Dissertation Prospectus

1. **The research problem**

My research focus lies primarily in the intersection of the field of machine learning and Alzheimer’s disease assessment. Using routinely collected clinical data, including magnetic resonance imaging scans, neuropsychological test results, functional assessment for daily activities etc., we investigated whether machine learning-based diagnostic systems can reach expert-level performance in a typical memory clinic setting. If so, these fully automatic machine learning-based diagnostic tools can be used to alleviate the issue of growing scarcity of clinical dementia specialists as the population ages.

To answer this central question, we need to get a comprehensive understanding of several questions regarding how experts like neurologists diagnose individuals who visits the neurologic clinics, for example, what clinical measures or tests they need to make a diagnosis with confidence, what are the common diagnostic outcomes, what are the common decision procedure to derive the final diagnostic conclusion. By working closely with neurologists and MD students, we as a team defined a set of clinically relevant tasks for the model to predict so that the machine learning-based system can be trained to operate in scenarios that mimic the real clinics to some extent.

Starting from training a model to conduct binary classification between the subjects with Alzheimer’s disease and the healthy control group, we expanded the diagnostic spectrum to include firstly different stages of cognitive decline, including the initial cognitive normal stage, the mild cognitive impairment stage, and the dementia stage. With these 3 stages of cognitive decline, we covered the full spectrum of cognitive ability in the longitudinal axis. Since there are various etiologies of dementia, we further classified whether the dementia is caused by Alzheimer’s disease or other etiologies. In the future, as we collect more data for each other etiologies, for instance, Parkinsonian dementia, Lewy Body dementia, Vascular dementia, and Frontotemporal lobe dementia, we might be able to diagnose each etiology independently. For this stage, Alzheimer’s disease is still our central focus, and we grouped all other etiologies as non-Alzheimer’s dementia given the situation of only having access to a finite amount of non-Alzheimer’s dementia cases.

With the prediction tasks being defined, we systematically exhausted the collection of datasets, to support the scope of diagnosis that we are aiming for, from multiple resources including publicly available databases, and in-house data from internal collaborators. Some datasets focus on a single type of dementia, while other more generic data centers collect cases from broad types of dementia. The extreme heterogeneity of what information was collected in which format across cohorts added extra challenges for systematically organizing, cleaning, and harmonizing the data. The complexity also introduced the risk of neglecting confounding factors that might unintentionally bias the model’s predictions. We collected and integrated multimodal data, including medical imaging scans, memory tests, functional assessment and many other measures as typically requested by neurologists, from multiple resources to feed machine learning models to predict the relevant tasks.

From the modeling perspective, deep learning models are usually considered as “black-box” models since they merely produce the final predictions from inputs in an end-to-end fashion. When making high-stake decisions like medical diagnosis, these “black-box” models can barely earn any trust from the medical community, and it also makes the diagnosis irresponsible due to the mysterious and untraceable decision process. I thus dedicated myself to the design, development, and application of interpretable deep learning methods throughout my PhD study. Instead of only measuring a model's performance based on accuracy, it is often beneficial to know how the model attributes its final prediction to each feature, and how the model quantifies the severity degree within the medical imaging scans in finer spatial granularity.

These interpretable results open many opportunities for us to understand how the model relates input values to output predictions, and thus allow us to validate the model’s underlying decision logics against rich domain knowledge from medical professionals. For us, the target task that the model was trained to predict is not a simple numerical label anymore. Instead, starting from how the disease was defined, we have dived deep into the pathological level and grounded our deep learning framework using rigorous cross-validation between the model’s interpretable results and neuropathological evidence. I believe that these carefully designed clinical-level validation helped relieve, to some extent, the uncertainty and distrust of AI prediction in the medical community.

**2. Method and techniques**

2.1 Convolutional neural network

Convolutional neural network is one of the core components of the modeling part of my PhD research that is commonly applied to analyze visual imagery. The information contained in an image, for instance a 3D magnetic resonance imaging scan of the brain, exists in a high-dimensional space characterized by all pixel values of the image. It thus became necessary to develop an efficient method to extract the abstract information that is relevant to the target task. For instance, if the target task is a binary classification between Alzheimer’s disease and healthy control, the convolutional neural network can be considered as a function that maps the original high-dimensional pixel space to a single scalar number that represents the probability that the subject might have Alzheimer’s disease.

Convolutional neural networks use convolutional operators to perceive the content of an image. The convolutional operator is parametrized by multiple kernel matrices each with a typical shape of 3 by 3 or 5 by 5, and the number of kernels defines the “width” of the convolutional layer. Inside a convolutional layer, the algorithm will traverse the kernel matrices systematically over the whole image line by line, and store as output the inner product between the kernel and the image patch of the same size at each location. Wherever the local patches are similar to the kernel matrix, the inner product yields a strong response at the corresponding location. Due to the existence of highly complicated visual signals, allowing multiple independent kernel matrices to learn various spatial features within a layer is beneficial for increasing the capacity of the meaningful patterns that the network can learn.

Similar to the vanilla artificial neural network, there are often used non-linear activation operations interleaving with the linear convolution operation to empower the model with a more versatile and expressive function space that the model can approximate. The parameters in the kernel matrices are not determined pre-hand by experts or programmers, instead, the model was trained to learn those parameters through the process of optimizing a certain target task as discussed in the section below.

2.2 Model training with stochastic gradient descent

Depending on the task that the model is predicting, we can define a loss function whose value characterizes the empirical performance of the model based on a group of data. For example, when conducting a classification task, the cross-entropy function between the predicted probabilities and the ground truth label provides a good numerical approximation of the agreement between the predictions and the truths.

With the loss function being defined, the model can be trained towards the direction that the loss function is descending. The stochastic gradient descent algorithm provides an effective and efficient strategy of training broad types of neural networks by consecutively applying a two-steps process. The first step is to feed forward a random batch of inputs and estimate the gradient of the loss function with respect to all parameters. The second step can be conducted to adjust all the parameters by a small step along the direction of the gradient.

Other than the convolutional neural network, we also trained other types of machine learning models, including decision tree models, random forest models, boosting tree-based models etc., that are suitable to handle non-imaging features by following the same training algorithm.

2.3 Interpretability

For a disease diagnostic model that is trained on medical imaging data, we are not satisfied by merely producing a predicted disease probability; instead, the information of what specific areas within the image are raising concerns and contributing to the final predictions is more valuable and informatic. One of the common approaches that has been used to generate the saliency map is by mutating a single area of the image, for instance, masking out the pixels, and to observe how the model’s prediction is changed accordingly. Another state of the art interpretable framework is inspired by the classical Shapley value which was originally designed to attribute fair value to each worker as they work together towards the same goal. The Shapley value can be derived by generating all subsets of workers and then separating the whole subsets into one group where the target person is included and the other group with the person being absent. By comparing the difference of the business metric between the two groups, we are able to draw a fair contribution from the person. In the context of developing machine learning models, input features can be considered as the workers, and the model’s output can be understood as the overall gain obtained with a coalition of input features.

**3. Outline of dissertation**

1. Introduction [This chapter introduces the basic concepts of machine learning and the contextual information about Alzheimer’s disease]

1.1 Alzheimer’s disease and dementia [Introducing the major challenges of diagnosing Alzheimer’s disease and dementia and the motivations of my PhD study]

1.2 Machine learning and deep learning

1.3 Interpretable machine learning

2. Machine Learning and Deep Learning [This chapter contains a more concrete introduction of machine learning and detailed descriptions of the techniques and methods used throughout my PhD study]

2.1 Traditional machine learning models

2.2 Artificial neural network

2.3 Convolutional neural network

2.4 Optimization and gradient backpropagation

2.5 Model training and evaluation

3. Interpretable Machine Learning in Healthcare [This chapter serves as a literature review on a special topic of machine learning called interpretability and its medical applications]

3.1 Introduction [Limitations of the “black-box” deep learning models and why interpretability is a critical component for medical decisions]

3.2 Inherent vs. post-hoc interpretability [Two common types of interpretable machine learning models]

3.3 Saliency approach for medical imaging [The application of interpretable machine learning on medical imaging analysis]

3.4 Feature importance and prediction attribution [The application of interpreting the importance of non-imaging features]

4. Interpretable Deep Learning for Alzheimer’s Disease [The chapter presents the novel interpretable saliency approach that we developed to understand Alzheimer’s disease-relevant brain atrophy patterns]

4.1 Introduction [Introducing the deep learning-based diagnostic methods for Alzheimer’s disease]

4.2 Dataset [Describing the data used for this study]

4.2.1 Dataset collection and feature selection

4.2.2 Data harmonization [The data cleaning steps that we conducted to handle the heterogeneity across cohorts]

4.3 Model development [Describing the interpretable saliency method that I developed]

4.3.1 Existing saliency methods [Discussing the limitations of the previous works]

4.3.2 Our saliency approach [Discussing the advantages of our method and how our method addressed those limitations]

4.3.3 Disease probability map (DPM) [Presenting the primary results on the interpretable outcome based on MRI]

4.3.4 Multi-modal data integration [Presenting the results on the overall diagnostic performance by integrating MRI-derived information with non-imaging features]

4.4 Validation of the DPM [Presenting the multi-disciplinary collaborations within our team towards validating the interpretable outcomes]

4.4.1 Local accuracy of the DPM [Measured the local accuracy of the DPM according to the clinical diagnostic results]

4.4.2 Neuropathological validation of the DPM [Compared the DPM with neuropathological evidence]

4.4.3 Global accuracy of the DPM [Compared the model’s performance on the individual-level classification with a group of neurologists]

4.5 Conclusion

5. Expert-level Deep Learning for Dementia Assessment [As an extension of the chapter 4, we expanded the diagnostic scope from Alzheimer’s disease to dementia diagnosis to make it broadly applicable]

5.1 Introduction [Introducing the complexity of real-world neurological diagnoses]

5.2 Dataset [To support the target diagnostic scope, we collected data from 8 independent cohorts]

5.2.1 Dataset collection and feature selection

5.2.2 Harmonization of MRI scans

5.2.3 Harmonization of non-imaging features

5.2.4 Confounding factors

5.3 Formulation of the prediction tasks [Defining the clinically relevant prediction tasks under the guidance of neurologists]

5.3.1 Neurologists’ decision process

5.3.2 Multi-task deep learning setting

5.4 Model development [Describing the details about the model development, including model architectures and model training]

5.4.1 Model architecture

5.4.2 Data split and model training

5.5 SHAP-based interpretability analysis [Presenting the major results on the SHAP-based interpretable analysis of both the MRI and the non-imaging data modalities]

5.5.1 Voxel-level interpretable saliency map

5.5.2 Feature importance ranking for dementia diagnosis

5.5.3 Data-driven insights on Alzheimer’s disease

5.6 Clinical-level validation [Validating the model’s predictions and the post-hoc SHAP-based interpretable outcomes using domain knowledge]

5.6.1 Radiologists’ atrophy rating and diagnosis

5.6.2 Neurologists’ diagnosis

5.6.3 Neuropathological validation

5.7 Conclusion

6. Knowledge Distillation for Dementia Diagnosis [This chapter presents a special training strategy of deep neural networks that support knowledge transferring from some models to other models]

6.1 Introduction [Introducing the basic concepts of knowledge distillation]

6.2 Existing methods for knowledge distillation [Presenting the standard ways of conducting knowledge distillation and recent advances in this direction]

6.3 Real-world applications [Discussing how knowledge distillation can be a powerful tool in real-world medical scenarios]

6.4 Decentralized knowledge sharing [Presenting a novel way of creating a decentralized knowledge distillation setting among many models]

6.5 Conclusion

7. Conclusion