 2005.

[178] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.

[179] 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*, pages 806–813, 2014.

[180] Joris Guérin, Olivier Gibaru, Stéphane Thiery, and Eric Nyiri. Cnn features are also great at unsupervised classification. *arXiv preprint arXiv:1707.01700*, 2017.

[181] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[182] Randy Ardywibowo, Guang Zhao, Zhangyang Wang, Bobak Mortazavi, Shuai Huang, and Xiaoning Qian. Adaptive activity monitoring with uncertainty quantification in switching gaussian process models. *arXiv preprint arXiv:1901.02427*, 2019.

[183] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[184] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

[185] Shiyu Liang, Yixuan Li, and R Srikant. Principled detection of out-of-distribution examples in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[186] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[187] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[188] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.

[189] Daniel Bauer, Lars Kuhnert, and Lutz Eckstein. Deep, spatially coherent inverse sensor models with uncertainty incorporation using the evidential framework. *arXiv preprint arXiv:1904.00842*, 2019.

[190] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*, pages 233–240. IEEE, 2010.

[191] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[192] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

# Appendix A

## *A.1 CHI model formulation*

For completeness of DL-CHI, here we present more details of the CHI formulation (3.1). The CHI formulation is convex but contains multiple non-smooth terms such as (3.1b), (3.1c), and (3.1f). To solve this formulation, we could merge the smooth terms and derive the dual optimization problem, and finally train it via the block coordinate descent algorithm. Specifically, we can simplify Eq. (3.1) in a quadratic forms by defining:

$$\begin{aligned}\|\mathbf{w}\|_Q^2 &:= \mathbf{w}^\top Q\mathbf{w} = \|\mathbf{w}\|^2 \\ &+ \frac{\lambda}{2}\left(\frac{1}{N^+}\sum_{n\in\{N^+|y_n=1\}}((\mathbf{x}_{n,T_n}-\bar{\mathbf{x}}_{T_n}^+)^T\mathbf{w})^2\right)+ \\ &+ \frac{\lambda}{2}\left(\frac{1}{N^-}\sum_{n\in\{N^-|y_n=1\}}((\mathbf{x}_{n,T_n}-\bar{\mathbf{x}}_{T_n}^-)^T\mathbf{w})^2\right).\end{aligned} \tag{A.1}$$

where $Q$ is defined as

$$\begin{aligned}Q := \mathbf{I} + \lambda\Bigg(&\frac{1}{N^+}\sum_{n\in\{n|y_n=1\}}(\mathbf{x}_{n,T_n}-\bar{\mathbf{x}}_{T_n}^+)(\mathbf{x}_{n,T_n}-\mathbf{x}_{T_n}^+)^\top \\ &\frac{1}{N^-}\sum_{n\in\{n|y_n=-1\}}(\mathbf{x}_{n,T_n}-\bar{\mathbf{x}}_{T_n}^-)(\mathbf{x}_{n,T_n}-\bar{\mathbf{x}}_{T_n}^-)^\top\Bigg).\end{aligned}$$

With that, Eq. (A.1) is simplified to Eq. (A.2) as follows:

$$\begin{aligned}\min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_Q^2 + \gamma\|\mathbf{w}\|_1 + \\ & \alpha \sum_{\substack{n\in\{1,\cdots,N\} \\ t\in\{1,\cdots,T_n-1\}}} \max(0, 1 - \mathbf{z}_{n,t}^\top \mathbf{w}) + \\ & \beta \sum_{n\in\{1,\cdots,N\}} \max(0, 1 - y_n(\mathbf{x}_{n,T_n}^\top \mathbf{w} + b)).\end{aligned} \tag{A.2}$$

By introducing two relaxation variables $\boldsymbol{\xi}$ and $\boldsymbol{\epsilon}$, Eq. (A.2) is equivalent to Eq. (A.3) as follows:

$$\begin{aligned}\min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_Q^2 + \alpha\mathbf{1}^\top\xi + \beta\mathbf{1}^\top\epsilon + \gamma\|\mathbf{w}\|_1 \\ \text{s.t.} \quad & \mathbf{1} - Z^\top\mathbf{w} - \boldsymbol{\xi} \leq \mathbf{0} \\ & \mathbf{1} - \hat{X}^\top\mathbf{w} - b\mathbf{y} - \boldsymbol{\epsilon} \leq \mathbf{0}\end{aligned} \tag{A.3}$$

where,

$$\begin{aligned}&\boldsymbol{\xi} = (\boldsymbol{\xi}_{1,1}, \cdots, \boldsymbol{\xi}_{1,T_1-1}, \cdots, \boldsymbol{\xi}_{N,1}, \cdots, \boldsymbol{\xi}_{N,T_N-1})^\top, \\ &Z = (Z_{1,1}, \cdots, Z_{1,T_1-1}, \cdots, Z_{N,1}, \cdots, Z_{N,T_N-1}), \\ &\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_N)^\top, \\ &\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^\top, \\ &\hat{X} = (\mathbf{y}_1 X_{1,T_1}, \cdots, \mathbf{y}_N X_{N,T_N}).\end{aligned}$$

We then can derive the dual formulation of (A.3) by substituting the $\ell_1$-norm penalty in (A.3) by its conjugate norm $\|\mathbf{w}\|_1 = \max_{\|\mathbf{s}\|_\infty \leq 1} \langle \mathbf{s}, \mathbf{w} \rangle = \max_{\|\mathbf{s}\|_\infty \leq 1} -\langle \mathbf{s}, \mathbf{w} \rangle$, and then introducing two new dual variables $\mathbf{u}$ and $\mathbf{v}$ which leads to the following formulation:

$$\min_{\substack{\mathbf{w},b \\ \epsilon\geq\mathbf{0} \\ \xi\geq\mathbf{0}}} \max_{\substack{\mathbf{u}\geq\mathbf{0} \\ \mathbf{v}\geq\mathbf{0} \\ \|\mathbf{s}\|_\infty\leq\gamma}} \quad \tfrac{1}{2}\|\mathbf{w}\|_Q^2 + \alpha\mathbf{1}^\top\xi + \beta\mathbf{1}^\top\epsilon - \langle \mathbf{w}, \mathbf{s} \rangle +$$

$$\langle \mathbf{u}, 1 - Z^\top\mathbf{w} - \xi \rangle + \langle \mathbf{v}, 1 - \hat{X}^\top\mathbf{w} - b\mathbf{y} - \epsilon \rangle.$$

This can be rewritten as the following constrained smooth convex optimization problem,

which can be solved efficiently:

$$\begin{aligned} \min_{\mathbf{s},\mathbf{u},\mathbf{v}} \quad & F(\mathbf{s},\mathbf{u},\mathbf{v}) := \frac{1}{2}\|\mathbf{s} + Z\mathbf{u} + \hat{X}\mathbf{v}\|^2_{Q^{-1}} - \langle \mathbf{1}, \mathbf{u}\rangle - \langle \mathbf{1}, \mathbf{v}\rangle \\ \text{s.t.} \quad & \mathbf{0} \leq \mathbf{u} \leq \alpha \mathbf{1} \\ & \mathbf{0} \leq \mathbf{v} \leq \beta \mathbf{1} \\ & \langle \mathbf{v}, \mathbf{y}\rangle = 0 \\ & \|\mathbf{s}\|_\infty \leq \gamma. \end{aligned} \tag{A.4}$$

Then the solution $\mathbf{w}^*$ to Eq. (A.4) can be obtained by:
$\mathbf{w}^* = Q^{-1}(\mathbf{s}^* + Z\mathbf{u}^* + \widehat{X}\mathbf{v}^*)$.

### *A.2 The block coordinate descent algorithm*

The block coordinate descent algorithm [192] to solve the dual problem in Eq. (A.4) is an iterative procedure as follows:

**Algorithm 3** Block Coordinate Descent for Solving CHI

**Require:** Problem parameters $\{Z, Q, \widehat{X}, \mathbf{y}, \alpha, \beta, \gamma\})$ and Optimization parameters $\eta_s, \eta_u, \eta_v, \rho \in (0, 1)$ (step sizes $\eta_s$, $\eta_u$, and $\eta_v$ can be adaptively decided using linear search alternatively)

**Ensure:** $\mathbf{w}^*, b^*$

1: Initialize $k = 0$

2: **While** not converge do

3:

4: $\mathbf{s}_{k+1} = \max(-\gamma\mathbf{1}, \min(\gamma 1, \mathbf{s}_k - \eta_s \nabla_s F(\mathbf{s}_k, \mathbf{u}_k, \mathbf{v}_k)))$

5: $\mathbf{u}_{k+1} = \max(\mathbf{0}, \min(\alpha\mathbf{1}, u_k - \eta_{\mathbf{u}} \nabla_{\mathbf{s}} F(s_{k+1}, \mathbf{u}_k, \mathbf{v}_k)))$

6: $\mathbf{v}_{k+1} = \text{Proj}_{\substack{\mathbf{0}\leq\mathbf{v}\leq\beta\mathbf{1} \\ \langle\mathbf{v},\mathbf{y}\rangle=0}}(\mathbf{v}_k - \eta_{\mathbf{v}} \nabla_{\mathbf{v}} F(\mathbf{s}_{k+1}, \mathbf{u}_{k+1}, \mathbf{v}_k))$

7: $k \leftarrow k + 1$

8: Recover the primal variables

$$\mathbf{w}^* = Q^{-1}(\mathbf{s} + Z\mathbf{u} + \widehat{X}\mathbf{v})$$

$$b^* = \sum_{\{i|\mathbf{v}_i\in(0,\beta)\}} \mathbf{y}_i - \mathbf{w}^{*T}\mathbf{x}_i$$

# Appendix B

## *B.1 Proof to Lemma 3.3.1*

*Proof.* By adding a set of dual variables, one for each constraint, the Lagrangian of the optimization problem in (3.3) can be written as:

$$\mathcal{L}\big(p(\mathbf{w},\lambda)\big) = KL\big(p(\mathbf{w})||p_0(\mathbf{w})\big) - \sum_n \lambda_n\big(y_n E_{p(\mathbf{w})}[\mathcal{D}\left(X_n|\mathbf{w}\right)] - 1\big), \tag{B.1}$$

In order to find the solution to Eq. (3.3), and given the definition of the KL-divergence in (3.5) we require,

$$\frac{\partial \mathcal{L}}{\partial p\left(\mathbf{w}\right)} = \log p(\mathbf{w}) - \log p_0(\mathbf{w}) - \sum_n \lambda_n y_n \mathcal{D}\left(X_n|\mathbf{w}\right) = 0, \tag{B.2}$$

The solution to the MED optimization problem has the following general form:

$$p(\mathbf{w}^*) = \frac{1}{Z(\lambda)} p_0(\mathbf{w}) \exp\left(\sum_n \lambda_n y_n \mathcal{D}\left(X_n|\mathbf{w}\right)\right). \tag{B.3}$$

Here, $Z(\lambda)$ is the normalization constant defined in (3.10), then the general exponential form of the solution becomes:

$$\begin{aligned} g(\lambda) &= \mathcal{L}\big(p(\mathbf{w}^*),\lambda\big) \\ &= \int \frac{1}{Z(\lambda)} p_0(\mathbf{w}) \exp\left(\sum_n \lambda_n y_n \mathcal{D}\left(X_n|\mathbf{w}\right)\right) \\ &\quad \left(\sum_n \lambda_n y_n \mathcal{D}\left(X_n|\mathbf{w}\right) - \log p_0(\mathbf{w}) - log(\mathbf{Z})\right) d\mathbf{w} - \\ &\quad \sum_n \lambda_n \left(y_n \int \frac{1}{Z(\lambda)} p_0(\mathbf{w}) \exp\left(\sum_n \lambda_n y_n \mathcal{D}\left(X_n|\mathbf{w}\right)\right)\right. \\ &\quad \left.\mathcal{D}\left(X_n|\mathbf{w}\right) d\mathbf{w} - 1\right) \\ &= \sum_n \lambda_n - \log Z(\lambda). \end{aligned} \tag{B.4}$$

□

Hence, the dual of the MED problem can be shown in (3.9).

### *B.2 Proof to Lemma 3.3.3*

*Proof.* Let $Z(\lambda)$ be the normalization constant defined in Eq. (3.10), given the constraints in (3.8) the normalization constant can be reformulated as follows:

$$Z(\lambda) = \int p_0(\mathbf{w}, \gamma) \tag{B.5a}$$

$$\exp\Bigg( \sum_{n\in\{1,\dots,N\}} \lambda_n [p(y_n)\mathcal{D}(\mathbf{x}_{n,Tn} \mid \mathbf{w}) - \gamma_n] + \tag{B.5b}$$

$$\sum_{\substack{n\in\{1,\dots,N\}\\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \left[\mathcal{M}\left(\mathbf{z}_{n,t}|\mathbf{w}\right) - \gamma_n\right] \Bigg) d\mathbf{w} d\gamma, \tag{B.5c}$$

$$= \int P_0(\mathbf{w}) \exp\Bigg( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{w}^T \mathbf{x}_{n,Tn} + \tag{B.5d}$$

$$\sum_{\substack{n\in\{1,\dots,N\}\\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{w}^T \mathbf{z}_{n,t} \Bigg) \tag{B.5e}$$

$$P_0(\gamma) \exp\Bigg( - \sum_{n\in\{1,\dots,N\}} \lambda_n \gamma_n \Bigg) d\mathbf{w} d\gamma, \tag{B.5f}$$

Given the priors in (3.7), each term in Eq. (B.5) can be reformulated as follows: For the term in (B.5d) and (B.5e) we have the followings:

$$
\begin{aligned}
Z_{\mathbf{w}}(\lambda) &= \int P_0(\mathbf{w}) \exp\Bigg( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{w}^T \mathbf{x}_{n,Tn} + \\
&\quad \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{w}^T \mathbf{z}_{n,t} \Bigg) d\mathbf{w} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\Big( -\frac{1}{2} \mathbf{w}^T \mathbf{w} \Big) \exp\Bigg( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{w}^T \mathbf{x}_{n,Tn} + \\
&\quad \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{w}^T \mathbf{z}_{n,t} \Bigg) d\mathbf{w} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\Bigg( -\frac{1}{2} \Bigg( \mathbf{w}^T \mathbf{w} - 2\mathbf{w}^T \Big( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{x}_{n,Tn} + \\
&\quad \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t} \Big) \Bigg) \Bigg) d\mathbf{w} \\
&= \exp\Bigg( \frac{1}{2} \Big( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{x}_{n,Tn} + \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t} \Big)^T \\
&\quad \Big( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{x}_{n,Tn} + \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t} \Big) \Bigg) \\
&\quad \int_{-\infty}^{\infty} \exp\Bigg( -\frac{1}{2} \Bigg( \mathbf{w} - \Big( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{x}_{n,Tn} + \\
&\quad \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t} \Big) \Bigg)^T \\
&\quad \Bigg( \mathbf{w} - \Big( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{x}_{n,Tn} + \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t} \Big) \Bigg) \Bigg) \\
&= \exp\Bigg( \frac{1}{2} \Big( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{x}_{n,Tn} + \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t} \Big)^T \\
&\quad \Big( \sum_{n\in\{1,\dots,N\}} \lambda_n p_n(y_n) \mathbf{x}_{n,Tn} + \sum_{\substack{n\in\{1,\dots,N\} \\ t\in\{1,\dots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t} \Big) \Bigg),
\end{aligned}
\tag{B.6}
$$

And for the last term (B.5f) we have the following:

$$\begin{aligned} Z_\gamma(\lambda) &= P_0\left(\gamma\right) \exp\big(- \sum_{n\in\{1,\ldots,N\}} \lambda_n\gamma_n\big) d\gamma \\ &= \prod_1^n \int P_0\left(\gamma_n\right) \exp\big(-\lambda_n\gamma_n\big) d\gamma_n \\ &= \int_{-\infty}^1 c \exp\big(-c + c\gamma_n\big) \exp\big(-\lambda_n\gamma_n\big) d\gamma_n \\ &= \frac{c}{c-\lambda_n} \exp\big(-\lambda_n\big) \\ &= \frac{1}{1-\lambda_t/c} \exp\big(-\lambda_n\big), \end{aligned} \tag{B.7}$$

Substituting the results from Eq. (B.6) and (B.7) in (B.5), results in Eq. (3.14). □

### *B.3 Proof to Lemma 3.3.5*

*Proof.* Given the marginal distribution $p\left(\mathbf{w}\right)$ in (3.17) and the convex combination of discriminant functions defined as $\int p(\mathbf{w})\mathcal{D}\left(\mathbf{x}|\mathbf{w}\right) d\mathbf{w}$ we have the following:

$$\begin{aligned} \int p(\mathbf{w})\mathcal{D}\left(\mathbf{x}|\mathbf{w}\right) d\mathbf{w} &= \int p(\mathbf{w})(\mathbf{w}^T\mathbf{x}_{new}) d\mathbf{w} \\ &= \frac{1}{Z_{\mathbf{w}}(\lambda)} \mathbf{w}^T\mathbf{x}_{new} p_0\left(\mathbf{w}\right) \\ &\quad \exp\Bigg( \sum_{n\in\{1,\ldots,N\}} \lambda_n p_n(y_n) \mathbf{w}^T\mathbf{x}_{n,Tn} + \\ &\quad \sum_{\substack{n\in\{1,\ldots,N\} \\ t\in\{1,\ldots,T_{n-1}\}}} \lambda_n \mathbf{w}^T\mathbf{z}_{n,t} \Bigg) d\mathbf{w}, \end{aligned} \tag{B.8}$$

Where, given the prior distributions, (B.8) can be written as:

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \frac{\mathbf{w}^T \mathbf{x}_{new}}{Z_{\mathbf{w}}(\lambda)\sqrt{2\pi}} \exp\Bigg( -\frac{1}{2}\mathbf{w}^T\mathbf{w} + \mathbf{w}^T\Bigg( \sum_{n\in\{1,\ldots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} + \\
&\sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Bigg)\Bigg) d\mathbf{w} \\
&= \int_{-\infty}^{\infty} \frac{\mathbf{w}^T \mathbf{x}_{new}}{\sqrt{2\pi}} \exp\Bigg( \quad \frac{-1}{2}\Bigg(\mathbf{w} - \Big( \sum_{n\in\{1,\ldots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} + \\
&\sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Big)\Bigg)^T \Bigg(\mathbf{w} - \Big( \sum_{n\in\{1,\ldots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} + \\
&\sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Big)\Bigg)\Bigg) d\mathbf{w} \\
&= \Bigg( \sum_{n\in\{1,\ldots,N\}} \lambda_n p_n(y_n)\mathbf{x}_{n,Tn} + \sum_{\substack{n\in\{1,\ldots,N\}\\ t\in\{1,\ldots,T_{n-1}\}}} \lambda_n \mathbf{z}_{n,t}\Bigg)^T \mathbf{x}_{new}.
\end{aligned}
\tag{B.9}
$$

□

# Appendix C

## *C.1 Uncertainty quantification*

**Analytical details of MEL** To further illustrate this, note that our context detection problem is also a classification problem where the response variable is denoted by $y$ taking values for different contexts. Let $\mathbf{x}_n = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be an input feature vector as an aggregate of all the measures from sensors for each window, and let $\mathcal{D}(\mathbf{x}_n|\mathbf{w})$ be the discriminant function parameterized by $\mathbf{w}$ implemented in the $\alpha$ network.

Traditional learning machines such as the max-margin methods estimate the optimal $\hat{\mathbf{w}}$ that minimizes the classification error in predicting the labels of training examples as:

$$\hat{y} = sign\mathcal{D}(\mathbf{x}_n|\mathbf{w}). \tag{C.1}$$

Based on this line of thought, we can classify margin as $y_n\mathcal{D}(\mathbf{x}_n, \mathbf{w})$, and learn the optimal parameter setting $\mathbf{w}$ by the empirical loss and the regularization penalty as:

$$\begin{aligned} &\min_{(\mathbf{w},\gamma_n)} R(\mathbf{w}) + \sum_n L(\gamma_n) \\ &\text{s.t.} \quad y_n\mathcal{D}(\mathbf{x}_n \mid \mathbf{w}) - \gamma_n \geq \mathbf{0}, \quad \forall n. \end{aligned} \tag{C.2}$$

where $L(\gamma_n)$ is the loss function, a non-increasing and convex function of the margin, and $R(\mathbf{w})$ is the regularization penalty.

Given $p(\mathbf{w})$, we can recast (C.2) as an integration where the classification constraints will also be applied in an expected sense. Instead of considering an expectation of the regularization penalty functions, we can apply a canonical penalty function for distributions, the negative entropy; minimizing the negative entropy is equivalent to maximizing the entropy. Hence,

we use the Shannon entropy defined as $H(p(\mathbf{w})) = -\int p(\mathbf{w}) \log p(\mathbf{w}) d\mathbf{w}$. This gives us the following objective function to learn the distribution $p(\mathbf{w})$ over the parameters $\mathbf{w}$:

$$\begin{aligned} &\min_{p(\mathbf{w})} H(p(\mathbf{w})) \\ &\text{s.t.} \quad \int p(\mathbf{w})[y_n \mathcal{D}(\mathbf{x}_n, \mathbf{w}) - \gamma_n] d\mathbf{w} \geq \mathbf{0}, \quad \forall n. \end{aligned} \tag{C.3}$$

As a result, MEL no longer finds a fixed set of the parameters, but a distribution over them. Learning such a distribution of model parameters does not rely on assumptions on the model's mathematical form. It also does not rely on knowing a particular distribution as is needed in Bayesian learning frameworks. Therefore, MEL is more flexible than typical Bayesian learning methods [91, 90] to characterize uncertainties associated with complex models such as the $\alpha$-$\beta$ network here.

To solve the MEL formulation (C.3), we could derive a Lagrangian $J(p)$, and take the derivatives with respect to $\mathbf{w}$ and set them to 0. To do that we first need to calculate the unconditional maximum of the problem (C.3) plus the constraints added with some multiplying factors (the Lagrange multipliers), which give the probabilities in a functional form with the Lagrange multipliers as parameters. Our UQ approach compares the uncertainty with a threshold to see whether a given sample should be detected as belonging to a new context or not. We selected the threshold based on a cross validation set, and we computed the resulting F-score, accuracy, sensitivity and specificity for different thresholds. We defined a classification with rejection option as $\hat{y}_i^{Rej}$, where if a sample is rejected $\hat{y}_i^{Rej} = 0$, and if it is accepted $\hat{y}_i = \hat{y}$, where $\hat{y}$ corresponds to the classification of the $i$th sample. Note that, a sample is rejected when $p(\mathbf{w}|x_i) < \epsilon$.