Integrating Traditional Machine and Deep Learning Methods for Enhanced Alzheimer's Detection from MRI Images

Shreyan Kancharla
Independent Researcher
Cary, United States of America
shreyansk6@gmail.com

***Abstract*— Alzheimer's Disease (AD) is a prominent progressive neurodegenerative disorder that causes impairments in cognition and physically affects the brain. As there is no cure for AD, early detection is pertinent to prevention and slowed progression of the disease. Current diagnostic methods involve the manual evaluation of MRI images by trained professionals. This is useful yet has limitations that could be overcome by utilization of machine learning classification models for early detection of AD, which is the aim of this research. To do this, an OASIS-3 dataset containing images of brain MRIs labeled as mild cognitive impairment and cognitively normal was split into testing, training, and validation data. The data was augmented, featurized, and trained on various algorithms. Featurization was done using the deep learning convolutional neural network ConvNextXLarge, which was pre-trained on ImageNet. The algorithms were tested on validation data and the best model was selected. MLP, KNN, and RF were models that had an accuracy of 0.979 and XGBoost had an accuracy of 0.959. MLP was selected as the final model and performed with a final accuracy of 0.953 on the testing data with a recall value of 1. The results of this study demonstrate that machine learning models can be used to aid in diagnosis of Alzheimer's disease, allowing for improved health conditions and treatment of AD.**

Keywords—Alzheimer's Disease, MRI, Machine Learning, Convolutional Neural Network, Featurization

I. Introduction

Alzheimer's Disease (AD) is a disorder that negatively affects memory and thinking and accounts for 60-80% of dementia cases. It is estimated that over six million Americans over the age of sixty-five are AD patients. Some of the main features of AD are amyloid-beta plaques, neurofibrillary/tau tangles, and loss of connections between neurons. AD impairs speaking, spatial orientation, and judgment. Additionally, Alzheimer's is correlated with an increase in age and is known to be more common in females than it is in males [1].

It is predicted that by 2050, there will be 153 million cases of dementia worldwide, compared to 57 million cases in 2019 [2]. There is no cure for this disease currently and with the increase in prevalence and mortality of AD, the need for early detection of this disease is crucial. By detecting the disease early, the progression of the disease can be delayed and stopped eventually [3]. Imaging techniques are widely used in mid to late diagnosis of AD in patients, with MRIs of the hippocampus being one of the best-established biomarkers for the disease [4]. MRI images can allow clinicians to identify shrinkage in certain regions of the brain, such as the hippocampus or the cerebral cortex, which are affected by AD.

While methods for AD diagnosis are established, machine learning applications to MRI can enable faster and more accurate detection of AD. Currently, brain MRI analysis is a tedious and time-consuming process. As this process requires manual labor, the turnover time of analyses can vary due to the volume of patients presenting with symptoms. Typically, a patient will receive the results of an MRI scan two weeks after the initial appointment [5]. Therefore, automated analysis of brain images provides a potential application of machine learning in clinical practice to reduce the time required for the process of MRI analysis. Additionally, using a machine learning model can eliminate some of the other limitations of conventional AD diagnosis as well. For example, detection done by an ML model could pick up on certain features and patterns that may not be easily discernible by human observers. Such identification of features could also be beneficial in the early detection of the disease. One major issue regarding human diagnosis is the apparent high prevalence of missed and delayed diagnoses of dementia. Factors contributing to missed or delayed diagnoses include lack of knowledge regarding dementia, personal concerns regarding misdiagnosis, and lack of feasibility of implementation of existing tools in practice [6]. These factors can be mitigated by using machine learning models, which, unlike humans, are less subjected to these limitations. This could provide better detection rates and opportunities for earlier interventions, allowing for significant improvements in patient outcomes.

The typical progression for AD ranges from mild cognitive impairment (MCI) to severe dementia. MCI is considered to be a transitional stage between normal aging and dementia [7]. Mild cognitive impairment is generally an indication of early-stage AD [8]. One physiological characteristic that occurs with mild cognitive impairment is Lewy bodies, which are clumps of protein commonly associated with Parkinson's Disease and Lewy Body Dementia, but also with Alzheimer's Disease as well. In addition to protein deposits, decreased hippocampus size and increased brain ventricle size are some

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00066

physiological markers of mild cognitive impairment [9]. Identification of MCI could provide another avenue for early detection, as it is present in the beginning stages of AD. This could be done using a binary classification machine learning model, which would classify data into two possible outcomes. In the context of early detection, a binary classification system could be trained on features extracted from MRI image data. The model can then predict whether an individual is likely to have MCI indicative of early-stage AD or not. The goal of this project is to develop an accurate and reliable binary classification model that leverages MRI image data to detect mild cognitive impairment indicative of early-stage Alzheimer's disease to provide a better method of AD detection than the current conventional diagnostic methods.

II. Methods

A. Dataset

This study utilized an open OASIS-3 dataset containing MRI images of the brain that were labeled either CN (no cognitive impairment) or MCI. This dataset contained a total of 657 MRI images. Approximately 20% of these images were grouped for testing the model, and 80% for training. In the testing group, 79 images were labeled CN, and 71 images were labeled MCI. In the training group, 168 images were labeled CN, and 339 images were labeled MCI.

B. Test-Train-Validation Split

As this study includes a comparison of various machine learning models, it is important that there is a way to evaluate the models prior to proceeding to final testing. To do this, the data was split into another category called validation. The validation data was created by randomly extracting 10% of images from data labeled CN and MCI in the training group and adding it to a new validation set.

C. Data Augmentation

Following the splitting, the number of images in the training set was 152 CN and 306 MCI. This represented a clear imbalance. Imbalance in data can result in classification models that are biased towards the majority class which result in a poor recall rate. To address this concern, data augmentation was applied. Data augmentation is a technique that uses existing data to create new data by making minor changes. To create a balanced training set, the total number of desired images in CN and MCI was set to 307 each. An Augmentor pipeline was used to apply rotation augmentation to the images with a probability of 90% and a maximum left and right rotation of 10 degrees. Random flipping augmentation was also applied to the images with a probability of 80%. The augmentation ensured that the results are more accurate by creating a balance in the previously imbalanced data set.

D. Featurization

The images in the training and validation sets were both converted into features and labels in a tabular format using a pre-trained model. The pre-trained model used is ConvNextXLarge, a convolutional neural network used for image classification, which has pre-trained weights from ImageNet. ImageNet is a large dataset, containing millions of labeled images. The use of ConvNextXLarge allows for the use of transfer learning to extract high-level features from ImageNet to allow for a more efficient and better performing model. The converted tabular data from featurization was then used to train the model.

E. Training Model and Testing on Validation Data

In this study, various algorithms were tested, and the performance of each algorithm was recorded. The algorithms used were K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), XGBoost (XGB), and Random Forest (RF), which are common for machine learning classification tasks. KNN classifies data points based on the majority class among their nearest neighbors, which is determined by a specified value indicated by *K*. MLP is a neural network that learns relationships in the data through weighted connections between neurons. RF is an ensemble learning method that uses multiple decision trees to combine their predictions. XGBoost uses a series of decision trees in sequential order and corrects mistakes made by previous trees. Each of these algorithms were trained on the training data with specified hyperparameters. They were then tested on the validation data for each of their parameterizations. The accuracies were then plotted for each algorithm and the best model for each algorithm was selected.

Methodology Flow Chart

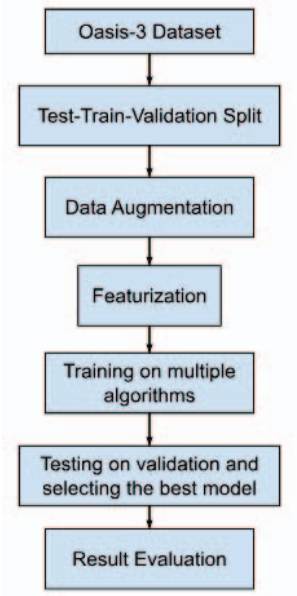

Fig. 1. Methodology Flow Chart

III. Results

A. Overview

The results from testing the various models on the validation data were considered when selecting the best model. The study found that MLP was the best model. The MLP model was then tested using the testing data to gather final results.

B. Statistical Significance

This experiment uses the confusion matrix to evaluate the performance of the models. A confusion matrix is a table that compares the labels predicted by the models to the true labels by including the number of true positives, true negatives, false positives, and false negatives. The following equations were used to calculate the accuracy, precision, recall, and F1 score:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$F1\ Score = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

C. Data

To select the most effective model for training, the best model was chosen based on accuracy scores from testing on the validation data. Multi-Layer Perceptron, Random Forest, and K-Nearest Neighbors all had accuracies of 0.979. XGBoost had the lowest accuracy of 0.959.

TABLE I

TESTING MODELS ON VALIDATION DATA

MODEL	ACCURACY	PRECISION	F1 SCORE	RECALL
Multi-Layer Perceptron	0.979	CN: 0.94 MCI: 1.00	CN: 0.97 MCI: 0.98	CN: 0.94 MCI: 1.00
XGBoost	0.959	CN: 0.89 MCI: 1.00	CN: 0.94 MCI: 0.97	CN: 1.00 MCI: 0.94
Random Forest	0.979	CN: 0.94 MCI: 1.00	CN: 0.97 MCI: 0.98	CN: 1.00 MCI: 0.97
K-Nearest Neighbors	0.979	CN: 1.00 MCI: 0.97	CN: 0.97 MCI: 0.99	CN: 1.00 MCI: 0.97

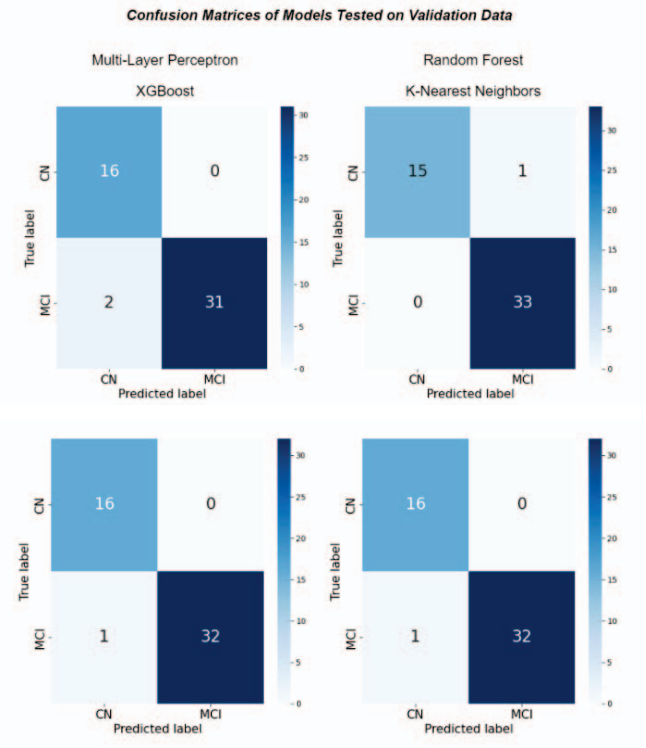

Figure 2. Confusion matrices of models tested on validation data

To select between MLP, RF, and KNN, the advantages and disadvantages of each model were considered. KNN tends to be sensitive to noise and outliers, therefore this model was not considered for the final model, despite the higher precision and F1 scores. RF and MLP are both robust, so both models were used on the testing data. Results from the test demonstrated that MLP has a higher accuracy of 0.953 than RF which had an accuracy of 0.906. Additionally, MLP had the highest recall. Therefore, MLP was selected as the final model, and a final accuracy of 0.953 was calculated.

TABLE II

MULTI-LAYER PERCEPTRON RESULTS ON TESTING DATA

Model	Accuracy	Precision	F1 Score	Recall
Multi-Layer Perceptron	0.953	CN: 0.97 MCI: 0.93	CN: 0.95 MCI: 0.95	CN: 0.94 MCI: 0.97

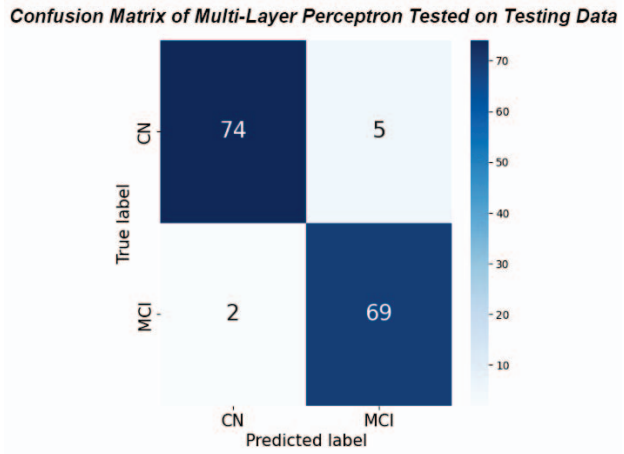

Figure 3. Confusion matrix of multi-layered perceptron tested on testing data

IV. Discussions

A. Interpretation of Results

Among the models evaluated, the MLP demonstrated the highest recall value for the MCI class, achieving a perfect score of 1.00. This indicates that the MLP model was able to correctly identify all individuals with MCI in the dataset. While other metrics such as accuracy, precision, and F1 score are important, in the context of disease detection, the ability to capture true positives (MCI cases) is paramount.

Both XGBoost and Random Forest models also performed well, with XGBoost achieving a recall value of 0.97 for MCI and Random Forest achieving a recall of 0.98 for MCI. While these models had high overall accuracy, their performance in correctly identifying individuals with MCI was slightly lower than that of the MLP model.

B. Comparison With Previous Studies

The results of this study demonstrate that an MLP binary classification model can be used to accurately identify significant indicators of early-stage AD. It has been shown that ensemble learning is an accurate model on the same OASIS-3 dataset [10]. This study shows that an individual binary classification model using transfer learning has similar accuracy.

Additionally, a comparison to other machine learning models used to classify MRI images based on characteristics of Alzheimer's disease shows that the model developed in this study is more effective than previous models. One study compared the performance of KNN, Random Forest, and linear regression. In that study, features were extracted using Gray Level Cooccurrence Matrix and Haralick Features. The accuracy of the model in that study was 84%, which is significantly lower than the accuracy of the model created in our study [11]. This comparison demonstrates that the model in our study is more effective at predicting AD characteristics from MRI images than some currently existing models with a similar goal.

To further expand on this finding, the model created in this study can be compared to other models that utilize pre-trained ConvNext features. One study was conducted by Techa et.al., in which a machine learning model was developed to classify Alzheimer's Disease using an Alzheimer's Disease Kaggle dataset containing MRI images [12]. The model in this study also utilized ConvNext for feature extraction. Both our study and the study conducted by Techa et. al utilized ConvNext as the convolutional neural network for feature extraction, however, our study specifically utilizes the ConvNextXLarge variant. When evaluated, the model performed with a 92.2% accuracy. Even among models utilizing similar methods of featurization, the model developed in our study exhibits superior accuracy.

Another similar area to investigate would be a comparison of performance between our model, which utilizes ConvNext for featurization, and a model that utilizes a different featurization technique. One study highlights a deep learning model in which DenseNet-169 and ResNet-50, two convolutional neural network architectures used for feature extraction, are used to classify Alzheimer's Disease using brain images [13]. After evaluation of the model, it was found that the DenseNet-169 architecture had an accuracy of 0.8382 on testing data, and the ResNet-50 architecture had an accuracy of 0.8192 on testing data. Based on these findings, it can be concluded that the model devised in our study has a better performance in the classification of Alzheimer's Disease compared with some existing models that utilize different featurization methods. It is important to note, however, that in our model, only two labels, CN and MCI, exist for classification. In the study conducted by Shehri, there were four labels: non-dementia, very mild-dementia, mild-dementia, and moderate dementia. As more labels make classification a more complex task for machine learning, this could potentially account for some aspects of the lower accuracy [13]. While this may present some discrepancies between the studies, the superior performance of our model demonstrates potential clinical utility in diagnosing AD in its early stages and improvement of patient outcomes.

C. Implications of Findings

The clinical diagnostic accuracy for AD, even among experts, is only 77% [14]. In fact, in a study with over 900 patients, it was found that 25% of patients were misdiagnosed [15]. Additionally, it has been previously seen that ML models overall do a better job of predicting dementia than humans. According to a review paper analyzing the predictive value of machine learning for dementia, "ML algorithms outperformed humans in predicting incident all-cause dementia within two years" [16]. This is important as it highlights the significance of the use of machine learning in identifying dementia in its early stages. The model we have created in this study can similarly increase the accuracy of diagnoses of Alzheimer's by being able to efficiently identify key features of AD pathology that may go unnoticed in standard clinical practice.

In the United States, dementia tends to be significantly underdiagnosed. Only approximately 50% of

individuals who exhibit the criteria for dementia are diagnosed with dementia by clinicians [17]. Missed diagnoses of dementia are also influenced by various factors, including race, education, and the number of people present at doctor's appointments. One study found that people of non-white race were more likely to have undiagnosed dementia. The study also found that people with at least a high-school level education were 46% less likely to have a missed diagnosis than those with lower level education. Additionally, it was found that those who visit the doctor alone are twice as likely to have a missed diagnosis than those who attend doctor's appointments with accompaniment [17]. The perfect recall value obtained in our study reflects that all cases of mild cognitive impairment were detected by the model. This demonstrates that our model has a significant potential application to clinical practice as it is not negatively influenced by the same many external factors that negatively affect a diagnosis made by human clinicians.

D. Limitations

The misclassifications seen in the confusion matrices can be explained by the possibility that some MCI MRIs did not exhibit features that were picked up by the machine. Likewise, CN images that were classified as MCI may have demonstrated some features common to MCI. Another variable that could have affected the accuracy of the model is the quality of images. If some images were of lower quality or were affected by excessive noise, important features may have been obscured. Despite these potential causes for the lower accuracy of the model, the final performance was superior.

E. Future Directions

Future experimentation could combine various models through ensemble learning. By aggregating the predictions from several models, ensemble learning allows for a reliable model that is more robust to noise. This could improve the model's accuracy and increase the detection of features indicative of early-stage AD. This was demonstrated in the transfer learning model developed by Grover et. al. Three of the best-performing models developed by their team had accuracies of 0.974, 0.974, and 0.982. Following this, the team used ensemble methods to combine the three models in various ways and the best-performing combination of models resulted in an accuracy of 0.989, which is significantly better than the individual models' performances. Similar techniques could be applied to the models created in our study to determine if combinations of these models result in a more effective classification.

Another area for improvement in the future is training the model on larger datasets to allow for more comprehensive learning of patterns and features found in the MRI images. Having a larger dataset could help prevent overfitting by offering more diverse examples that prevent the learning of noise or irrelevant patterns. The synergistic effect provided by larger datasets of better pattern recognition and less overfitting can increase the capability of the model to make accurate classifications.

V. CONCLUSION

The binary classification machine learning model created in this study used MRI image data balanced through image augmentation and applied featurization using a pre-trained model to extract features significant to the identification of indicators of Alzheimer's Disease, followed by training and testing of various models and finally selecting the best model based on results. The results from this study support the goal of the work, which was to develop a model that can accurately predict mild cognitive impairment suggesting early-stage AD. The model devised in this study performed with similar and in some cases better accuracy than other AD-detection models that utilize various convolutional neural networks, including ConvNext, ResNet, and DenseNet. This study takes the novel approach of utilizing the ConvNext variant ConvNextXLarge, a model pre-trained on ImageNet, with training, testing, and validation done on the OASIS-3 dataset. This model could be applied in real-life clinical settings to aid clinicians in the diagnosis of patients with early-stage Alzheimer's disease using MRI images for more accurate results.

REFERENCES

[1] National Institute on Aging. Alzheimer's disease and related dementias public health roadmap. NIH National Institute on Aging, 2021. Retrieved from https://www.nia.nih.gov/alzheimers/publication/2020-2023-alzheimers-disease-related-dementias-public-health-road-map

[2] The Lancet Public Health. Estimation of the global prevalance of dementia in 2019 and forecasted prevalance in 2050: an analysis for the Global Burden of Disease Study, 2019, Feb 2022, vol 7, issue 2, E105-E125. https://doi.org/10.1016/S2468-2667(21)00249-8

[3] Rasmussen, J., & Langerman, H. A neurodegenerative disease research framework: Expanding the discovery pipeline while prioritizing patient needs. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 2019, 5, 793–798. https://doi.org/10.1016/j.trci.2019.10.012

[4] Van Oostveen, W. M., & de Lange, E. C. M. Therapeutic potential of magnetic resonance imaging-guided focused ultrasound in Alzheimer's disease. Journal of Controlled Release, 2021, 329, 1118-1130. https://doi.org/10.1016/j.jconrel.2020.12.043.

[5] NHS. How its performed, MRI Scan, Reviewed July 6 2022. https://www.nhs.uk/conditions/mri-scan/what-happens/

[6] Bradford, A., Kunik, ME., Schulz, P., Williams, SP., Singh, H. Missed and delayed diagnoisis of dementia in primary care: prevalance and contributing factors. Alzheimer Dis Assoc Discord, 2009, Oct-Dec; 23(4):306-14. https://doi.org/10.1097/wad.0b013e3181a6bebc

[7] Geda, Y. E. Mild cognitive impairment in clinical practice. Mayo Clinic Proceedings, 2014, 89(10), 1452–1459. https://doi.org/10.1016/j.mayocp.2014.07.003

[8] Morris, J. C., Storandt, M., Miller, J. P., McKeel Jr, D. W., Price, J. L., Rubin, E. H., & Berg, L. Mild cognitive impairment represents early-stage Alzheimer disease. Archives of Neurology, 2001, 58(3), 397–405. https://doi.org/10.1001/archneur.58.3.397

[9] Mayo Clinic. Mild Cognitive Impairment, Feb 13, 2024. https://www.mayoclinic.org/diseases-conditions/mild-cognitive-impairment/symptoms-causes/syc-20354578#:~:text=Brain%2Dimaging%20studies%20show%20that,glucose%20in%20key%20brain%20regions.

[10] Grover, P., Chaturvedi, K. Zi, X. Saxena, A., Prakash, S., Jan, T., Prasad, M. Ensemble Transfer Learning for Distinguishing Cognitively Normal and Mild Cognitive Impairment Patients Using MRI. Algorithms, 2023, 16, 377. https://doi.org/10.3390/a16080377

[11] Uma Rani, K., Sharvari, S. S., Umesh, M. G., Vinay, B. C. "Binary Classification of Alzheimer's disease using MRI images and Support Vector Machine," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India, 2021, pp. 423-426 https://ieeexplore.ieee.org/document/9641661

[12] Techa, C., Ridouani, M., Hassouni, L., Anoun, H. Automated Alzheimer's Disease Classification from Brain MRI Scans Using ConvNeXt and Ensemble of Machine Learning Classifiers. In: Abraham, A., Hanne, T., Gandhi, N., Manghirmalani Mishra, P., Bajaj, A., Siarry, P. (eds) Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022). SoCPaR 2022. Lecture Notes in Networks and Systems, vol 648. Springer, Cham. https://doi.org/10.1007/978-3-031-27524-1_36

[13] Al Shehri W. Alzheimer's disease diagnosis and classification using deep learning techniques. PeerJ Comput Sci. 2022 Dec 20;8

[14] Sabbagh, M. N., Lue, L. F., Fayard, D., Shi, J., & Aisen, P. S. Alzheimer disease research in the 21st century: Past and current failures, new perspectives and funding priorities. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 2017, 3(3), 389–397. https://doi.org/10.1016/j.trci.2017.07.001

[15] DocPanel. Alzheimer's disease and the benefit of radiology second opinion, 2018. https://www.docpanel.com/alzheimers-disease-and-benefit-radiology-second-opinion/

[16] Javeed, A., Dallora, A.L., Berglund, J.S., Ali, A, Ali, L, Anderberg, P. Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. J Med Syst, 2023, 47, 17. https://doi.org/10.1007/s10916-023-01906-7

[17] Amjad, H., Roth, DL., Sheehan, OC., Lyketsos, CG., Wolff, JL., Samus, QM. Underdiagnosis of Dementia: an Observational Study of Patterns in Diagnosis and Awareness in US Older Adults. J Gen Intern Med. 2018 Jul;33(7):1131-1138. https://doi.org/10.1007/s11606-018-4377-y