Deep Kernel Learning with Temporal Gaussian Processes for Clinical Variable Prediction in Alzheimer's Disease

Vasiliki Tassopoulou VTASS@SEAS.UPENN.EDU
Fanyang Yu YFY@SEAS.UPENN.EDU.COM
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, USA

Christos Davatzikos CHRISTOS.DAVATZIKOS@PENNMEDICINE.UPENN.EDU
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, USA
Department of Radiology, Perelman School of Medicine, University of Pennsylvania, USA

Abstract

Longitudinal prediction of Alzheimer's disease progression is of high importance for early diagnosis and clinical trial design. We propose to predict the longitudinal changes of neuroimaging biomarkers and cognitive scores by leveraging the expressivity of Deep Kernel Learning with single-task Gaussian Processes. The temporal function that describes the progression of the biomarker is learned through a Gaussian Process. By learning these temporal functions we can predict any future value of a clinical variable. We apply our method for extrapolation of neuroimaging biomarkers, SPARE-AD index, and cognitive metric Adas-Cog13, both significant predictors for the pathological and cognitive changes of Alzheimer's Disease. The method has been validated in two cohorts, ADNI and BLSA, where the results show that the proposed method significantly outperforms baselines and state-of-the-art models in AD progression prediction both on providing point estimates and quantifying uncertainty.

Keywords: Deep Kernel Learning, Gaussian Processes, Alzheimer's Disease Progression

1. Introduction

Alzheimer's disease (AD) is the most prevalent form of dementia and is estimated to reach 300 million cases worldwide by the year of 2050 (Rathore et al., 2017). Early detection and patient stratification are critical for disease treatment and clinical trial design. To this end, an automated approach that could forecast future clinical variables would be of great value during assessment procedures, and could potentially improve clinical trial design and early detection of at-risk subjects. Throughout the years, various predictors such as imaging-derived biomarkers and cognitive scores have been developed for that purpose. Particularly, we use the following two predictors for the longitudinal assessment of the disease progression. The Spatial Pattern of Abnormality for Recognition of Early Alzheimer's Disease (SPARE-AD) index (Davatzikos et al., 2009) is a neuroimaging biomarker derived from a high-dimensional pattern classification approach to indicate spatial patterns of brain atrophy. It is a highly sensitive predictor for subtle brain change to discriminate among cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD) subjects. The positive value of SPARE-AD reflects AD-pathological brain structure, while the negative value implies normal structure. The Adas-Cog13 is a widely used version of Alzheimer's Disease Assessment Scale-Cognitive Subscale (Adas-Cog) cognitive score, which is a gold standard for estimating the level of cognitive symptoms in AD. The score provides a fine-grained scale of 0–85 with higher value implying more severe cognitive dysfunction. The major challenges for the accurate longitudinal predictions of these biomarkers include missing values and noisy data, especially in a realistic clinical setting where the patient visit times can be highly limited and irregular.

In this paper, we propose a novel deep kernel learning framework with temporal single-task Gaussian Processes. Specifically, we aim to learn the temporal function that describes the evolution in time of two important clinical variables, the SPARE-AD score and the Adas-Cog13 score. Our approach tailors the deep architecture to handle multi-modal data including imaging, genomics and clinical information to learn a common embedding as the input to the Gaussian Process (GP) kernel. This approach leads to an end-to-end training scheme where the network parameters will be together optimized with respect to the log marginal likelihood from the GP inference. We evaluate our method across two neuroimaging study cohorts, Alzheimer's Disease Neuroimaging Initiative (ADNI) and Baltimore Longitudinal Study of Aging (BLSA) to demonstrate its generalizability and superior performance over other state-of-the-art methods in both aspects of providing the prediction but also with respect to uncertainty quantification. The code to reproduce the results of this paper will be made publicly available upon acceptance.

Section 2 briefly mentions the background work on forecasting clinical variables in the context of Alzheimer's Disease. Section 3 describes the data, the processing and the preparation scheme. Our methodology is then presented in Section 4. In Section 5, we report the experiments on the test sets, as well as additional analysis of the results.

2. Related Work

We briefly review here the related work on the forecasting of cognitive scores and clinical status for AD assessment, with the focus on the ADNI dataset. Most existing approaches (Schmidt-Richberg et al., 2016; Gavidia-Bovadilla et al., 2017; Guerrero et al., 2016), focus on modeling subjects based on their clinical status: CN, MCI, and AD. (Guerrero et al., 2016) use mixed effects modeling to derive global and individual biomarker trajectories for a training population, which was later used to instantiate subject-specific models for unseen subjects. Some of the modeling techniques (Guerrero et al., 2016; Schmidt-Richberg et al., 2016; Gavidia-Bovadilla et al., 2017) require cohorts with known disease onset and are prone to bias due to the uncertainty of the conversion time. Most of these works attempt forecasting the changes in subjects' clinical status, which deals with a limited number of future outcomes (i.e. either binary or a three-class). On the contrary, we aim to extrapolate variables with multiple levels as the Adas-Cog13 with 85 levels, and continuous variables such as the SPARE-AD score.

Recent work (Petersen et al., 2010; Utsumi et al., 2019) utilize Gaussian Processes to extrapolate the ADAS-Cog13 6 months and up to 24 months ahead respectively. In (Utsumi et al., 2019) they extrapolate 4 values at the same time with 6-month interval at each. Having equal time intervals among consecutive acquisitions is not a realistic scenario since obtaining acquisitions is not trivial and in numerous cases the subjects skip visits. This constrain is also critical when constructing the training set. For example, in (Utsumi et al., 2019) the model was trained only on approximately 100 subjects, from thousands of subjects in ADNI cohort. This reduced number of subjects is due to the strict filtering based on time intervals between consecutive visits. Additionally, in (Utsumi et al., 2019) the authors utilized multimodal features, using only a single kernel for all modalities. Furthermore, they extrapolate the Adas-Cog13 score up to two years ahead and have the same covariance function for all the four values, which indicates that the uncertainty is modeled to remain the same with respect to the time.

In our work, we use Deep Kernel Learning (DKL) (Wilson et al., 2016) to integrate the multimodal data in a single representation as input to the Gaussian Process layer. Deep neural networks (DNNs) offer the ability of representation learning which can be utilized for downstream tasks. Naturally, DKL has been proposed for combining both the representational power of DNN and uncertainty estimation of GP with benefits in expressivity and scalability. In this paper, we use DKL in order to learn the temporal functions that describe the progression of the biomarkers. With our method, we learn the whole temporal function and not only the function with 6-month intervals. Therefore, our task is harder than the one compared to (Utsumi et al., 2019).

3. Datasets and Preprocessing

3.1. Datasets

The Alzheimer's Disease Neuroimaging Initiative (ADNI, http://www.adni-info.org/) study is a public-private collaborative longitudinal cohort study which has recruited participants categorized as CN, MCI, and AD subjects through 4 phases (ADNI1, ADNIGO, ADNI2) (Weiner et al., 2017). ADNI has acquired longitudinal MRI, cerebrospinal fluid (CSF) biomarkers, and cognitive testing. The Baltimore Longitudinal Study of Aging (BLSA) has been following participants who are cognitively normal at enrollment with imaging and cognitive exams since 1993. Detailed information of enrollment criteria can be found in Petersen et al. (2010) for ADNI and in Resnick et al. (2003) for BLSA. Participants provided written informed consent to the ADNI and BLSA studies. The protocol of this study was approved by the University of Pennsylvania institutional review board.

3.2. Processing

1.5 T and 3T MR data were acquired from both ADNI and BLSA study.An automated preprocessing pipeline applies to T1 structural MRIs. At first, the T1-weighted scans of each participant are corrected for intensity inhomogeneities (Sled et al., 1998). A multi-atlas skull stripping algorithm was applied for the removal of extra-cranial material (Doshi et al., 2013). For the ADNI study, 145 anatomical regions of interest (ROIs) were identified in gray matter (GM, 119 ROIs), white matter (WM, 20 ROIs) and ventricles (6 ROIs) using a multi-atlas label fusion method (Doshi et al., 2016). For the BLSA study, this method was combined with harmonized acquisition-specific atlases (Erus et al., 2018) to derive the same 145 ROIs. Phase-level cross-sectional harmonization was applied on regional volumes of the 145 ROIs to remove site effects (Pomponio et al., 2020). Before being used as features for the Temporal Deep Kernel model, ROI volumes were residualized and variance-normalized. To correct age and sex effects while keeping disease-associated neuroanatomical variations, we estimated ROIs-specific age and sex associations among CN participants using a linear regression model. All cross-sectional and longitudinal data were then residualized by age and sex effects. Then, all ROI volumes were further normalized to ensure a mean of 1 and standard deviation of 0.1 for each ROI.

After preprocessing, we use 10-fold cross validation to create 10 train and test sets for model training and evaluation. We created two different versions of the datasets. The first one had no time constraint among consecutive acquisitions and we accepted subject with more than one samples, whereas the second one accepted only subjects with more than 5 samples and equal time intervals approximated to one year. From both datasets, we excluded subjects that remained CN the whole course of the study, leading up to 1211 subjects from both ADNI and BLSA studies. In Section 4 we elaborate how the two versions of datasets were used.

Regarding the clinical modality, we used the APOE4 allele and the baseline diagnosis, which are categorical variables. For the genomic modality we used SNPs. The SNPs data are 54 most AD-related SNPs pre-selected from Genome-Wide Association Studies (GWAS) based on existing literature with categorical values of $\{0, 1, 2\}$. We perform the data imputation using the information of nearest date available in the past.

4. Methodology

Central to our methodological development is the idea that each biomarker is described with a temporal function $\mathbf{F}$. The time of the first acquisition is considered the reference time, and the temporal variable is the time difference between any future acquisition and the baseline. The function $\mathbf{F}$ takes as input the structural image of the brain (structural MR) at the baseline along with the clinical variable (in our case SPARE-AD score or Adas-Cog-13), and the time-shift, calculated in months, which we want to extrapolate. If we symbolize the temporal function that describes the progression of a biomarker as $\mathbf{F}$, and a subject j has 5 acquisitions in total, then we are having 5 samples from this function $\mathbf{F}$, namely:

$$F_0, F_1, \ldots, F_4$$

where F_0 is the value of the biomarker at time zero, F_1 is the value at time t_1 and so on.

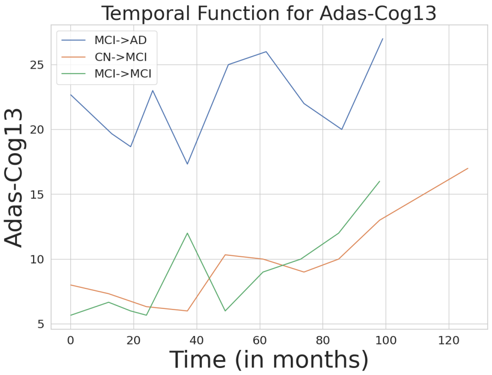


Figure 1: Example of temporal functions for Adas-Cog13 for three different progression paths

In Figure 11, we see an example of a temporal function for Adas-Cog13 for three different subjects with three different progression trends. We notice that subjects that convert to AD have higher Adas-Cog13, whereas subjects that remain MCI have lower cognitive scores. Moreover, all the temporal functions are characterized by an increasing trend. Regardless the increasing trend, we observe several fluctuations in the cognitive score. These indicate that Adas-Cog13 is a noisy variable which is also confounded by several factors such as the intelligence, educational level or even behavioral ones. Each subject i in our dataset is characterized by longitudinal features and longitudinal biomarkers respectively:

$$X_i = [x_i(0), x_i(1), ..., x_i(T_i)]$$

$$Y_i = [Y(i, 0), Y(i, 1), ..., Y(i, T(i))]$$

The longitudinal features X_i can be imaging, clinical and genomic. The T_i is the last acquisition of the subject i and the number of acquisitions can vary per subject. The data $x_i = [m_{1i}, m_{2i}, m_{3i}]$ contain the modalities m_i that can be imaging, clinical and genomics data. In our dataset, the genomic information is SNPs. It is important to highlight that SNP data, are not changing in time but we still need them to condition the potential different biomarker progression patterns on the subject's genomic signature.

We assume that there is a temporal function that describes the progression of a biomarker $\mathbf{B}$ in the population. We symbolize this function as $F_{\mathbf{B}}$. This function takes input the baseline biomarker metric, and time (t) and outputs the value of biomarker at time t. The baseline value is given by the $F_{\mathbf{B}}(t = 0)$ We will use Gaussian Processes to learn an approximation of the progression function $F_{\mathbf{B}}$, out of a limited number of samples that we have. We denote the measurements of the samples as y.

$$y_t = f(x,t) + \epsilon \ \text{ where } \epsilon \sim N(0, \sigma^2) \qquad (1)$$

If we symbolize the input space Z_t as the following tuple:

$$Z_t = (x_0, Y_0, t)$$

then the GP will learn the following mapping:

$$Z_t \longrightarrow Y_t$$

We split our subjects into population and test subjects. The population set $D = \{(Z_t^j, Y_t^j) | j = 1, \cdots L\}$ is used for training the proposed GP models, where L denotes the number of observations. The test subjects are used for evaluation purposes. The split into population-test subjects occurs with mentioned the 10-fold cross validation.

4.1. Temporal Gaussian Processes Model

The problem that we tackle falls into the class of supervised learning and specifically regression. With Gaussian Processes we define a prior distribution over the temporal functions and after inference using the subjects of the population, we have the posterior distribution of the temporal function. We gradually build our model by defining, at first, the Temporal GP model (TempGP). To this end, we are using a single-task Gaussian Process with an RBF kernel as covariance function and a constant mean function. The model is trained with Exact Inference using GPytorch (Gardner et al., 2018).

4.2. Temporal Deep Kernel Gaussian Process Model

In this part, we leverage Deep Kernel Learning so as to be able to capture the population correlations using a deep neural network. Using the deep kernel, we perform a synthesis between the RBF kernel that is used in the GP Layer and the deep network. We assume the deep kernel is going to increase the expressivity of the model, which can lead to a better representation of the population, especially when we use multiple modalities as input. Similarly, the covariance in the single-task GP with deep kernel is RBF. The model is trained with the log marginal likelihood of the GP. If we symbolize the DNN as g then the deep kernel is:

$$K_{input} = K(g(x), g(x'))$$

$$\frac{dL}{d\theta} = \frac{dL}{dK}\frac{dK}{d\theta}$$

$$\frac{dL}{dw} = \frac{dL}{dK}\frac{dK}{dg}\frac{dg}{dw}$$

where $\theta = \{l^2, \sigma^2\}$ are the hyperparameters of the RBF kernel $K(x, x') = \sigma^2 e^{\frac{||x-x'||^2}{l^2}}$

5. Experiments and Results

In this section, we will study extensively the performance of our proposed method in comparison with baseline algorithms and the pGP Experts (pGPE) presented in (Utsumi et al., 2019). Due to the different formulation and dataset requirements in our comparison, we curate two different versions of the initial dataset (ADNI and BLSA). One is for the baseline comparison and the other is for the comparison with the pGPE. The pGPE requires equal time intervals among acquisitions and subjects with more than 5 longitudinal samples. These constraints shrink the initial dataset. In Table 1 we see the number of subjects that we have in the two versions for the experiments for the Adas-Cog13. We are gradually building our model starting from Temporal GP to Temporal DKGP and thus, we decide to include Temporal GP in the experiments as well, with the intention to highlight the effects of the Deep Kernel. Regarding the training details, for both Temporal GP and Temporal DKGP we used Adam optimizer with different learning rates for Adas-Cog13 and SPARE-AD. The optimal learning rate was found through hyperparameter search in a hold-out validation set. The models were all trained for maximum 300 epochs. In the Temporal DKGP model we used early stopping with patience of 20. If the validation loss was not improved for 20 epochs then the model was no further trained and the best model was returned. The detailed architecture for the Deep Kernel is presented in Table 5. All the models were built with the GPytorch framework (Gardner et al., 2018).

Table 1: Number of subjects for Adas-Cog13

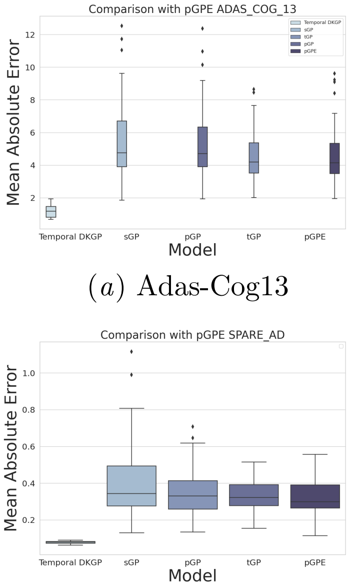
Dataset	Subjects
Dataset 1	943
Dataset 2	92

5.1. Evaluation Metrics

Our evaluation strategy comprises of two schemes. First, we examine the performance of the model as a predictor. We use the Mean Absolute Error (MAE) between the groundtruth biomarker and the prediction. Secondly, we evaluate the uncertainty estimation between the Temporal GP, the Temporal Deep Kernel GP and the pGPE, which is measured by the interval and the coverage. Interval is the upper and lower bound of the predictions and coverage indicates whether the groundtruth is within the interval. A model provides satisfactory uncertainty quantification when the interval is narrow and the coverage is high. Full coverage does not indicate good uncertainty quantification if the intervals are wide. We need both metrics to evaluate the quality of the uncertainty. In all the experiments below, 10-fold cross validation is applied in subject level, which indicates a leave-subjects-out validation scheme.

5.2. Baseline Methods

Our initial experiment is to compare our method, the Temporal DKGP and the Temporal GP with some baseline algorithms. We select the Ordinary Least Squares (OLS), the Ridge and the Elastic regression as three baselines. Additionally, we developed a Long short-term memory (LSTM) architecture and a Temporal Convolutional, composed of one-dimensional convolutions. However, both deep learning models failed to reach performance close to OLS, Elastic, and our models', after performing a thorough hyperparameter tuning. These models were trained with fixed intervals, and thus the available data was limited and not enough to train the deep networks. Additionally, LSTM is unstable in training, leading to no replicable results. Thus, we decided not to include them in the plots. Furthermore, we evaluate the uncertainty quantification only between the Temporal GP and the Temporal DKGP since we have not defined a frequentist notion of uncertainty in point estimate models. In Figure **??** we see the MAE for both the Adas-Cog13 and the SPARE-AD to be the best in the case of Temporal DKGP. Also in Table 2, the first two rows provide the uncertainty quantification for Temporal GP and Temporal DKGP. We observe that Temporal DKGP manages to provide a satisfactory coverage with narrower intervals.



(a) Adas-Cog13

(b) SPARE-AD

Figure 2: Comparison with pGPE models and TempDKGP in MAE

Table 2: Interval and Coverage for Experiments for Adas-Cog13

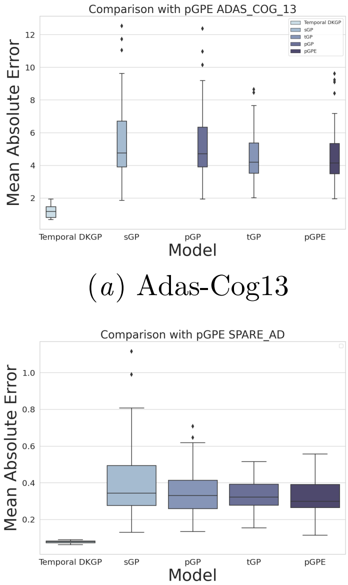
Model	Interval	Coverage
TempGP	9.215 ± 0.0040	0.724
TempDKGP	$\mathbf{7.173 \pm 0.0032}$	0.650
TempDKGP	$\mathbf{3.68 \pm 0.0005}$	0.820
pGPE	36.98 ± 30.0000	1.000

5.3. Comparison with pGP Experts

In this part, we compare the Temporal DKGP model with the pGPE. Utsumi et al. (2019) present four different GP models. We follow the notation of their paper and we symbolize as sGP the population model, the pGP, the tGP and the pGPE as the first, the second and the third version of their personalized GPs. We compare these models with the Temporal DKGP. In order to have a fair comparison we reformulated the dataset (Dataset 2). Both models are trained with the same train subjects and tested with the same test subjects. The Temporal DKGP model is now trained with longitudinal samples that have equal time intervals, but no constraint in the number of samples. On the contrary, the pGPE, due to the multi-output approach that uses, requires subjects with more than 5 acquisitions. In Table 1, we can see the number of subjects and samples used in this experiment for Adas-Cog13. In Figure 3 we see that Temporal DKGP manages to provide more accurate predictions in comparison with sGP, pGP, tGP and pGPE models. It is important to note that the task of extrapolation to equal time intervals and thus specific time-shifts is easier because the model does not learn not from random time-shifts, but from constrained ones. This is the reason we see lower MAE in Adas-Cog13 and SPARE-AD in the Temporal DKGP in comparison with the MAE in the Baseline Methods.

5.4. Ablation Studies

We perform an ablation study on the different input modalities that we use to extrapolate the metrics. We have three different modalities. The first is the structural MRI , the second is the clinical, including the baseline diagnosis and the APOE4 allele. The third modality is the genomic one, the 54 SNPs that are related to AD. We examine



(*a*) Adas-Cog13

(*b*) SPARE-AD

Figure 3: Comparison with pGPE models and TempDKGP in MAE

the performance of the Temporal DKGP in the different modalities in comparison with the Temporal GP and the pGPE, with the goal to showcase the effectiveness of the Deep Kernel when we integrate different modalities in the input. In Figure 4 we see that Temporal DKGP achieves the lowest MAE in comparison with the rest GP methods, the Temporal GP and the pGPE.

5.5. Analysis

5.5.1. Uncertainty as the time-shift increases

We examine how the uncertainty quantification changes as the time-shift increases. It is important to be able to estimate how the model performs if we increase the time-shift to up to 10 years ahead and further. In the boxplots below we see the MAE and the interval for Adas-Cog13. For Adas-Cog13, the error is maintained low and is in most cases comparable with the Temporal GP. The interesting difference that we observe between the two models is that the Temporal DKGP achieves to maintain the intervals narrow while keeping low the error. In the Appendix D, we attach the same plots for SPARE-AD, where we observe the same behavior.

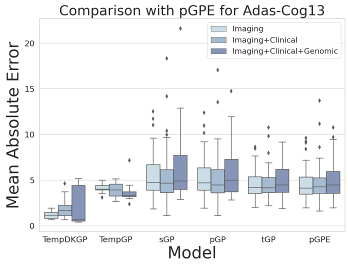
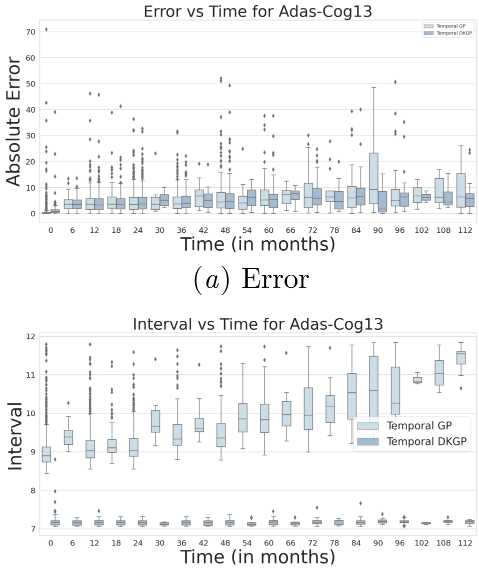


Figure 4: Comparison between GP and Deep Kernel Learning for Adas-Cog13

6. Conclusion and Future Work

6.1. Conclusion

In this paper we present a simple approach for extrapolating biomarkers related to Alzheimer's disease, the SPARE-AD and the Adas-Cog13, while at the same time quantifying the uncertainty of the predictions. We extensively show that Temporal DKGP successfully predicts both clinical variables by providing a generative model, which based on the time, can extrapolate the clinical variable at the given timepoint. Our method does not impose any time constraints or a specific curation of the longitudinal data according to acquisition time. Thus, it can be applied in any clinical variable that progresses over time in a longitudinal study. Importantly, we have shown the potential of deep kernel as a drop-in replacement for the kernel function in conventional



(a) Error

(b) Interval

Figure 5: Error and Interval for increasing time-shifts for Adas-Cog13

GP formulations. Our results prove the hypothesis that deep kernel learning provides with expressivity and can learn the data representation well. Since GPs have the issue of scalability to high-dimensional data, we have demonstrated that the deep kernel formulation can pave the way for more real-world applications for traditional GP methods. Our work confirms the feasibility of the proposed model through extensive experiments on two types of neuroimaging biomarker/cognitive score with different scales across two cohorts.

6.2. Future Work

For our future work, we will examine if we could use a different encoding of the time in the input of the Temporal DKGP. Also, experimenting with the structure of Deep Kernel and perform an additional analysis on how the model is conditioned based on some clinical variables, such as the clinical status, the APOE4 allele will help us interpret the behavior of the model. Additionally, we want to explore multitask GPs for the extrapolation of multiple clinical variables or structural features in time. Such an approach can be extended to parse the disease heterogeneity by clustering the subjects, while at the same time performing extrapolation, based on the inter-task similarities on the extrapolated longitudinal trajectories captured by the multitask GP.

References

Christos Davatzikos, Feng Xu, Yang An, Yong Fan, and Susan M Resnick. Longitudinal progression of alzheimer's-like patterns of atrophy in normal older adults: the spare-ad index. *Brain*, 132(8):2026–2035, 2009.

Jimit Doshi, Guray Erus, Yangming Ou, Bilwaj Gaonkar, and Christos Davatzikos. Multi-atlas skull-stripping. *Academic Radiology*, 20:1566–1576, 12 2013. ISSN 10766332. doi: 10.1016/j.acra.2013.09.010.

Jimit Doshi, Guray Erus, Yangming Ou, Susan M. Resnick, Ruben C. Gur, Raquel E. Gur, Theodore D. Satterthwaite, Susan Furth, and Christos Davatzikos. Muse: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage*, 127:186–195, 2 2016. ISSN 10959572. doi: 10.1016/j.neuroimage.2015.11.073.

Guray Erus, Jimit Doshi, Yang An, Dimitris Verganelakis, Susan M. Resnick, and Christos Davatzikos. Longitudinally and inter-site consistent multi-atlas based parcellation of brain anatomy using harmonized atlases. *NeuroImage*, 166:71–78, 2 2018. ISSN 10959572. doi: 10.1016/j.neuroimage.2017.10.026.

Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. *CoRR*, abs/1809.11165, 2018. URL http://arxiv.org/abs/1809.11165.

Giovana Gavidia-Bovadilla, Samir Kanaan-Izquierdo, María Mataró-Serrat, Alexandre Perera-Lluna, and for the Alzheimer's Disease Neuroimaging Initiative. Early prediction of alzheimer's disease using null longitudinal model-based classifiers. *PLOS ONE*, 12(1):1–19, 01 2017. doi: 10.1371/journal.pone.0168011. URL https://doi.org/10.1371/journal.pone.0168011.

R. Guerrero, A. Schmidt-Richberg, C. Ledig, T. Tong, R. Wolz, and D. Rueckert. Instantiated mixed effects modeling of alzheimer's disease markers. *NeuroImage*, 142:113–125, 2016. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2016.06.049. URL https://www.sciencedirect.com/science/article/pii/S1053811916302981.

R C Petersen, P S Aisen, L A Beckett, M C Donohue, A C Gamst, D J Harvey, C R Jack, W J Jagust, L M Shaw, A W Toga, J Q Trojanowski, and M W Weiner. Alzheimer's disease neuroimaging initiative (adni) clinical characterization, 2010. URL www.neurology.org.

Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M. Nasrallah, Theodore D. Satterthwaite, Yong Fan, Lenore J. Launer, Colin L. Masters, Paul Maruff, Chuanjun Zhuo, Henry Völzke, Sterling C. Johnson, Jurgen Fripp, Nikolaos Koutsouleris, Daniel H. Wolf, Raquel Gur, Ruben Gur, John Morris, Marilyn S. Albert, Hans J. Grabe, Susan M. Resnick, R. Nick Bryan, David A. Wolk, Russell T. Shinohara, Haochang Shou, and Christos Davatzikos. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208, 3 2020. ISSN 10959572. doi: 10.1016/j.neuroimage.2019.116450.

Saima Rathore, Mohamad Habes, Muhammad Aksam Iftikhar, Amanda Shacklett, and Christos Davatzikos. A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages. *NeuroImage*, 155:530–548, 2017.

Susan M Resnick, Dzung L Pham, Michael A Kraut, Alan B Zonderman, and Christos Davatzikos. Longitudinal magnetic resonance imaging studies of older adults: A shrinking brain, 2003.

Alexander Schmidt-Richberg, Christian Ledig, Ricardo Guerrero, Helena Molina-Abril, Alejandro Frangi, Daniel Rueckert, and on behalf of the Alzheimer's Disease Neuroimaging Initiative. Learning biomarker models for progression estimation of alzheimer's disease. *PLOS ONE*, 11(4):1–27, 04 2016. doi: 10.1371/journal.pone.0153040. URL https://doi.org/10.1371/journal.pone.0153040.

John G. Sled, Alex P. Zijdenbos, and Alan C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, 17:87–97, 1998. ISSN 02780062. doi: 10.1109/42.668698.

Yuria Utsumi, Ricardo Guerrero, Kelly Peterson, Daniel Rueckert, Rosalind W Picard, et al. Meta-weighted gaussian process experts for personalized forecasting of ad cognitive changes. In *Machine learning* *for healthcare conference*, pages 181–196. PMLR, 2019.

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, John C. Morris, Ronald C. Petersen, Andrew J. Saykin, Leslie M. Shaw, Arthur W. Toga, and John Q. Trojanowski. Recent publications from the alzheimer's disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials, 4 2017. ISSN 15525279.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

Appendix A. Datasets

In this section, we appose some details about the datasets. The Dataset 1 (D1) is the version of the temporal data that was used for the comparison with the baseline models. The Dataset 2 (D2) is the version of the temporal data that was used to train the Temporal GP, and Temporal DKGP for comparison with the pGP Experts. In D2, we set temporal constraints of equal time intervals among consecutive acquisitions to be one year. The Dataset 3 (D3) is the version of the data that was used to train the pGP Experts. Again, we imposed time intervals to be one year and accepted subjects with more than 5 samples to be able to construct the multitask dataset.

Appendix B. Deep Kernel Model

The deep neural network architecture of our Deep Kernel is composed of 4 linear layers along with ReLU non-linearities in-between. Table 5 contains the detailed structure of the Deep Kernel.

Table 3: Number of Subjects used in our experiments for Adas-Cog13

Dataset	**Subjects**
Img (D1)	943
Img+Cl (D1)	943
ADNI Img+Cl+G (D1)	876
Img (D2)	907
Img+Cl (D2)	907
ADNI Img+Cl+G (D2)	876
Img (D3)	92
Img+Cl (D3)	92
ADNI Img+Cl+G (D3)	91

Table 4: Number of Subjects used in our experiments for SPARE-AD

Dataset	**Subjects**
Img (D1)	1175
Img+Cl (D1)	1175
ADNI Img+Cl+G (D1)	1105
Img (D2)	1162
Img+Cl (D2)	1162
ADNI Img+Cl+G (D2)	1093
Img (D3)	454
Img+Cl (D3)	454
ADNI Img+Cl+G (D3)	457

Appendix C. Ablation Study on Modalities

We present the ablation study on the Dataset 1 when comparing Temporal DKGP with Temporal GP. Both models, in this comparison, are trained and tested with Dataset 1, namely not time and sample constraints.

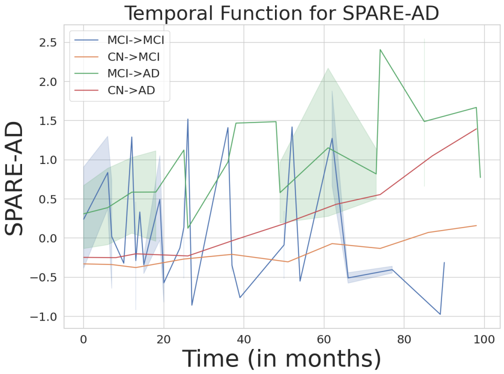


Figure 6: Example of a temporal function for SPARE-AD for four different progression paths

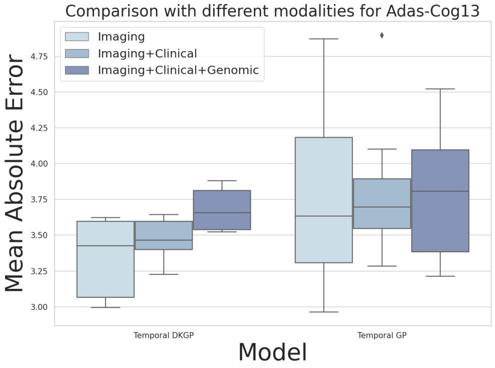


Figure 7: Ablation Study for Modalities: Adas-Cog13

Table 5: Deep Kernel

Deep Kernel
Linear(dim, 1000)
ReLU
Linear(1000,500)
ReLU
Linear(500,50)
ReLU
Linear(50,25)
Dropout

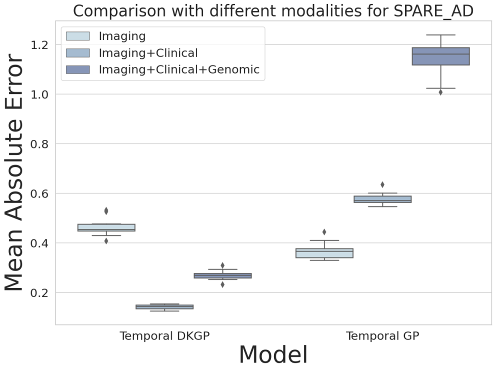


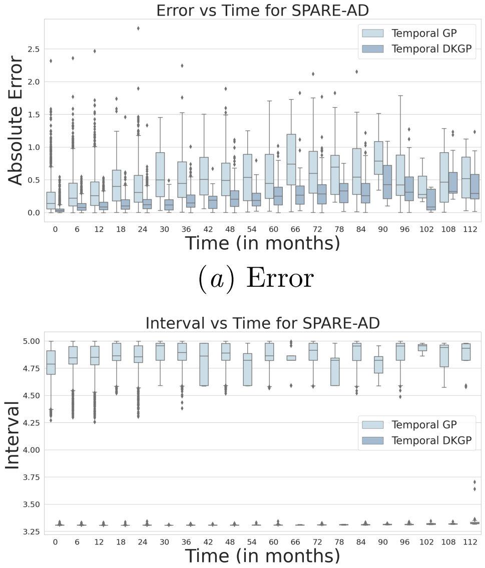
Figure 8: Ablation Study for Modalities: SPARE-AD

Appendix D. Analysis

D.1. Uncertainty

We attach the uncertainty-time analysis for the SPARE-AD.

D.2. Prediction Examples



(*a*) Error

(*b*) Interval

Figure 9: Error and Interval for increasing time-shifts for SPARE-AD

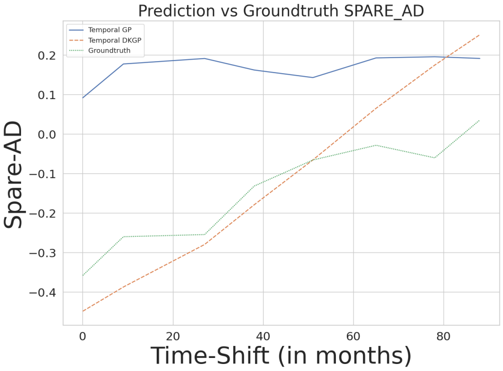


Figure 11: Example of a temporal functions learned by the Temporal models

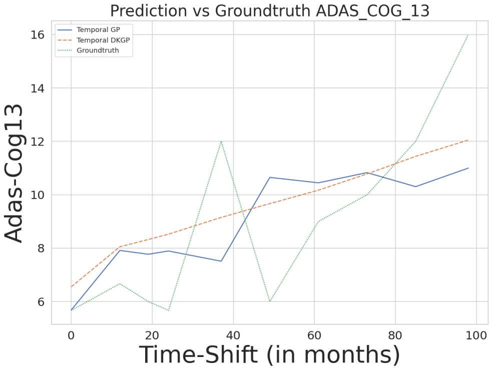


Figure 10: Example of a temporal functions learned by the Temporal models