Dissertation Prospectus

1. **Research problem**

Alzheimer’s disease (AD) is an irreversible progressive disease that causes memory impairment and decline of other important mental functions. AD is the major cause of dementia and there are nearly 10 million new cases of dementia annually worldwide. Tremendous efforts have been made globally to alleviate the burden of AD, including the development of biomarkers for better diagnosis and drug discoveries to slow down the AD progression. However, there is no cure for AD yet and AD as a growing epidemic is projected to have an even larger social impact.

Due to the rapidly growing burden of AD and the scarcity of clinical specialists, new methods and tools are urgently needed to keep pace with the increasing demand of timely AD diagnosis. New biomarker-based diagnostic tests using blood or cerebrospinal fluid have been investigated but the availability of these tests remain limited in research contexts. It is thus still critical to figure out less invasive and broadly available options to diagnose AD.

The convergence of digitization of medical records and rapid advancements in artificial intelligence (AI) opens many opportunities to apply AI in medicine. A rapidly increasing number of research teams started applying machine learning or deep learning models to solve medical problems as public datasets and machine learning libraries became readily available. However, there remains a large gap between these massively produced research papers and the systems that have been successfully translated to clinical practice. One of the major missing pieces from these rudimentary projects is the lack of clinical relevance which is the core value that our team is aiming to achieve.

My research focuses on the development and validation of deep learning diagnostic systems for AD and dementia. To reach the goal of clinical relevance, we involved medical professionals through different stages of my research, i.e., starting from defining clinically meaningful tasks for the model to solve, to deciding on what input information to include based on practical considerations, then to the validation stage of the model under a holistic view of the disease beyond simple labels. In addition, we aim to deliver rigorous machine learning solutions that generalize well on unseen data.

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 scenario that mimics memory clinics. 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, which in combination covers the full spectrum of cognitive status. We further classified whether the dementia is caused by AD or other etiologies in the following task if a subject was predicted as demented. The diagnostic scope spanned by these two tasks supports the diagnosis of AD under most of the cases in neurologic clinics.

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 contain cases with broad etiologies of dementia. The heterogeneity of data storage and availability across cohorts added extra challenges for organizing, cleaning, and harmonizing the data, and also introduced the risk of neglecting confounding factors that might unintentionally bias the model’s predictions. We thus spent a tremendous amount of effort on data processing and ruling out confounding factors. Reserving multiple independent data cohorts to test the model plays a key role of justifying the generalizability of the model when facing new data.

From the modeling perspective, deep learning models are usually considered as “black-box” models since they produce the final predictions from inputs in an end-to-end fashion. When making high-stake decisions like medical diagnoses, 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 and development of interpretable deep learning methods. Interpretable results derived from the model open the door of debugging the system, validating the model against domain knowledge, and generating better understanding of the disease. Different from simply evaluating models’ performance in comparison with the predicted label, we grounded the model using clinical level knowledge, for example, from neuropathological examinations.

In summary, I have been conducting my PhD research under the motivation of delivering clinically relevant machine learning systems to diagnose AD. We as a multi-disciplinary team have been aiming at a high standard of earning trust from the medical community through efforts towards involving medical professionals on problem formulations, developing easy-to-interpretable machine learning models, testing the model’s generalizability over many independent cohorts, and grounding the model with evidence at the clinical level.

**2. Method and techniques**

2.1 Convolutional neural network

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

CNN uses convolutional operators to perceive the content of an image which is parametrized by multiple kernel matrices. Inside a convolutional layer, the algorithm traverses the kernel matrices systematically over the whole image line by line, and store as output the inner product between the kernel and the image patches. Wherever the local patches present similar pattern to the kernel, a large response is yielded at the corresponding location. Allowing a bandwidth of multiple independent kernel matrices equips the model with the adequate capacity of learning various spatial features.

Like the vanilla artificial neural network, there are non-linear activations interleaving with linear convolution operations to construct a more expressive function space in which the model exists. The parameters in the kernel matrices are not determined pre-hand by experts or programmers, instead, the model was trained to learn those parameters directly from the data by optimizing a certain target function 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. For example, when conducting a classification task, the cross-entropy function between the predicted probabilities and the ground truth label provides a good numerical proxy of the classification performance.

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 CNN, 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 probability; instead, it is more informative to show the specific areas within the image that raised concerns. One of the common approaches that has been used to characterize region-specific impact is by mutating a small area of the image and then observing how much the prediction changed accordingly. Another state of the art interpretable framework is inspired by the classical Shapley value which was originally designed to fairly distribute the gain from a coalition to each individual’s contribution. In the context of analyzing 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 the 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. Conclusion