Predicting Alzheimer's Progression in Mild Cognitive Impairment: Longitudinal MRI with HMMs and SVM Classifiers

Deep Himmatbhai Ajabani
Application Developer Lead, Source InfoTech Inc, Atlanta, Georgia

***Abstract*—The number of elderly people has increased due to the huge growth in human life expectancy over the past few decades. As a result, age-related illnesses and ailments have become more prevalent, including Alzheimer's Disease (AD). A notable deterioration in cognitive functions, particularly memory and thinking skills, characterizes Mild Cognitive Impairment (MCI), a condition that lies in the middle of normal aging and dementia. Therefore, MCI carries a noticeably higher chance of developing into AD and frequently serves as a prelude to dementia. However, using cutting-edge image processing and machine learning techniques, it is possible to examine and find underlying patterns in these complex diseases. By using these techniques, it is possible to separate groups, identify the causes of such separation, and create disease prediction models. Clinical trials, mostly using cross-sectional Magnetic Resonance Imaging (MRI) data, have extensively looked into the use of MRI for the early identification of AD and MCI. On the other hand, longitudinal studies follow the same subjects over an extended period, giving researchers the chance to investigate cross-sectional trends as well as the development of the disease. Three different techniques are put forth in this study for the analysis and assessment of the structural data found in longitudinal MRI scans. Without considering any other diagnostic measures, this information is used to forecast the progression of those who have been diagnosed with MCI. These techniques utilize Hidden Markov Models (HMMs), which capitalize on the advantages of Support Vector Machine (SVM) classifiers.**

Keywords*—Alzheimer's disease; image processing; Magnetic Resonance Imaging; Mild Cognitive Impairment; machine learning

I. Introduction

The extraordinary organ known as the human brain is in charge of controlling every aspect of the body, including breathing, blood circulation, digestion, and digesting. Additionally, it acts as the control center for conscious functions including thinking, memory formation, thought retrieval, and decision-making while facilitating conscious behaviors like walking, talking, and visual perception. The brain is an equally fascinating phenomenon when seen from an anatomical standpoint. It is thought that it has about 100 billion neurons and a mind-boggling 100 trillion synapses, which are the connections between neurons that allow for communication. The network of blood arteries in the brain is essential to maintaining its normal operation. Surprisingly, the brain controls an astounding 20% of the body's blood flow despite making up just around 2% of the total body weight. Around 400 billion capillaries make up this complex circulatory system, which works ceaselessly to deliver oxygen, glucose, and other nutrients necessary for the survival of brain cells. The brain's large number of neurons plays a crucial role in preserving optimal function. The long lifespan of neurons, which begins during fetal development and lasts for up to a century, makes them unique. In the extremely rare case that they perish, neurons can regenerate, highlighting the significance of routine maintenance and repair. Individual differences in these alterations' scope and timing can have a significant impact on their impact levels. A diminished ability to learn new information, problems recalling memories, and increased difficulty performing tasks that were once simple to complete are common signs of aging. Importantly, these talents are not completely restricted because cognitively healthy senior people can still carry out these tasks, albeit somewhat more slowly than their younger counterparts. With 50–80% of dementia cases being caused by Alzheimer's Disease (AD), it becomes clear that AD is the most common type of dementia. The age of diagnosis or onset of the disease has a significant impact on the life expectancy of AD patients, which ranges from three to ten years. Although structural brain atrophy, pathological amyloid deposits, and metabolic alterations in the brain are thought to be related with Chronic Traumatic Encephalopathy (CTE), it is unclear whether these elements are the disease's causes or the results of its progression. Due to its effects on memory, Mild Cognitive Impairment (MCI), also known as amnestic MCI, is seen as an early stage of AD. Even though MCI is not considered a true disease, the early stages of AD are very similar to it. Although not severe enough to interfere with their daily lives, people who struggle with MCI experience memory, language, and judgment issues that are noticeable to others and distinct from usual aging symptoms.

The increased risk of acquiring AD in the future that MCI patients face compared to people with cognitively normal brains emphasizes the importance of MCI. As a result, MCI is a topic that medical professionals are quite interested in. There are significant ambiguities in the distinctions between the three phases of cognitive health - Cognitively Normal (CN), Moderate Cognitive Impairment (MCI), and AD - and no clear-cut standards for determining an individual's stage are in place. Nevertheless, decades of study have produced a range of approaches intended to assess the condition of brain health such as [1]–[3]. Science and medicine have placed a lot of emphasis on the study of the brain and its anomalies [4].

However, there has been a barrier to non-invasively studying it for a very long time. Even today, the only time an exact diagnosis for AD can be made is post-mortem, during the autopsy, when amyloid plaques produced in the brain and other indicators of brain degeneration can be studied by a doctor. Early identification and detection of AD and MCI are crucial because they can help patients and their families get ready for illness and start treatment as soon as feasible. In exchange, this can provide patients the chance to take part in clinical trials where the most recent medicines can be used, and they can generally manage the illness better [5]. Scientists can simultaneously study the early phases of AD and MCI to understand the disease's origin, which could result in better techniques of therapy or prevention. According to estimates, MCI patients get AD at a rate of 10–15% while cognitively healthy people develop dementia at a rate of 1%–2%. With the use of Magnetic Resonance Imaging (MRI), researchers may now conduct non-invasive in vivo examinations of the human body. This means that the brain can be monitored and evaluated to establish a baseline of what it should resemble at various stages of the disease's course or even in cognitively normal brains. It is now possible to study the brain and the changes that take place because of either the normal aging process or particular disorders thanks to the combination of computer science and machine learning.

This study's goal is to examine and interpret the structural data obtained from longitudinal brain MRI images. To investigate the gradient of anatomical and morphological changes occurring in the brain as MCI progresses to AD. Despite the widespread use of MRI scans in this sector, the goal is to forecast the possible development of AD using only this information, without the addition of other biomarkers or clinical and cognitive assessments. We use a longitudinal series of MRI scans from different people who have been given different diagnoses (CN, MCI, AD) and who are transitioning to different diagnoses (CN, MCI, AD). The goal is to avoid the possibility for human judgment errors that can occur in complicated and time-consuming processes by just using the data derived from structural (volumetric) brain changes. Given that individuals often seek medical advice after the onset of symptoms and that the diagnostic process needs some time to complete, this technique enables a quicker forecast of the condition. This study also aims to evaluate and investigate the accuracy of longitudinal MRI scans in foretelling the transition from MCI to AD in this domain. The longitudinal MRI scans are viewed as a series of observations, after which Hidden Markov Models (HMMs) [6]–[10] are used for modeling. Then, either the HMMs alone or a Support Vector Machine (SVM) [11]–[13] classifier that has been trained using the data that the HMMs used to represent the data are used to make the predictions. It's crucial to understand that this study does not try to improve, extend, or implement any one predefined technique. This technique is unusual because it uses longitudinal MRI images that are obtained one year apart and uses only the structural data that was derived from those scans. As a result, a direct comparison between the performance and results of the trials carried out for this study and the most recent findings is not possible.

The paper is as follows: In Section II we will see the related works. In Section III, an empirical study has been presented consisting of dataset description and evaluation metrics. In Section IV, the proposed models have been discussed. In Section V, the experimental results and analysis have been done. In Section VI, the thought of the paper has been presented and we conclude the paper in Section VI with some conclusions and future works.

II. Related Works

A quick overview of current research on AD and longitudinal data in the domains of computer science and machine learning is provided in this section.

A. Brain

A substantial amount of research has been focused on identifying and extracting the elements from an MRI scan that are the best diagnostic predictors in order to facilitate additional diagnosis [14]–[16]. The features that are most frequently used involve the assessment of both grey and white matter volumes [17], either over the whole brain or in particular areas such the frontal, temporal, parietal, and hippocampal cortex [18]. Furthermore, cortical thickness is a common characteristic [16], as are CSF density maps [19], [20]. Manually extracting and choosing characteristics from MRI scans is a very difficult and time-consuming process. When some feature parameters change and the feature-extraction process needs to be repeated, it often leads to the possibility of inaccurate data or complexity in re-extracting features. As a result, several tools have been created, such as FreeSurfer[1], FSL[2], and SPM[3], which let scientists and medical experts handle and interpret MRI scans in different ways and accomplish accurate feature extraction. With little to no human oversight, these technologies can carry out extraction and selection tasks.

B. Alzheimer Disease

Considerable advancements have also been achieved in the effort to control and make use of these attributes. Using different classification techniques, it is possible to distinguish between a brain that is cognitively normal and one that is impaired, either by AD or MCI. MRIs and f-MRIs are commonly used in this type of research to get anatomical and physiological brain features, which are then used to determine any pathological or normal changes occurring in the brain. These include the assessment of cortical thickness and the density of cerebrospinal fluid, as previously indicated, as well as volumetric measures of several brain areas, such as the cingulate cortex, hippocampus, and parahippocampal gyrus. These assessments make it easier to identify anomalies in the brain and offer vital information about whether dementia of any kind is present. Biomarker characteristics taken from MRI and Positron Emission Tomography (PET) scans are used in a study focused on the classification of AD and MCI [37] to enable an SVM classifier to differentiate between CN and MCI or CN and AD patients. For MCI and AD classification, the systems achieve 76.4% and 93.2% accuracy, respectively.

[1] https://surfer.nmr.mgh.harvard.edu/
[2] https://http//fsl.fmrib.ox.ac.uk/
[3] https://www.fil.ion.ucl.ac.uk/spm/

Brain biomarkers and Region-of-Interest (ROI)-based morphological parameters, such as cortical thickness values, volumes of the cerebral cortical grey matter, and cortical-associated white matter, are employed in a related investigation [16]. To create characteristics and detect abnormality patterns, this study presents the idea of correlated abnormalities, which is achieved by associating different ROIs with one another. Additionally, CN and MCI, CN and AD, and MCI and AD are separated using an SVM classifier that has been trained; the corresponding classification accuracies are 83.75%, 92.35%, and 79.24%. An Orthogonal Projection to Latent Structures (OPLS) is another technique that has been developed [21]. This technique, which was initially created for the modeling of complicated data, combines the concepts of Orthogonal Signal Correction (OSC) [22], [23] with Partial Least Squares (PLS) regression. The methodology is predicated on the idea that the observations are produced by latent variables. Nevertheless, systematic differences in the independent variables unrelated to the class labels seem to have a negative impact on it. As a result, the OPLS approach was created to deal with this problem. When separating AD from CN individuals, the OPLS classifier achieves a sensitivity of 86.1% and a specificity of 90.5%. Though they produce less-than-ideal outcomes, less-common alternatives include decision trees, Artificial Neural Networks (ANN) [24], [25], and other techniques for regression or classification. These techniques use cross-sectional MRI images and focus exclusively on the brain structure information found in the present scan to identify or forecast the disease.

C. Longitudinal Data

Even while the analysis of merely cross-sectional MRI scans has shown encouraging and effective results, it is limited in its ability to provide information about a single point in a disease's progression or the overall health of the brain. It is unable to reveal information about changes that occur over time, recognize patterns, or create connections with various circumstances. The research community has been highly interested in longitudinal studies as a result, which involve a series of data, like brain MRI scans, that are taken at regular intervals (e.g., every six months or annually). Some studies that handle longitudinal data have been centered around f-MRI scans [26], [27], focusing on the responses that particular brain areas display. Regression techniques, namely linear and modified least squares models, were primarily employed in these investigations. It is important to remember, though, that these studies cannot be classified as using longitudinal MRI scans in the sense that this study intends, since f-MRI captures brain activity over seconds, while the longitudinal data we are attempting to use investigates changes in the brain occurring over much longer time frames. The use of HMMs was first implemented to try to identify mild Alzheimer's disease, or early dementia, in older people [18]. In this work, characteristics taken from a series of MRI scan slices are combined into a time series, which is subsequently, subjected to HMM analysis and classification. With accuracy reaching up to 97.8% in some tests, the suggested strategy shows great promise in the early identification of dementia. However, the main goal of this study is to identify AD based on a single brain snapshot; it does not attempt to anticipate or explore the disease's progression across a number of years, and the data is still not truly longitudinal. Similarly, HMMs are used to predict the age of people who do not have dementia in a different study [20], but longitudinal data is not used in this instance. Prediction inaccuracy on average is as low as 2.57 years. The study in [28] introduces the use of longitudinal MRI scans to investigate changes and correlations between nine-year scans of cognitively normal and demented brains. By utilizing 9-year longitudinal MRI scans, scientists examine the data obtained from individual scans, opening a promising field with enormous potential for brain study. Even though we use different characteristics and datasets, this study is really important to our work since it shows how much information longitudinal MRI scans can provide.

III. Empirical Study

A. Dataset

The Alzheimer's Disease Neuroimaging Initiative (ADNI) provided the dataset utilized in this study. The National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), a few private pharmaceutical companies, non-profit organizations like the Alzheimer's Association (AA) and the Institute for the Study of Aging (ISA), and other organizations are among the sponsors of the ADNI research initiative, which was started in 2003. It functions in partnership with the National Institutes of Health (NIH) [29]. The main goal of ADNI is to collect and make use of longitudinal data from people who have been diagnosed with CN, MCI, or AD. Whether serial MRI scans, PET imaging, other biological markers, clinical and neuropsychological evaluations, and other data can be combined to track and characterize the development of MCI and early AD is the purpose of this study. Moreover, ADNI seeks to offer a freely available database of clinical and imaging information that clarifies changes over time in brain metabolism and structure, cognitive performance, and biomarkers in CN, MCI, and AD patients. More than 50 research facilities in the United States and Canada provide ADNI with subjects. A longitudinal MRI scan dataset containing 631 individuals was made available for this study. After their initial MRI scans, 192 of these people were diagnosed with CN, 309 as MCI, and 130 as AD. Consequently, 189 were classified as CN, 202 as MCI, and 240 as AD at the time of their most recent scans. Every person had one to three follow-up scans, spaced a year apart, with a variable number of follow-ups performed. A total of 1913 MRI scans, including 1.5T sagittal 3D T1-weighted MPRAGE MRI scans, were included in the dataset. The Freesurfer pipeline, an open source set of tools for the thorough and automated examination of important aspects of the human brain, was used for the preparation of these MRI data. The analysis included mapping of cortical grey matter thickness, estimation of architectonic boundaries from in vivo data, segmentation of hippocampal subfields, volumetric segmentation of most macroscopically visible brain structures, and several other functions. It also included inter-subject alignment based on cortical folding patterns. Given that manual study of such a vast dataset would require a lot of labor and time, automation processing was essential.

Consequently, each MRI scan yielded 55 MRI-derived regional measures, comprising 21 subcortical volumes and 34 cortical thickness values.

B. Evaluation Metrics

In this section, we employ a set of crucial metrics to meticulously evaluate the efficacy and precision of the classification and prediction techniques developed throughout this investigation. These metrics play a pivotal role in gauging the performance of the models, providing a comprehensive understanding of their capabilities. True Positives (TP) measure the subjects accurately identified as having AD, while True Negatives (TN) count those correctly classified as CN or having MCI. On the flip side, False Positives (FP) represents instances where subjects are incorrectly classified as having AD, and False Negatives (FN) denote subjects wrongly classified as CN or MCI when they indeed have AD. Sensitivity, or the True Positive Rate (TPR), showcases the proportion of TP samples (AD) correctly identified, expressed as $TPR = TP\ /\ (TP + FN)$. Specificity, or the True Negative Rate (TNR), quantifies the proportion of TN samples (CN/MCI) correctly classified, calculated as $TNR = TN\ /\ (TN + FP)$. Precision ($Positive\ Predictive\ Value - PPV$) signifies the accuracy of positive classifications (AD) among all positive predictions, defined by $PPV = TP\ /\ (TP + FP)$. The F1 Score, also used combines precision and sensitivity through their harmonic mean, offering a balanced assessment of the model's performance: $F1 = 2 * (PPV * TPR)\ /\ (PPV + TPR)$. Notably, we calculate the harmonic mean of specificity (TNR) and sensitivity (TPR) for a comprehensive evaluation. The Receiver Operating Characteristic (ROC) Curve provides a graphical overview of the model's performance across varying parameters. It plots sensitivity against 1 - specificity, with a superior model closer to the upper-left corner. The Area Under the Curve (AUC) quantifies the overall model performance. The Diagnostic Odds Ratio (DOR), an essential metric in medical research, gauges the odds of a positive test result when the disease is present compared to when it's absent. Calculated as $DOR = (Sensitivity * Specificity)\ /\ [(1 - Sensitivity) * (1 - Specificity)]$, higher DOR values signify better discriminatory test performance, ranging from 0 to infinity. Collectively, these metrics form a robust framework for the precise evaluation of the developed models, enabling a comprehensive assessment of their classification and prediction capabilities.

IV. Proposed Models

The techniques employed in this work are based on HMMs. The selection of HMMs was based on their innate capacity to efficiently interpret sequential data. Their architecture is a good representation of markov chains since it includes hidden states and their emissions, which maps to data that can be observed. Although HMMs are mainly used for markov chains, they are also widely used to capture sequential relationships in time-sequential data, like speech processing. They are a useful tool in our situation for processing the longitudinal MRI scans as observations and identifying the relationships between them, with an emphasis on the markov chain, a hidden structure. Three different techniques are presented in this study, each expanding on the preceding one. These techniques will be covered in detail and with thorough explanations in the sections that follow. To maintain clarity, we will now outline the data partitioning and utilization process that will be used in the upcoming sections. As was previously mentioned, the dataset consists of an assortment of MRI images from different people. A series of scans are available for each participant, including an initial cross-sectional scan and one to three follow-up scans. There are two basic ways to partition the data. The initial technique focuses on the diagnosis made from the first cross-sectional scan, which is known as the "subject-initial-group". The participants are classified as having MCI, being CN, or having been diagnosed with AD. No follow-up diagnoses are considered in this category; only the baseline diagnosis is taken into account. The "subject-end-group", which is the last follow-up scan diagnostic, is used to categorize individuals in the second technique. Similar to the first technique, this grouping yields the same diagnoses/categories as the subject-initial-group (CN, MCI, & AD) and only considers the diagnosis obtained from the most recent follow-up scan. Reconfiguring this data separation makes more sense in the context of this study. Although their labels have changed, the subject-initial-group and subject-end-group remain the same. The CN and AD groups are joined in the subject-initial group to produce two alternative categories: CN/AD or MCI. This modification makes more sense because the main goal is to investigate the development of the MCI subject-initial group, which is a high-risk group. Our goal is to ascertain whether MCI will progress to AD. As a result, the subject-end group falls into one of two categories: CN/MCI or AD. The training and testing sets for our models are defined by the subject-initial group; the training set is CN/AD, and the testing set is MCI. The particular interest in MCI patients and the investigation of their possible long-term progression are the driving forces behind this tactic. The wide range of cognitive impairments associated with MCI makes it a particularly important group in the field of medical research because it can progress into several disorders, including AD. The main goal is to assess how well our systems can anticipate outcomes for this population. The attempt to see how well an HMM can extract basic and generic structural changes that indicate progression toward AD or CN/MCI (conversion to CN or stability) is the rationale behind using non-MCI participants in the training set. Next, we evaluate the applicability of these derived features to the MCI group (a subset of the MCI group is used for training during experimentation to evaluate its effect on the overall performance of all techniques).

A. Technique 1: HMM Classification

In the first technique, HMMs are only used to evaluate how well they can extract and represent temporal structural changes in the brain throughout normal aging or as it moves closer to AD. The next step is to find out if these alterations are like those shown by a brain that has been diagnosed with motor cortex injury. HMMs are trained to maximize the probability $P(O|\lambda)$, where $O = [o_1, o_2, \ldots, o_T]$ denotes a series of observations, and λ denotes the HMM model. An observation series, represented by the letter *O* in our dataset, is equivalent to a subject's longitudinal MRI scan sequence. The volumetric data retrieved from each scan, represented by the

vector o_t, has a size of [1 × 55] for each observation [2, 4]. First, the [1 × 55] sized vectors are aggregated to create observation sequences for each subject ($[n \times 55], n \in [2, 4]$). Subsequently, the non-MCI subject-initial-group data are employed to train λ_{AD} and $\lambda_{CN/MCI}$, two HMMs. Only the observations from individuals in the AD and CN/MCI subject-end groups are considered for each HMM. Following the effective training of these two HMMs, testing is conducted on the MCI subject-initial-group. Two probabilities, $P_{AD}(O_i|\lambda_{AD})$ and $P_{CN/MCI}(O_i|\lambda_{CN/MCI})$, are calculated for every observation sequence using the forward algorithm. These probabilities show how likely it is that the matching HMM might produce each sequence. The sequences are predicted based on this probability.

$$y_i = \begin{cases} AD, \ if \ P_{AD}(O_i|\lambda_{AD}) \geq P_{\frac{CN}{MCI}}(O_i|\lambda_{\frac{CN}{MCI}}) \\ CN/MCI, \ if \ P_{AD}(O_i|\lambda_{AD}) < P_{CN/MCI}(O_i|\lambda_{CN/MCI}) \end{cases} \quad (1)$$

It is crucial to stress that we do not assign any semantics to the states in the HMMs. When using an HMM conventionally, the number of states is usually selected so that each state, or combination of states, corresponds to a unique "logical state" of the underlying process. Nevertheless, modeling the sequences in this way is not practical in our scenario because of the sparse nature of the scans. From here on, we think of the states as an independent variable in the system that we can work with and examine to see how it affects the system's overall performance.

B. Technique 2: HMM Modelling SVM Classification

There is a premise that there is room for improvement even when the initial technique has produced good results. The data has shown that it contains useful information, and the HMM's ability to extract and represent this information has been validated. The investigation of whether this data may be organized in a way that makes it compatible with alternative models and techniques is therefore of particular interest. There is also some interest in the possibility of improving performance by adding a strong classifier. Although the HMM's states were not given explicit meanings in the previous discussion, it is agreed that the transition matrix shows that the HMM may implicitly assign certain meanings to its states after training. HMMs regard their states as markov chains by definition. As a result, the data structure for the observation variable (time) is examined during the training phase. Patterns or anomalies that appear repeatedly in the observation sequences are identified, and the start, transition, and emission probabilities are set up to correspond with these patterns. To shed further light on the reasoning behind this approach, look at the following example: Let's say the goal is to create a model of adults' everyday activities using observations made at predetermined times of the day. These findings differentiate between those who are employed and those who are not. Because of this, these two groups' observation sequences are different from one another, reflecting their different lifestyles. After being trained in this scenario, an HMM adjusts its probability to construct state markov chains that produce observation sequences that match each adult's lifestyle. The transition matrix might have been initialized to direct the HMM in giving the states particular meanings to predefine state transitions. State definitions for "working", "commuting to/from work", "sleeping", and "resting" may have been applied in the preceding case. The HMM would organize the state markov chains in a way that is easier to understand by initializing higher probabilities for state transitions like "commuting to/from work to working", "working to commuting to/from work", and "commuting to/from work to eating", and lower probabilities for transitions like "working to sleeping" and "sleeping to working". It is important to remember that while this alignment of states to actions could make sense more naturally to humans, the HMM itself may not necessarily gain from it. The HMM may nevertheless arrange the markov chains in a way that makes sense for the data's behavior even in situations where states and actions aren't directly connected. This is true even when it's not immediately obvious to human observers. The same action may even be assigned to several states by the HMM. Either way, the general organization of the states and how they behave is tailored to the features of the training set. The states of our HMMs may theoretically represent the three cognitive states of the participants (CN, MCI, and AD), much like the example given. It could have been possible to initialize the HMMs in a way that directed them to ascribe states based on the predetermined cognitive conditions right from the start. Nevertheless, we have chosen not to initialize the HMMs because of the characteristics of the data and the inherent unpredictability of these circumstances. Rather, we let them independently determine the state structure without any prior knowledge. The phrase "nature of data" refers to a range of characteristics that are taken out of every MRI scan, including the corresponding diagnosis, which is prone to inaccuracy. The three cognitive states are also rather inclusive. In particular, the MCI condition has a broad range of severity fluctuations, as was explained in previous sections; this feature also applies to the AD condition. As a result, it makes sense to believe that there are intermediate states between the states that exactly match the predetermined conditions. Because of this, we do not initialize the HMM's matrices; instead, we allow the training process to define the probabilities related to the state and emission structures. We specifically want to use this characteristic of HMM modeling in this technique. We hypothesize that important information contained in the state sequences corresponding to our observation sequences can be used as features for an alternative model or classifier. Our goal is to produce state sequences, or features, by extending the logic and foundation set by the prior technique. These features will then be used to train an SVM classifier. As previously mentioned, the procedure creates observation sequences ($O = [o_1, o_2, \ldots, o_T]$) for each subject. Then, using the CN/AD subject-initial-group, an HMM is trained. Unlike the prior technique, which trained multiple HMMs depending on the subject-end-group for each observation, this technique simply trains one HMM. Our goal in making this decision is to investigate the inherent capacity of the HMM to define and model discriminative information on its own, as well as to extract more generic characteristics from the data. Following training, all observation sequences (including the subject-initial-groups CN/AD and MCI) have state sequences produced by the HMM λ. These state sequences function as

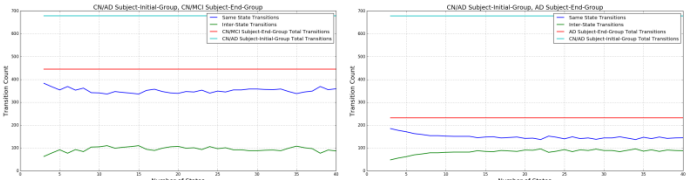
characteristics for the technique's next step. An SVM classifier is trained using feature sequences from the CN/AD subject-initial-group. After being trained for binary classification, the SVM divides the data into two categories: AD and CN/MCI, depending on which subject-end group belongs to which sequence. After training, the efficacy of the SVM is determined by analyzing its ability to categorize the MCI subject-initial-group sequences into the two designated classes (AD or CN/MCI).

C. Technique 3: HMM Modelling SVM Classification 2

Due to the intrinsic properties of the data, the state sequences produced by the HMM show a significant amount of volatility and fluctuation, especially when the number of states rises. Additionally, the original state sequences are features, but they invariably have varying durations based on how long the observation sequence was. This unpredictability makes the final prediction more difficult to make and adds instability to the classification process. After producing state sequences for different state counts, it is clear that, in the CN/AD and MCI datasets, the state sequences for the CN/MCI subject-end-group exhibit a more consistent pattern than those from the AD subject-end-group. There are few state transitions in the case of CN/MCI participants. This difference between the more stable conduct of CN/MCI subjects and the more erratic behavior of AD subjects may be due to structural similarities in the brain throughout normal and pathological aging. The brain regions under study typically undergo constant changes during the ordinary aging process, frequently at a slow and steady rate (e.g., a specific brain area steadily decreases in volume over the years). But because aberrant aging is characterized by unpredictable aging, these changes become more sudden and difficult to monitor precisely. There is clearly a substantial correlation even though it is still unknown if the irregular changes are the result of atypical aging or the cause of it. It may even be a combination of the two. We are mostly interested in transitions that either keep the present state or change it in the analysis of state transitions:

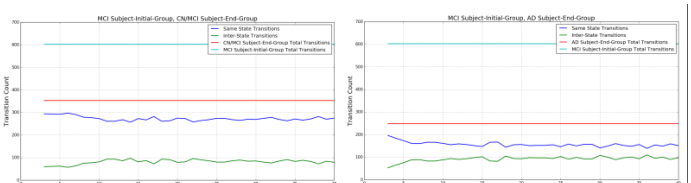
$$s_t \longrightarrow s_{t+1}: \begin{cases} same - state\ transition, & if\ s_t \equiv s_{t+1} \\ inner - state\ transition, & if\ s_t \not\equiv s_{t+1} \end{cases} \quad (2)$$

As such, we track and log the total number of transitions as well as the intra-state (same-state) and inter-state transitions that occur inside each group and end-group. The above Eq. (2) is used to determine the percentages of same-state and inter-state transitions that take place inside the AD subject-end-group of the CN/AD individuals. While 75% of the transitions are inter-state, 25% are same-state transitions. Similar computations can be made for every group and end-group to find interesting and possibly useful differences that the HMMs have brought to light. Using HMMs with different numbers of states, Fig. 1 and Fig. 2 show the counts of same-state and inter-state transitions for various groups. As a reference frame, the overall transition counts for the relevant groups are also displayed. Because they are correlated with sequence lengths—that is, the number of follow-ups scans the participants receive—rather than the structural analysis of the HMMs, these total transitions stay constant across the graphs. Interestingly, the numbers show that in the CN/MCI subject-end-groups, whether they are the CN/AD or MCI initial-groups, the number of same-state transitions is much larger (about three to four times) than the number of inter-state transitions. The AD subject-end-groups, on the other hand, show a less noticeable difference (about 1.5 to 2 times). Figs, which show the percentages of same-state and inter-state transitions compared to all transitions, support this conclusion. The Figures also compute these percentages' mean and variation across a growing number of HMM states. The information in the table highlights the fact that roughly 22% of CN/MCI end-group transitions are inter-state and 78% of them are same-state. As opposed to the CN/MCI groups, the AD end-group generates sequences where roughly 63-64% of the transitions are same-state, supporting our initial claim that the AD end-group's sequences follow a more regular pattern.



(a) CN/MCI end-group (b) AD end-group

Fig. 1. Number of state transitions that the CN/AD group of HMMs trained with a growing number of states experienced.

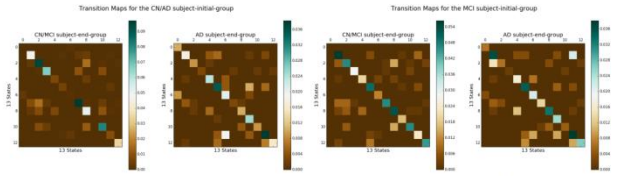


(a) CN/MCI end-group (b) AD end-group

Fig. 2. Number of state transitions that the MCI group of HMMs trained with a growing number of states experienced.

Now, we want to take use of this property and remove the variability caused by the length of the generated features. As shown in Fig. 3, we create transition frequency maps to accomplish this. These maps are represented as $[N \times N]$ matrices, where N is the total number of HMM states. As a counter, each element in the matrix, a_{ij}, counts the number of transitions from state i to state j. Interestingly, the elements of the matrix diagonal represent inter-state transitions, whereas the components of the same-state diagonal correspond to same-state transitions. Reiterating that we treat the number of HMM states as a variable that can be adjusted to investigate its effect on system performance is crucial. Consequently, Fig. 3—13 states are just meant to serve as an illustration. These matrices are used as feature vectors for the subjects after being serialized in a row-wise manner. This technique avoids practical issues by releasing our features from the temporal element, which is intrinsic to their structure but has no bearing on their length. It is clear from the initial matrices that the feature vectors are extremely sparse, with few non-zero elements that frequently take values in the range of $a_{ij} \in [1, 2, 3]$. The feature vectors are primarily composed of zeros. As such, the non-zero components' placements are more significant than their exact values. This greatly reduces the

work for an SVM classifier in comparison to the previous technique's separation of state sequences. In particular, this issue is made simpler by the fact that it may be handled by an SVM as a spatial separation problem for 2D data. The procedural stages in this technique are like those in the prior way. First, the data is prepared, and for each subject, observation sequences ($O = [o_1, o_2, \ldots, o_T]$) are created. Once more, using the CN/AD subject-initial-group, another HMM is trained. Then, as previously said, state sequences are generated for each observation series, which are then utilized to make transition maps and, ultimately, converted into feature vectors. The SVM classifier is then trained using these feature vectors, with an emphasis on the CN/AD training set. Based on the end-groups of the patients, the classifier is trained to classify data into AD and CN/MCI. Ultimately, the MCI testing set is used to assess the classifier's performance. The main difference between Techniques 2 and 3 is the type of characteristics that are sent into the SVM classifier.



(a) Transition Frequency Maps of CN/AD subject-initial-group (b) Transition Frequency Maps of MCI subject-initial-group

Fig. 3. Mapping transition frequencies using a 13-State HMM.

V. Experimental Analysis

This section presents the results and evaluation of the experiments with comparison, contrast, and discussion of the different techniques.

A. Experimental Setup

Python, the hmmlearn toolbox, scikit-learn, and a number of machine learning techniques are used in the research. The construction of HMMs and SVM classifiers for the purpose of classifying subjects into various groups is the main goal of these investigations. Those without CN, those suffering from AD, and those with MCI are among these subjects. HMM models are made with the help of the hmmlearn toolkit. Since these models are fully coupled upon startup, it is possible to move between any states. Based on Gaussian emission distributions, the emission probabilities have a "spherical" covariance, which means that a single covariance value is applicable to every feature. The implementation and test run can affect how many states the HMMs have. We decided to configure the hyperparameters in advance and maintain consistency across numerous techniques and approaches for the SVM classifier. The kind of kernel, the kernel coefficient (γ), the penalty parameter (C), and the independent term for polynomial kernels are the hyperparameters that are being examined. These parameters include a polynomial kernel of degree 3, a penalty parameter C of 63.26, a γ value of 0.001, and an independent term of the polynomial function of 3.

The way the training data is handled in the experiments is one intriguing feature. The MCI subject group is first left out of the training process for the HMM and the SVM classifier by us. This technique is predicated on the idea that a greater variety and quantity of training data improve model performance and lower the likelihood of overfitting. However, choose to carry out more research to see if adding any MCI data to the training set can enhance the system's functionality. We use a technique often used in machine learning, called *k*-fold cross-validation, to assess the models' performance. To evaluate how well the models generalize their behavior to new data using this technique. There are *k* subsets of the data; *k* - 1 subsets are utilized for training, and the remaining subset is used for testing. The ultimate performance measure is calculated by averaging the evaluation metrics or errors generated in each run of this process, which is performed *k* times. We use a variant of cross-validation to incorporate MCI data into the training procedure. CN/AD and MCI participants are first given different training and testing sets. The MCI group is then divided into *k* = 3 folds, of which 2 are chosen as testing sets and 1 is combined with the training set. To significantly influence the process, this method yields about 25% of the training set as MCI individuals. Importantly, this modified cross-validation strategy is considered "semi-blind", whereas tests that are carried out without using MCI data in the training set are referred to as "blind". The SVM training procedures in Techniques 2 and 3 employ a conventional cross-validation procedure. The training data is split into *k* folds (with *k* values of 5, 7, or 10), either as CN/AD solely or as CN/AD plus one-third of MCI data. F1-scores are computed after training and testing several SVM classifiers. The testing set, which consists of MCI data, is classified by the SVM classifier with the greatest F1-score in preparation for the assessment.

B. Result

The following graphics contain the metrics of the various procedures under test. These metrics are assessed with and without the use of 5, 7, and 10-fold cross-validation on the SVM training, as well as with and without cross-validation on the training data. Furthermore, metrics are generated and provided for the participants who have undergone the maximum number of follow-up scans, which are three follow-ups.

1) Random classifier: Since the technique used in this study does not build upon an earlier approach, there are no state-of-the-art outcomes to compare with. As a result, the outcomes will be contrasted with a random classifier, whose performance threshold is set at the lowest possible value. Based on the values shown in Section III (A), the prior probabilities for the two classes (CN/MCI and AD) in the dataset can be defined as:

$$P_{CN/MCI} = \frac{Number\ of\ CN\ and\ MCI\ diagnoses\ at\ Last\ Scan}{Number\ of\ all\ subjects} = \frac{391}{631} = 0.62 \quad (3)$$

$$P_{AD} = \frac{Number\ of\ AD\ diagnoses\ at\ Last\ Scan}{Number\ of\ all\ subjects} = \frac{240}{631} = 0.38 \quad (4)$$

Any data point is given a class using a random classifier based on a predetermined probability:

$$P_{random} = \begin{cases} p_{class},\ where\ class\ \in\ [\frac{CN}{MCI}, AD] \\ \frac{1}{2},\ equal\ probability\ of\ assigning\ either\ class \end{cases} \quad (5)$$

In the first case, the classifier classifies each data point with a probability equal to the priors of the two classes, achieving the highest possible classification accuracy overall. In the second scenario, the minor class—in our case, AD—is marginally favored by the classifier. By increasing the system's sensitivity, this technique seeks to maximize the detection of AD at the cost of poor specificity, which increases the number of AD cases detected but also raises the possibility of FP results. In all scenarios, the recall, which gauges the proportion of accurately categorized data points in a particular class, stays random. According to our experimental findings, specificity is correlated with the memory of the CN/MCI class and sensitivity with the recall of the AD class. Thus, we would get the following results from the first random classifier: $Sensitivity = 0.38,\ Specificity = 0.62, and\ F1 = 0.4712$, while we would gain the following results from the second classifier: $Sensitivity = Specificity = F1 = 0.5$. The DOR for both iterations of the random classifier is 1. Currently, we have chosen the maximum values for sensitivity and specificity as our lower bounds to maximize performance optimization. As so, the following cutoff points are determined:
$Specificity_{min} = 0.62, F1_{min} = 0.554, and\ Sensitivity_{min} = 0.5.$

2) Technique 1: HMM Classification*:* Fig. 4, 5, 6, and 7 show the harmonic mean of the two (F1-score) and the sensitivity and specificity measures for increasing numbers of states. These graphs show that the number of states does not rise along with the performance of the system. The Fig. 4 and 6 show that, although the effect is not significant, there is a tendency for sensitivity to rise and specificity to fall while the F1-score remains stable. Striking for near and high values for both sensitivity and specificity is generally accepted as the standard technique. While reaching the highest possible level for both is desirable, these two metrics frequently show an adverse connection, even in the case of an absolute classifier. Positive and negative data points are accurately classified by a highly effective classifier with little to no FP and FN. Reduced specificity and sensitivity are the results of these FP and FN. A classifier's sensitivity will be almost perfect, but its specificity will be quite poor if it overclassifies one class, for example, classifying all data points as positive. In order to obtain a higher performance overall, it is wise to establish a balance between these criteria. It is clear from the blind experiment (see Fig. 4) that the sensitivity is higher than the allowable limit and the specificity first reaches the limit before declining as more states are used. Subjects with three follow-ups show a similar tendency (see Fig. 6), but the metrics are improved, leading to improved specificity performance and a delayed divergence between the two metrics. Both times, the F1-score greatly exceeds the upper bound. On the other hand, noticeably more stable metric graphs are produced by the semi-blind experiment (see Fig. 5). With very few exceptions, specificity usually reaches or exceeds the 0.62 threshold while sensitivity stays high. The stability of the graphs includes both the proximity of the two measures when compared and fluctuations in each metric separately. Similar behavior is shown in participants who have had three follow-ups, with higher metrics and a particularly noticeable improvement in specificity (see Fig. 7). The evolution of the DOR for the blind and semi-blind tests, with the full MCI group or just participants who had three follow-ups, is shown in Fig. 8. With all ratios remaining well over 2 (with 1 as the limit) and exhibiting little volatility, excellent results are seen in this case. When subjects receive three follow-ups, the blind experiment performs best in terms of DOR.

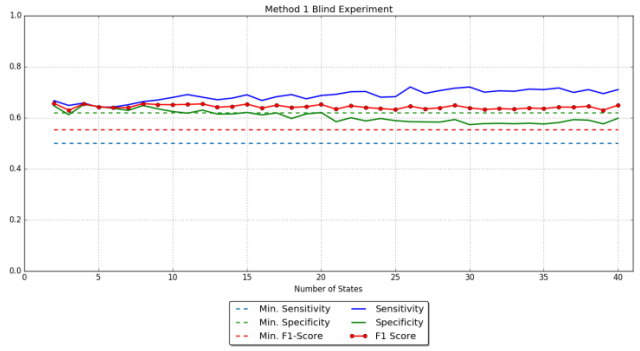


Fig. 4. Technique 1 with blind experiments: sensitivity and specificity for increasing number of HMM states (F1 score on average is 0.64).

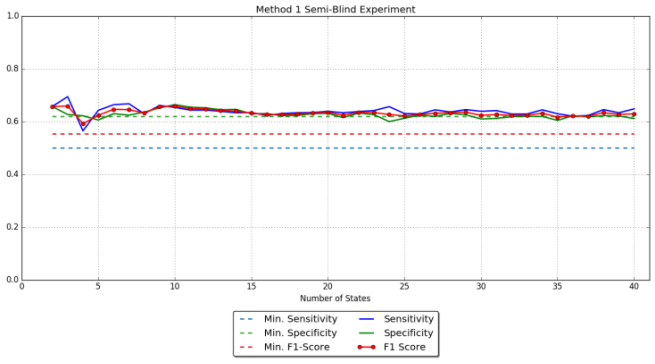


Fig. 5. Technique 1 with semi-blind experiments: sensitivity and specificity for increasing number of HMM states (F1 score on average is 0.63).

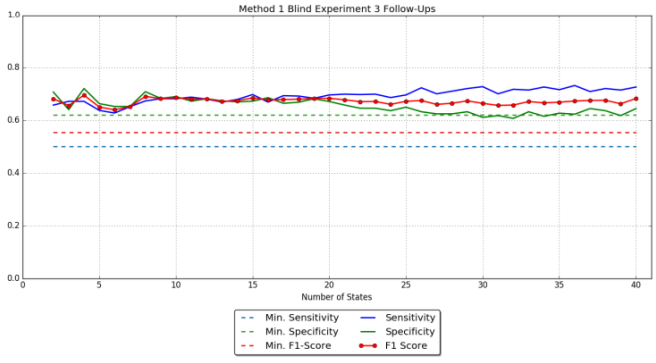


Fig. 6. Technique 1 with 3 blind experiments: sensitivity and specificity for increasing number of HMM states (F1 score on average is 0.67).

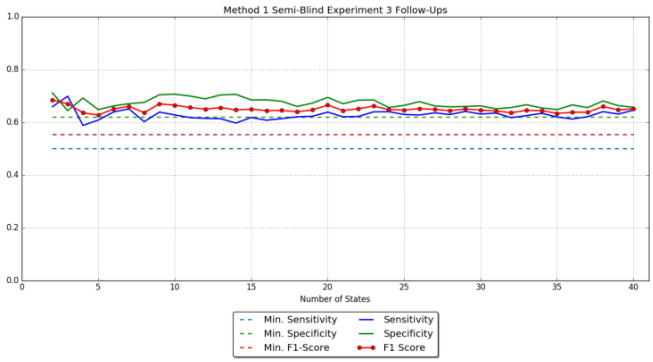


Fig. 7. Technique 1 with 3 semi-blind experiments: sensitivity and specificity for increasing number of HMM states (F1 score on average is 0.64).

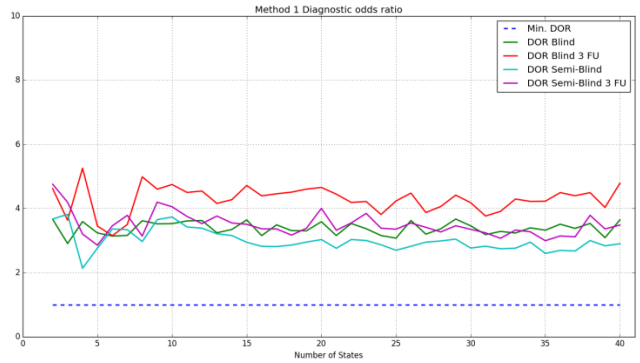


Fig. 8. DOR for the different approaches of technique 1.

3) Technique 2: HMM Modelling SVM Classification*:* The graphs take on greater interest when the second technique is examined in Fig. 9 to Fig. 20. This technique shows the F1 scores, sensitivity, and specificity for every experiment variant—Blind/Semi-Blind, all/only three follow-up scans, and each unique number of SVM folds (5, 7, & 10). Fig. 9, 10, and 11 show that the graphs for the various folds within the same type of trial are nearly identical, suggesting that SVM cross-validation has no discernible impact on system performance. Notably, the second technique shows extremely low sensitivity and very high specificity (approaching 1 for participants with three follow-up scans), which results in a low F1 score. The preceding section covered the phenomena of inverse behavior between sensitivity and specificity. After doing a thorough study of the data and examining the confusion matrices generated (see Fig. 24), it is apparent that the classifier primarily classifies most of the data as CN/MCI, which accounts for the remarkably high specificity and low sensitivity. The significant variety of the state sequences produced by the HMMs and utilized as feature vectors for the SVM makes them non-separable data, as was previously discussed. As a result, the SVM finds it difficult to identify an appropriate separating hyperplane. Upon reviewing the DORs see Fig. 21, 22, and 23) for this technique, it is evident that the system's behavior has not been affected by cross-validation for SVM, since all DOR graphs remain almost the same for varying numbers of folds. The DOR values are extremely low—much lower than those obtained using technique I. DORs in the blind tests may fall to levels less than 1. With this technique, adding MCI cases to the training set results in a marginal improvement in performance, which is mainly manifested in a smaller sensitivity/specificity divergence. Even yet, the overall outcomes are still unimpressive. Fig. 24 show two examples of confusion matrices for several experiment runs that show how many data points were categorized into each class. The confusion matrix's decimal numbers can be explained by the fact that all trials, as mentioned in Section III (A), are run ten times in order to reduce the impact of outliers. The average number of data points classified in each class over ten experiments with the same settings is effectively represented by these decimal figures. The creation of Technique 3 was spurred by the significant differences in sensitivity and specificity found in this technique.

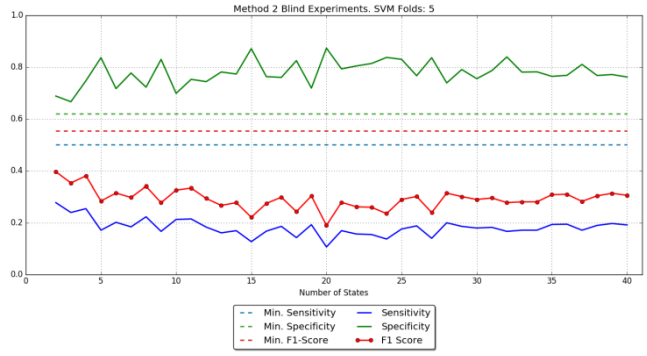


Fig. 9. Technique 2 with blind experiments: sensitivity and specificity for increasing number of HMM states with 5 SVM folds (F1 score on average is 0.29).

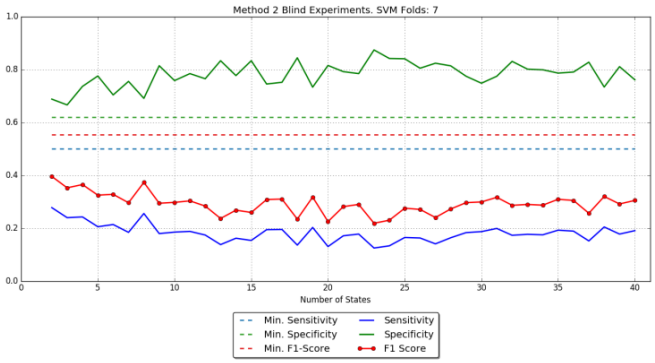


Fig. 10. Technique 2 with blind experiments: sensitivity and specificity for increasing number of HMM states with 7 SVM folds (F1 score on average is 0.29).

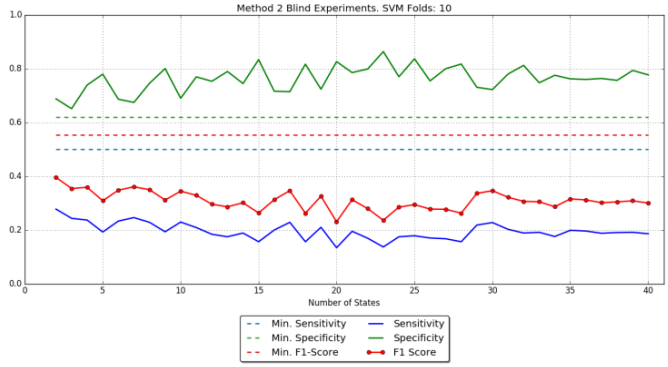


Fig. 11. Technique 2 with blind experiments: sensitivity and specificity for increasing number of HMM states with 10 SVM folds (F1 score on average is 0.30).

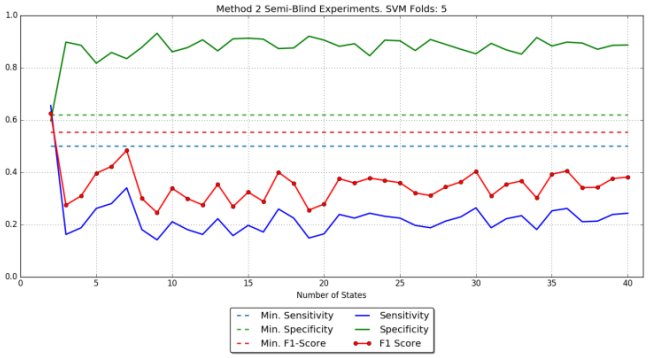


Fig. 12. Technique 2 with semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 5 SVM folds (F1 score on average is 0.34).

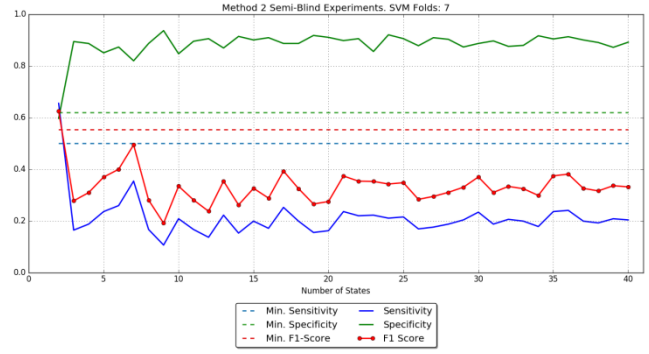


Fig. 13. Technique 2 with semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 7 SVM folds (F1 score on average is 0.33).

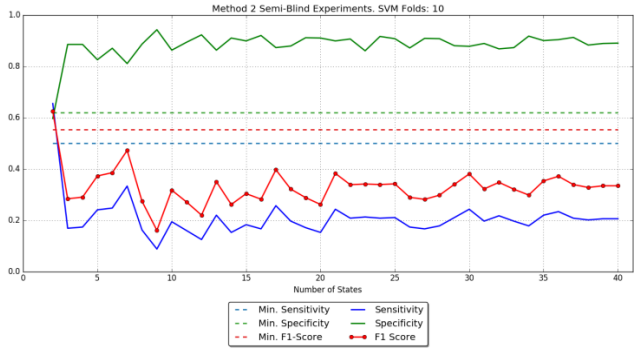


Fig. 14. Technique 2 with semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 10 SVM folds (F1 score on average is 0.32).

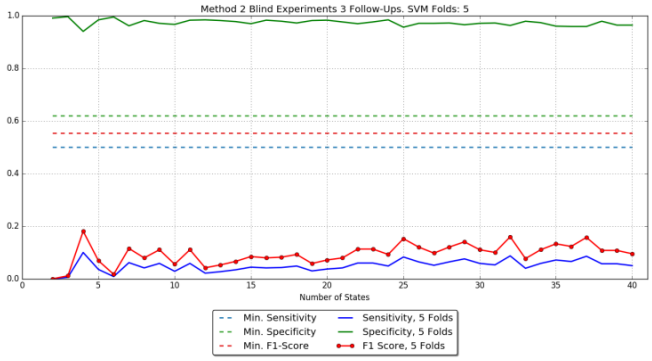


Fig. 15. Technique 2 with 3 blind experiments: sensitivity and specificity for increasing number of HMM states with five SVM folds (F1 score on average is 0.09).

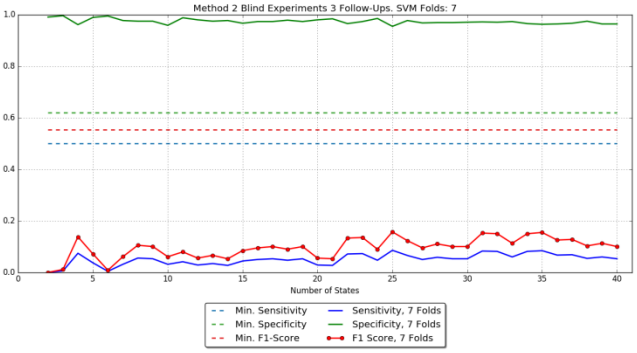


Fig. 16. Technique 2 with 3 blind experiments: sensitivity and specificity for increasing number of HMM states with 7 SVM folds (F1 score on average is 0.09).

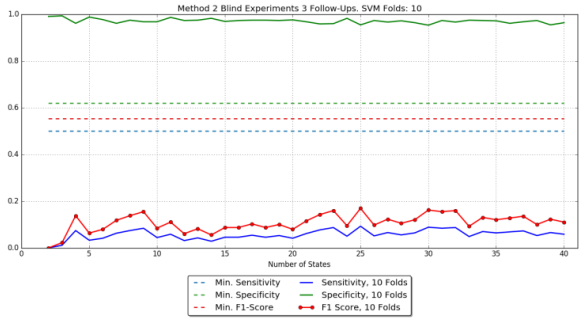


Fig. 17. Technique 2 with 3 blind experiments: sensitivity and specificity for increasing number of HMM states with 10 SVM folds (F1 score on average is 0.15).

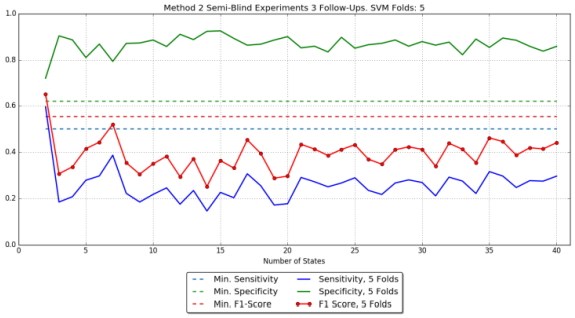


Fig. 18. Technique 2 with 3 semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 5 SVM folds (F1 score on average is 0.39).

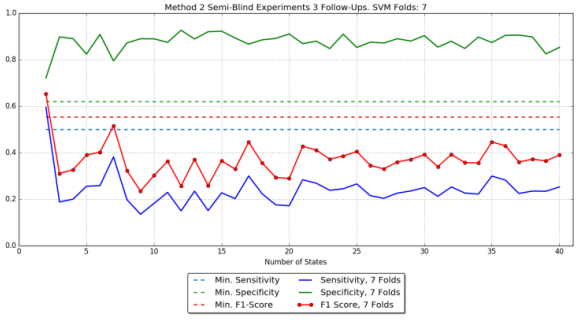


Fig. 19. Technique 2 with 3 semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 7 SVM folds (F1 score on average is 0.36).

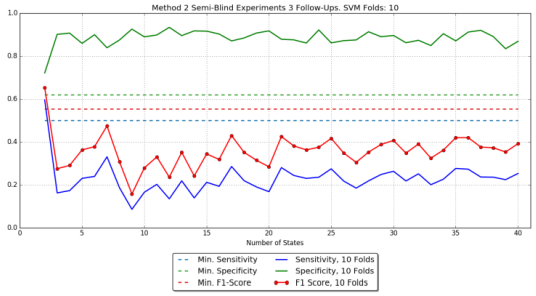


Fig. 20. Technique 2 with 3 semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 10 SVM folds (F1 score on average is 0.35).

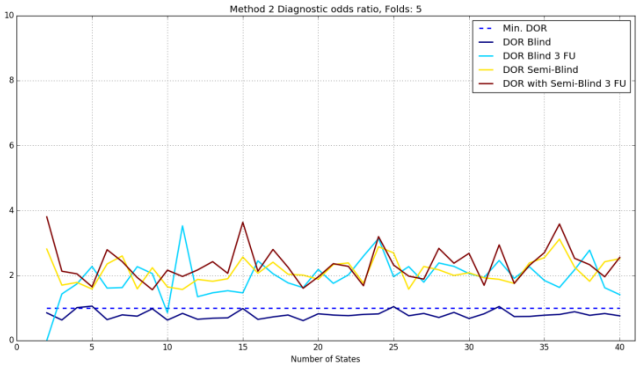


Fig. 21. DOR for the different approaches of Technique 2 with 5 folds.

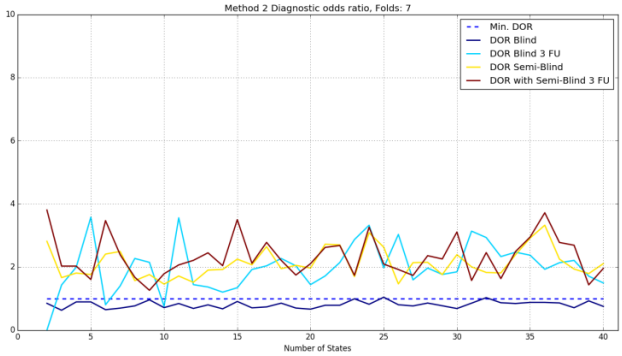


Fig. 22. DOR for the different approaches of Technique 2 with 7 folds.

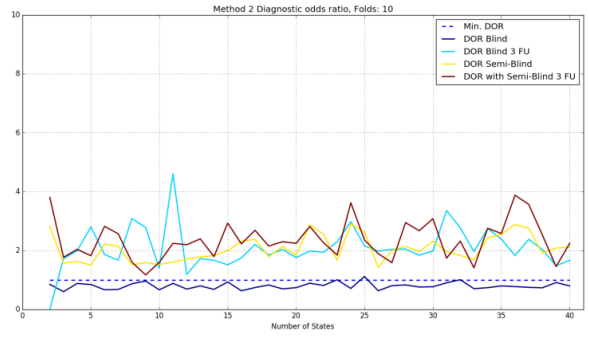
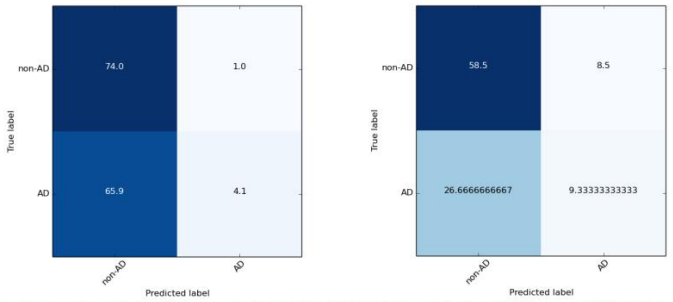


Fig. 23. DOR for the different approaches of Technique 2 with 10 folds.



(a) 11 States without Cross-Validation, 10 SVM Folds &(b) 17 States with Cross-Validation, 5 SVM Folds & with Subjects with 3 Follow-Ups all MCI Subjects

Fig. 24. Confusion matrix for Technique 2.

4) Technique 3: HMM Modelling SVM Classification 2: In the context of the third technique, the sensitivity, specificity, and F1 score graphs for the different tests and varying numbers of SVM folds are displayed once more (see Fig. 25- Fig. 36). SVM cross-validation is often found to have minimal impact on system performance when the number of folds is changed. It is important to note that the semi-blind trials are greatly impacted by the participation of MCI participants. As demonstrated in Technique 1, the disparity between sensitivity and specificity is reduced (see Fig. 28, 29, 30, 34, 35 and 36). Not only is there no divergence here, but convergence is observed. Both metrics cross over in each of the Figures and then stay rather close after that. Fig. 31– Fig. 33 show how, during the experiment, sensitivity and specificity closely coincide with one another. Sensitivity/specificity graphs and DOR graphs are provided, much like in the previous two techniques. Specificity faces difficulties in the blind trial, falling to and remaining at the lowest threshold. But there is a noticeable improvement when compared to the second technique, suggesting that using frequency maps instead of the real state sequences makes the data easier to separate and preserves important details about the evolution of the condition. The semi-blind studies do, in fact, help to lessen the sensitivity/specificity gap, although the F1 score, and both measures show a modest reduction (see Fig. 28–30). With values far above 3.0, the DOR graphs in Fig. 37– Fig. 39 show a discernible improvement over the previous technique. When MCI participants are included in the training process, these data show a reduction that is comparable to that seen in the sensitivity/specificity graphs.

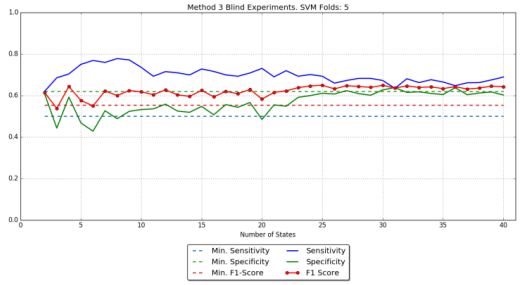


Fig. 25. Technique 3 with blind experiments: sensitivity and specificity for increasing number of HMM states with 5 SVM folds (F1 score on average is 0.62).

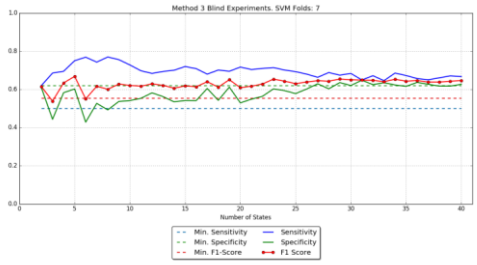


Fig. 26. Technique 3 with blind experiments: sensitivity and specificity for increasing number of HMM states with 7 SVM folds (F1 score on average is 0.62).

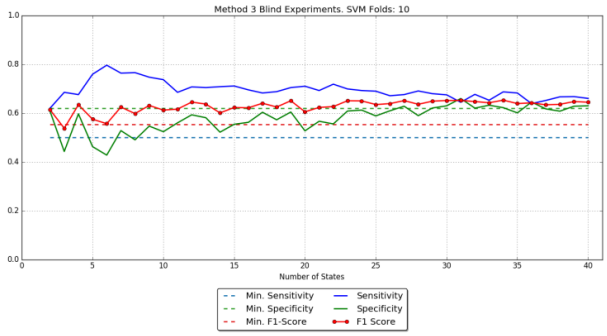


Fig. 27. Technique 3 with blind experiments: sensitivity and specificity for increasing number of HMM states with 10 SVM folds (F1 score on average is 0.62).

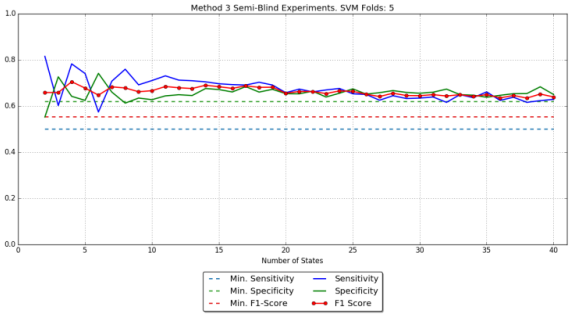


Fig. 28. Technique 3 with semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 5 SVM folds (F1 score on average is 0.66).

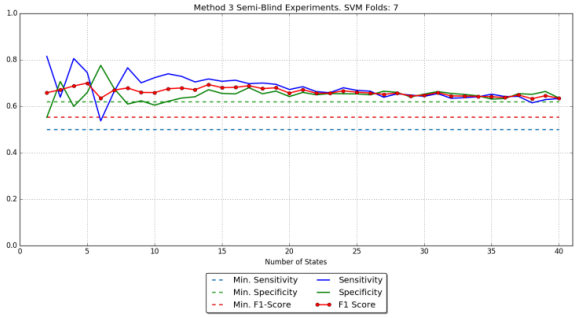


Fig. 29. Technique 3 with semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 7 SVM folds (F1 score on average is 0.66).

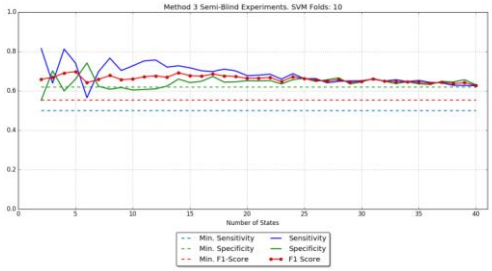


Fig. 30. Technique 3 with semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 10 SVM folds (F1 score on average is 0.66).

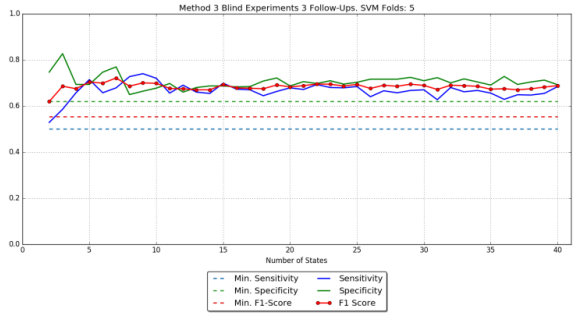


Fig. 31. Technique 3 with 3 blind experiments: sensitivity and specificity for increasing number of HMM states with 5 SVM folds (F1 score on average is 0.68).

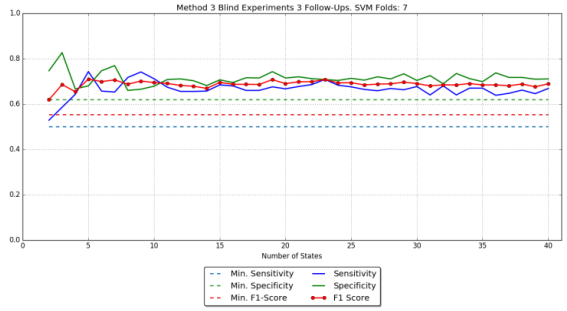


Fig. 32. Technique 3 with 3 blind experiments: sensitivity and specificity for increasing number of HMM states with 7 SVM folds (F1 score on average is 0.68).

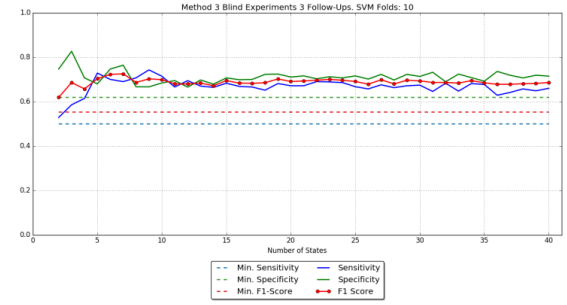


Fig. 33. Technique 3 with 3 blind experiments: sensitivity and specificity for increasing number of HMM states with 10 SVM folds (F1 score on average is 0.68).

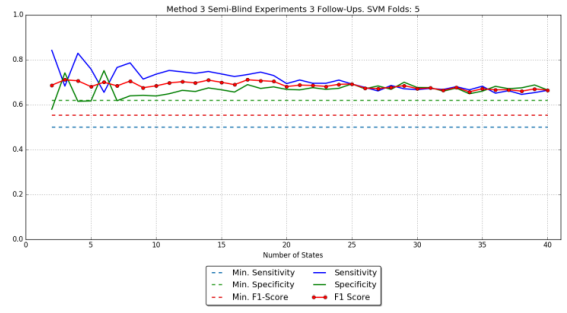


Fig. 34. Technique 3 with 3 semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 5 SVM folds (F1 score on average is 0.68).

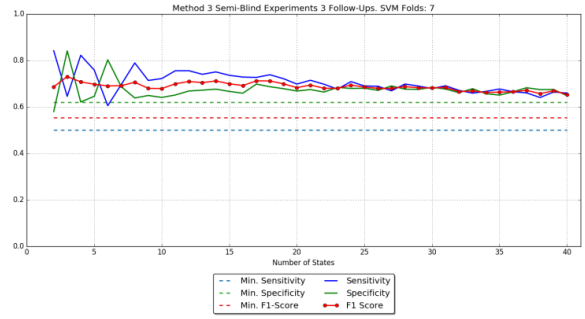


Fig. 35. Technique 3 with 3 semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 7 SVM folds (F1 score on average is 0.68).

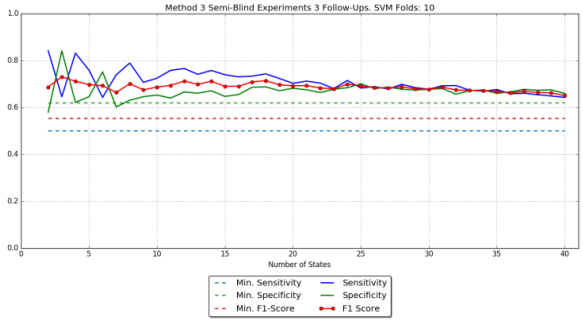


Fig. 36. Technique 3 with 3 semi-blind experiments: sensitivity and specificity for increasing number of HMM states with 10 SVM folds (F1 score on average is 0.68).

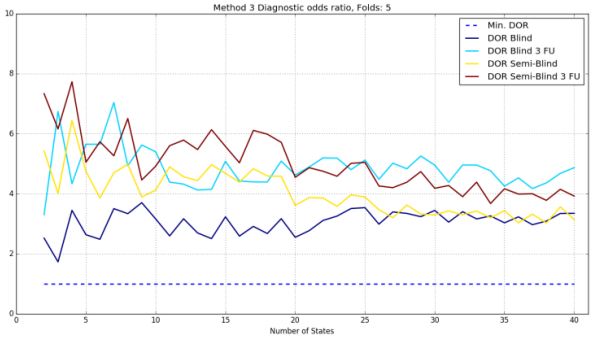


Fig. 37. DOR for the different approaches of Technique 3 with 5 folds.

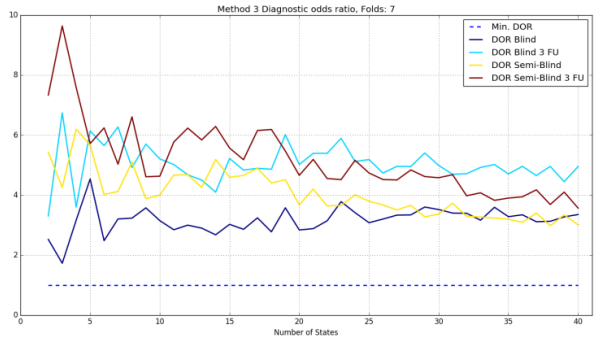


Fig. 38. DOR for the different approaches of Technique 3 with 7 folds.

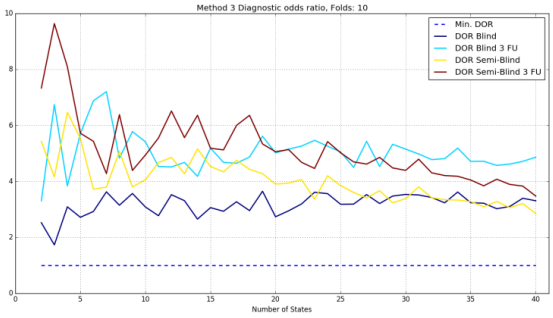
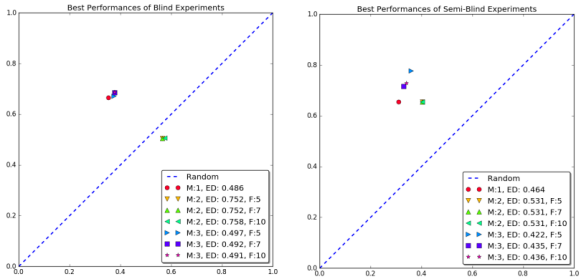


Fig. 39. DOR for the different approaches of Technique 3 with 10 folds.

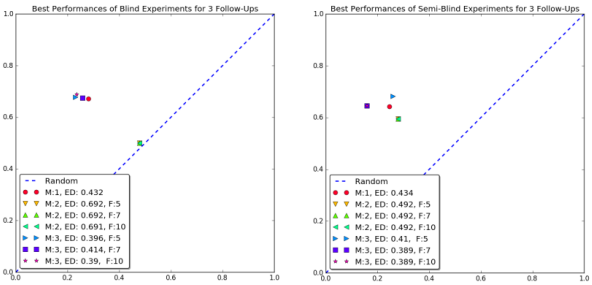
VI. Discussion

Thus far, the analyses and findings have primarily operated at a theoretical level, focusing on the techniques from a broader scientific perspective and integrating sensitivity, specificity, and DOR graphs generated across multiple states. On the other hand, specific instances of the trained and evaluated classifiers would be of interest to us if these techniques were to be applied in real life. Put another way, we want to assess the best "runs" of each technique rather than comparing results for the HMMs across various numbers of states (the major variable). To do this, we created Fig. 40 and Fig. 41, which show a ROC space filled with data points that indicate the greatest examples of each technique. For every technique, these data points represent the ideal run's TPR and 1−TNR. The various techniques are denoted by the letter *M* in the legend of these figures. Furthermore, we present the Euclidean distance (ED) of every point as measured from the upper-left corner; smaller EDs denote better performance. Tables I and II also provide a summary of these numerical values. The various methodologies and techniques can be more easily compared and contrasted thanks to this tabulation and visualization, which presents the performances visually. These Figures clearly show that adding MCI participants to the training set improves all tested techniques' peak points, with technique 2 showing the greatest improvement. Although Technique 2's initial performance was lower than that of a random classifier, it eventually obtains performance comparable to the other two (see Fig. 40). Furthermore, Technique 1 and 3 yield almost identical results, with Technique 3 slightly outperforming, especially for participants who receive three follow-up scans. An additional important inference from these Figures, which is further supported by the results of the separate technique, is the importance of prolonged MRI sequences. We regularly see that participants with three follow-up scans had better results than the overall results in all the presented Figures. This implies that there may be a greater likelihood of detecting AD progression, MCI progression, or perhaps conversion to CN in these individuals. It emphasizes how important it is for each person to get as many follow-up scans as possible to provide more precise projections. Moreover, selecting a different number of folds for SVM training has little effect on performance, consistent with other findings. One noteworthy observation from the data shown in Tables I and II is that although the metrics for sensitivity and specificity increase for participants who receive three follow-ups, the relative importance of the measures is inverted. When classifying all individuals, our technique shows higher sensitivity; but, when applied to participants with longer MRI sequences, it shows higher specificity. The unequal distribution of diagnosis across various numbers of follow-up scans can be used to justify this.



(a) All Methods Without Cross-Validation (b) All Methods With Cross-Validation

Fig. 40. Best performance of all techniques.



(a) All Methods Without Cross-Validation, Subjects with 3 Follow-ups (b) All Methods With Cross-Validation, Subjects with 3 Follow-ups

Fig. 41. Best performance of all techniques with 3 follow-ups.

TABLE I. An Overview of Each Method's Best Outcomes in Relation to the Point on the Roc Space That is Closest to the Upper Left Corner

Technique	Type	Specificity 5, 7, 10 Folds	Sensitivity 5, 7, 10 Folds	Distance 5, 7, 10 Folds	Avg. F1 5, 7, 10 Folds	Avg. DOR 5, 7, 10 Folds
1	Blind	0.646	0.665	0.486	0.643	3.381
1	Semi-Blind	0.689	0.655	0.463	0.633	2.998
2	Blind	0.434, 0.434, 0.425	0.504, 0.504, 0.505	0.752, 0.752, 0.757	0.293, 0.293, 0.31	0.794, 0.809, 0.797
2	Semi-Blind	0.596, 0.597, 0.597	0.655, 0.654, 0.654	0.530, 0.530, 0.530	0.35, 0.333, 0.329	2.135, 2.123, 2.066
3	Blind	0.626, 0.622, 0.622	0.672, 0.685, 0.686	0.496, 0.491, 0.490	0.622, 0.629, 0.628	3.054, 3.172, 3.173
3	Semi-Blind	0.641, **0.67**, 0.659	**0.776**, 0.715, 0.728	**0.422**, 0.535, 0.66	**0.662,** 0.661, 0.66	4.018, **4.04,** 4.015

Note: The point's corresponding specificity and sensitivity are displayed in the table along with its Euclidean distance from the upper left corner. The type of column is in line with the kind of experiment that was conducted using the training data. We also provide the average DOR and F1 Scores for each approach as a point of comparison.

TABLE II. AN OVERVIEW OF EACH METHOD'S BEST OUTCOMES IN RELATION TO THE POINT ON THE ROC SPACE THAT IS CLOSEST TO THE UPPER LEFT CORNER WITH THREE FOLLOWS UPS

Technique	Type	Specificity 5, 7, 10 Folds	Sensitivity 5, 7, 10 Folds	Distance 5, 7, 10 Folds	Avg. F1 5, 7, 10 Folds	Avg. DOR 5, 7, 10 Folds
1	Blind	0.719	0.671	0.431	0.672	4.289
1	Semi-Blind	0.753	0.642	0.434	0.649	3.495
2	Blind	0.521, 0.521, 0.522	0.5, 0.5, 0.5	0.691, 0.691, 0.691	0.095, 0.096, 0.108	1.939, 1.998, 2.089
2	Semi-Blind	0.72, 0.72, 0.72	0.595, 0.595, 0.595	0.491, 0.491, 0.491	0.391, 0.369, 0.356	2.347, 2.326, 2.357
3	Blind	0.769, 0.742, 0.764	0.677, 0.675, **0.688**	0.395, 0.414, 0.39	0.684, **0.688, 0.688**	4.855, 5.045, 5.043
3	Semi-Blind	0.74, **0.84**, **0.84**	0.682, 0.645, 0.645	0.410, **0.388, 0.388**	0.685, 0.687, 0.687	5.01, **5.177**, 5.171

VII. CONCLUSION AND FUTURE WORKS

To sum up, the main goal of this study was to create a model that could be used to predict how patients with MCI would progress based only on the examination of their longitudinal MRI scans. Another goal of this study was to use MRI data to derive useful diagnostic information without the need for further diagnostic instruments like cognitive tests. Predicting whether patients with MCI would develop AD was the third goal. Three different techniques based on HMMs were developed, one building on the other. It is clear from looking at the experimental findings that Techniques 1 and 3 have generated models that work well. These techniques have produced results that are both much higher than the preset criteria for sensitivity and higher than even the harsher predefined threshold for specificity. Particularly, technique 3 performs the best, closely followed by Technique 1. Technique 3, the best classifier, can identify 77.6% of participants who advance to AD and 64.1% of people who remain stable with MCI or return to normal cognitive status. It uses a semi-blind method with 5-fold SVM training. Furthermore, the findings highlight the structural information contained in the MRI images, which provides important information on how a patient's cognitive state is developing. Notably, the first technique relies on HMMs as the primary classifier without the requirement for a secondary classifier, using only the structural and temporal information from the scans for categorization. The findings further highlight the importance of the longitudinal MRI sequences' duration. Longer sequences consistently result in greater system performance, especially those with three follow-up scans. This emphasizes how crucial it is to get more follow-up scans for every person to improve prediction accuracy. The dataset's limitations present difficulties for future investigation. A primary concern pertains to the size of the dataset, which is determined by the expense and duration of MRI scans as well as the preprocessing done with Freesurfer. One can focus on decreasing processing time, increasing the dataset with longer sequences, and optimizing preprocessing efficiency through parallelization. A flag to denote confirmed diagnoses might also be introduced to alleviate the uncertainty around diagnoses. This would enable weighting of subjects with certain diagnoses and enable semi-supervised learning. Additional preprocessing procedures, like clustering or dimensionality reduction, to simplify and improve the feature vectors might be added in the future. It may also be worthwhile to investigate more sophisticated machine learning methods, such as Convolutional Neural Networks (ConvNets), Deep Neural Networks (DeepNets), and Recurrent Neural Networks (RNNs). Without a set input length, RNNs excel at modeling sequential and temporal data. Conversely, DeepNets and ConvNets provide high precision modeling of complicated data, which may minimize the need for intensive preprocessing. ConvNets may be able to operate directly with raw MRI scans, omitting the feature extraction stage, albeit their use may necessitate much bigger datasets. Although these methods offer fascinating directions for future study, their applicability in medicine will depend on the availability of data and developments in neural network technology.

VIII. DECLARATIONS

Funding: No funds, grants, or other support was received.

Conflict of Interest: The author declare that they have no known competing for financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability: Data will be made on reasonable request.

Code Availability: Code will be made on reasonable request.

REFERENCES

[1] H. Yoo, “Genetics of Autism Spectrum Disorder: Current Status and Possible Clinical Applications,” Experimental Neurobiology, vol. 24, no. 4, pp. 257–272, Dec. 2015, doi: 10.5607/en.2015.24.4.257.

[2] K. D. Miller, M. Fidler‐Benaoudia, T. H. Keegan, H. S. Hipp, A. Jemal, and R. L. Siegel, “Cancer statistics for adolescents and young adults, 2020,” CA: A Cancer Journal for Clinicians, vol. 70, no. 6, pp. 443‐459, Nov. 2020, doi: 10.3322/caac.21637.

[3] A. A. Adegun, S. Viriri, and R. O. Ogundokun, “Deep Learning Approach for Medical Image Analysis,” Computational Intelligence and Neuroscience, vol. 2021, 2021, doi: 10.1155/2021/6215281.

[4] M. M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, and P. Vandergheynst, “Geometric Deep Learning: Going beyond Euclidean data,” IEEE Signal Processing Magazine, vol. 34, no. 4. pp. 18–42, 2017. doi: 10.1109/MSP.2017.2693418.

[5] J. Shaw, F. Rudzicz, T. Jamieson, and A. Goldfarb, “Artificial Intelligence and the Implementation Challenge,” J Med Internet Res 2019;21(7):e13659 https://www.jmir.org/2019/7/e13659, vol. 21, no. 7, p. e13659, Jul. 2019, doi: 10.2196/13659.

[6] N. Marwah, V. K. Singh, G. S. Kashyap, and S. Wazir, “An analysis of the robustness of UAV agriculture field coverage using multi-agent reinforcement learning,” International Journal of Information Technology (Singapore), vol. 15, no. 4, pp. 2317–2327, May 2023, doi: 10.1007/s41870-023-01264-0.

[7] S. Wazir, G. S. Kashyap, and P. Saxena, “MLOps: A Review,” Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.10908v1

[8] M. Kanojia, P. Kamani, G. S. Kashyap, S. Naz, S. Wazir, and A. Chauhan, "Alternative Agriculture Land-Use Transformation Pathways by Partial-Equilibrium Agricultural Sector Model: A Mathematical Approach," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.11632v1

[9] H. Habib, G. S. Kashyap, N. Tabassum, and T. Nafis, "Stock Price Prediction Using Artificial Intelligence Based on LSTM– Deep Learning Model," in Artificial Intelligence & Blockchain in Cyber Physical Systems: Technologies & Applications, CRC Press, 2023, pp. 93–99. doi: 10.1201/9781003190301-6.

[10] G. S. Kashyap, D. Mahajan, O. C. Phukan, A. Kumar, A. E. I. Brownlee, and J. Gao, "From Simulations to Reality: Enhancing Multi-Robot Exploration for Urban Search and Rescue," Nov. 2023, Accessed: Dec. 03, 2023. [Online]. Available: https://arxiv.org/abs/2311.16958v1

[11] G. S. Kashyap, K. Malik, S. Wazir, and R. Khan, "Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing," Multimedia Tools and Applications, vol. 81, no. 25, pp. 36685–36698, Oct. 2022, doi: 10.1007/s11042-021-11558-9.

[12] G. S. Kashyap, A. E. I. Brownlee, O. C. Phukan, K. Malik, and S. Wazir, "Roulette-Wheel Selection-Based PSO Algorithm for Solving the Vehicle Routing Problem with Time Windows," Jun. 2023, Accessed: Jul. 04, 2023. [Online]. Available: https://arxiv.org/abs/2306.02308v1

[13] S. Wazir, G. S. Kashyap, K. Malik, and A. E. I. Brownlee, "Predicting the Infection Level of COVID-19 Virus Using Normal Distribution-Based Approximation Model and PSO," Springer, Cham, 2023, pp. 75–91. doi: 10.1007/978-3-031-33183-1_5.

[14] Y. Chen and T. D. Pham, "Sample entropy and regularity dimension in complexity analysis of cortical surface structure in early Alzheimer's disease and aging," Journal of Neuroscience Methods, vol. 215, no. 2, pp. 210–217, May 2013, doi: 10.1016/j.jneumeth.2013.03.018.

[15] S. Duchesne, A. Caroli, C. Geroldi, D. L. Collins, and G. B. Frisoni, "Relating one-year cognitive change in mild cognitive impairment to baseline MRI features," NeuroImage, vol. 47, no. 4, pp. 1363–1370, Oct. 2009, doi: 10.1016/j.neuroimage.2009.04.023.

[16] C. Y. Wee, P. T. Yap, and D. Shen, "Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns," Human Brain Mapping, vol. 34, no. 12, pp. 3411–3425, Dec. 2013, doi: 10.1002/hbm.22156.

[17] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," NeuroImage, vol. 55, no. 3, pp. 856–867, Apr. 2011, doi: 10.1016/j.neuroimage.2011.01.008.

[18] Y. Chen and T. D. Pham, "Development of a brain MRI-based hidden Markov model for dementia recognition," BioMedical Engineering Online, vol. 12, no. SUPPL 1, pp. 1–16, Dec. 2013, doi: 10.1186/1475-925X-12-S1-S2.

[19] S. M. Resnick, D. L. Pham, M. A. Kraut, A. B. Zonderman, and C. Davatzikos, "Longitudinal magnetic resonance imaging studies of older adults: A shrinking brain," Journal of Neuroscience, vol. 23, no. 8, pp. 3295–3301, Apr. 2003, doi: 10.1523/jneurosci.23-08-03295.2003.

[20] B. Wang and T. D. Pham, "MRI-based age prediction using hidden Markov models," Journal of Neuroscience Methods, vol. 199, no. 1, pp. 140–145, Jul. 2011, doi: 10.1016/j.jneumeth.2011.04.022.

[21] G. Spulber et al., "An MRI-based index to measure the severity of Alzheimer's disease-like structural pattern in subjects with mild cognitive impairment," Journal of Internal Medicine, vol. 273, no. 4, pp. 396–409, Apr. 2013, doi: 10.1111/joim.12028.

[22] J. Trygg and S. Wold, "O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter," in Journal of Chemometrics, Jan. 2003, vol. 17, no. 1, pp. 53–64. doi: 10.1002/cem.775.

[23] S. Wold, J. Trygg, A. Berglund, and H. Antti, "Some recent developments in PLS modeling," in Chemometrics and Intelligent Laboratory Systems, Oct. 2001, vol. 58, no. 2, pp. 131–150. doi: 10.1016/S0169-7439(01)00156-3.

[24] C. Aguilar et al., "Different multivariate techniques for automated classification of MRI data in Alzheimer's disease and mild cognitive impairment," Psychiatry Research - Neuroimaging, vol. 212, no. 2, pp. 89–98, May 2013, doi: 10.1016/j.pscychresns.2012.11.005.

[25] J. Escudero, J. P. Zajicek, and E. Ifeachor, "Machine Learning classification of MRI features of Alzheimer's disease and mild cognitive impairment subjects to reduce the sample size in clinical trials," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2011, pp. 7957–7960. doi: 10.1109/IEMBS.2011.6091962.

[26] K. Katanoda, Y. Matsuda, and M. Sugishita, "A spatio-temporal regression model for the analysis of functional MRI data," NeuroImage, vol. 17, no. 3, pp. 1415–1428, Nov. 2002, doi: 10.1006/nimg.2002.1209.

[27] A. Quirós, R. M. Diez, and D. Gamerman, "Bayesian spatiotemporal model of fMRI data," NeuroImage, vol. 49, no. 1, pp. 442–456, Jan. 2010, doi: 10.1016/j.neuroimage.2009.07.047.

[28] Y. Wang, S. M. Resnick, and C. Davatzikos, "Spatio-temporal analysis of brain MRI images using hidden Markov models," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6362 LNCS, no. PART 2, pp. 160–168. doi: 10.1007/978-3-642-15745-5_20.

[29] S. G. Mueller et al., "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI)," Alzheimer's and Dementia, vol. 1, no. 1. No longer published by Elsevier, pp. 55–66, Jul. 01, 2005. doi: 10.1016/j.jalz.2005.06.003.