Difficulty-Aware Diagnosis: A Two-Step Approach to Classification of Heterogeneous Datasets

Jimmy Mulder[1,2][0000−0001−9681−863X] ⋆, Toon Renssen[2][0000−0002−2609−1102], Sietske Tacoma[1][0000−0002−9662−8489], Henry Maathuis[1][0009−0002−5542−0478], Roelant Ossewaarde[1,2][0000−0002−7036−522X], and Yolande Pijnenburg[2][0000−0003−2464−1905]

[1] HU University of Applied Sciences Utrecht, Research Group Artificial Intelligence, Utrecht, The Netherlands
[2] Alzheimercenter UMC Amsterdam, The Netherlands

Abstract. In machine learning for clinical diagnosis, poor model performance can be difficult to interpret: is the issue with the algorithm, the features, or the data itself? This question is prevalent in neurodegenerative disease research, where diagnostic labels may be uncertain, clinical presentations are heterogeneous, and patient features often overlap across conditions. In this study, we use an ensemble of classifiers to probe the internal structure of a multi-class diagnostic dataset. By examining patterns of disagreement across models and folds, we identify a distinct subgroup of individuals that are consistently misclassified with regards to their diagnosis — suggesting these cases may be inherently difficult due to label ambiguity or clinical complexity. We build on this insight with a two-stage classification pipeline. A triage model first predicts whether a case is likely to be "easy" or "difficult" to classify, based on demographic and clinical data. Easy cases are routed through a standard multiclass model; difficult cases can be flagged for further testing or more specialized pipelines. Our approach improves accuracy for a large group of patients while referring patients with a diagnosis that is difficult to make in a cost-effective way. It also provides a general strategy for managing heterogeneous datasets without requiring manual relabeling or task-specific heuristics.

Keywords: Ensemble Learning · Multi-Stage Models · Mixture of Experts · Medical Data

1 Introduction

Accurate diagnosis of neurodegenerative diseases such as Alzheimer's Disease (AD) and Frontotemporal Dementia (FTD) remains a major challenge in clinical neurology [7], particularly in the early stages of disease progression [4]. Diagnosis

⋆ Corresponding author: `jimmy.mulder@hu.nl`

is even more complicated for atypical presentations within the spectrum of these diseases [10].

Compared to typical AD, the atypical phenotypes associated with dementia have a much lower prevalence, often occur at a younger age, and present with different cognitive deficits where memory is relatively spared [10]. Within the AD spectrum a visual variant and a language variant are recognized [2, 8], and within the FTD spectrum there are two language variants and a behavioral variant [8, 16]. To aid the clinician in the diagnostic decision making process, clinical criteria have been formulated for every phenotype. Although each of these criteria lists specific cognitive deficits as evidence, neuropsychology –the trade of assessing cognitive strenghts and weaknesses– has difficulties objectifying these specific profiles in its traditional methods [11].

Machine learning offers powerful tools for advancing this diagnostic process and disease subtyping [21, 14], especially when faced with high-dimensional data such as neuroimaging, cognitive performance scores, and clinical metadata [20, 17]. However, traditional approaches often assume that training datasets are reasonably clean and that class boundaries are well-defined. In the reality of clinical practice, these assumptions frequently break down: labels may be noisy due to diagnostic uncertainty, feature distributions may differ across subpopulations, and some patient profiles may fall between or outside canonical diagnostic categories. Analysis of our dataset showed that the datapoints were not separable by any linear approach, such as principal component analysis, regression or linear SVM's. Traditional classification models were stuck at barely above-chance performance. This inspired us to take an alternative approach.

By analyzing misclassifications across an ensemble of diverse machine learning models, we identify which individuals are consistently misclassified [3]. These cases likely represent the most heterogeneous or diagnostically ambiguous individuals. The underlying assumption is that some patients are inherently harder to classify than others, regardless of which model is used.

Building on this idea, we train a separate model to predict whether an individual's diagnosis is likely to be easy or difficult to classify. This allows us to implement a triage strategy: easy cases are routed through a high-confidence diagnostic model, while harder cases can be flagged for further testing, additional data collection, or more specialized models.

2 Methods

Our research consists of three components: an ensemble-based baseline model for understanding dataset difficulty, followed by a two-stage diagnostic pipeline consisting of a triage model for predicting classification confidence, and a final model specialized in confidently classifiable cases. This section details the architecture, training setup, and rationale for each component.

2.1 Dataset and Preprocessing

The data used in this study is clinical data of 2,012 patients who visited the Alzheimer Center of Amsterdam University Medical Center between 2010 and 2022 [5]. Within this dataset six clinical phenotypes of dementia are recognized, with marked differences in prevalence that reflect their occurrence in routine clinical settings. The phenotypes are typical (probable) Alzheimer's Disease (AD, $n = 1372$ [13]), behavioral variant of frontotemporal dementia (bvFTD, n=317 [16]) semantic variant of Primary Progressive Aphasia (svPPA , $n = 118$ [8]), posterior cortical atrophy (PCA, $n = 105$ [2]), logopenic variant of Primary Progressive Aphasia (lvPPA, $n = 57$ [8]), and the nonfluent variant of Primary Progressive Aphasia (nfvPPA, $n = 43$ [8]).

Data is collected in the following way: during a one day in-clinic visit a patient receives a clinical interview, a separate informant-based history, a neurological assessment, an Electro-Encephalogram, a Magnetic Resonance Imaging scan, analysis of blood and cerebrospinal fluid, and a neuropsychological assessment (NPA). A diagnosis is then given in a multidisciplinary consensus meeting where the results of the various examinations are discussed [18]. For this paper, only the NPA was used.

The neuropsychological assessment consists of a one hour standard paradigm with 11 neuropsychological tests to gauge the cognitive functions of an individual within various cognitive domains (which are memory, attention, language, visuospatial, and executive functioning) [1]. These 11 tests produce 44 columns of raw neuropsychological data. Next to these cognitive outcome measures we included demographic data (i.e., age, gender and education level) for each sample. Intra-class variability and diagnostic uncertainty are known issues, making the dataset inherently heterogeneous. We normalize all numerical features and encode categorical variables.

Only those patients who gave informed consent are included in our sample. The protocol for this study has been submitted and accepted by the Medical Ethical Board of our hospital.

2.2 Ensemble Classifier for Difficulty Profiling

To quantify classification difficulty at the sample level, we constructed an ensemble of six diverse machine learning models:

- Logistic Regression$_1$
- Linear Support Vector Classifier$_1$
- Random Forest (300 trees)$_2$
- Gradient Boosting$_2$
- k-Nearest Neighbors (k=15)$_3$
- Multilayer Perceptron$_4$

These models were selected for their computational efficiency, ease of implementation, and complementary modeling characteristics. Together, they represent a diverse set of algorithmic families: linear models$_1$, tree-based methods$_2$,

instance-based learning$_3$ and neural networks$_4$. This diversity helps reduce individual model biases by incorporating varied decision-making strategies across the ensemble. Moreover, all models are readily available in the standard machine learning library scikit-learn [15] and can be trained within reasonable time on a typical workstation, making them suitable for rapid prototyping and evaluation.

To address the class imbalance, the scikit-learn parameter `class weight= "balanced"` was used in any models that supported this, which automatically adjusts the loss function weights to be inversely proportional to class frequencies in the input data. Oversampling of minority classes was considered, but with such small sample sizes the risk of overfitting was deemed too great [9].

Each model was trained and evaluated using 5-fold stratified cross-validation, repeated five times to increase reliability. This procedure produced a total of 30 predictions per sample. Based on these, we counted the number of times each patient's diagnosis was misclassified. This yielded a distribution of misclassification frequency across the dataset. We computed the accuracy and F1 scores of each model separately, as well as the scores of the ensemble as a whole (using majority voting).

Data is processed in two separate stages. To avoid influence of training labels across the stages (leakage), difficulty labels were constructed exclusively from predictions on held-out test data during cross-validation. This procedure ensures that the second stage model was trained on labels uninfluenced by the first model's training data.

2.3 First stage: Triage Model: Predicting Case Difficulty

The first stage of our diagnostic pipeline involves predicting, based solely on demographic and neuropsychological features, whether a patient is likely to be *easy* or *hard* to classify. We refer to this as the triage model, and evaluate two variants:

Diagnosis-Agnostic Triage This version of the model excludes the diagnosis label entirely. Since the diagnosis is unknown for unseen individuals, and is ultimately what the second stage of the pipeline should produce, using it as a feature in the first stage would create a chicken-egg problem.

Binary-Diagnosis-Aware Triage In the second variant, we include a simplified diagnostic cue: whether the true diagnosis label was Alzheimer's Disease (AD) or not. This is implemented as a binary feature `is_AD`. This addition is motivated by the clinical intuition that practitioners often make a rough early judgment of whether a case fits an AD profile or represents an atypical or non-AD pathology. Including this feature significantly improved performance while remaining defensible from a practical standpoint. However, we acknowledge that it remains a form of label-as-a-feature, and are only using it to demonstrate the potential of future improvements.

Both triage variants were implemented in a Random Forest classifier which was trained and evaluated using 5-fold stratified cross-validation. The definition of *easy* and *hard* to classify cases was varied by adjusting the threshold t, representing the number of allowed misclassifications m out of 30 predictions. Cases were labeled *easy* if $m \leq t$, and *hard* if $m > t$. Thresholds of $t = 2, 4, 6, 8, 10, 15, 20, 25$ were examined. For each of these we create a precision/recall curve. The model decision bias that produced the highest F1 score was then used to make a final *easy/hard* prediction, used in the next stage.

2.4 Second stage: Classifier for Easy Cases

The second and final stage of our pipeline is a classifier trained exclusively on cases predicted to be *easy* by the triage model. These patients are expected to match familiar diagnostic profiles, producing a more homogeneous dataset that can be used to train a model with higher accuracy. We used a Random Forest classifier for this model because it showed competitive performance while offering lower model complexity and greater interpretability compared to gradient boosting methods.

Training and evaluation were done using 5-fold stratified cross-validation. Only samples predicted as *easy* by the triage model were included. This simulates a real-world deployment scenario where uncertain samples are routed to more intensive diagnostic processes, while confident cases are handled by automated tools.

This two-stage system enables nuanced handling of uncertainty, improves interpretability, and aligns with practical clinical workflows.

3 Results

3.1 Ensemble Model

Table 1 summarizes the accuracy and F1 scores of the individual models and the ensemble on the entire dataset. The ensemble model outperformed all individual classifiers with an overall accuracy of 73.8% and an F1 score of 0.712, although Gradient Boosting performed almost equally. It should be noted that due to the class imbalance, an accuracy score of 68% is trivially acquired by solely predicting the majority class. Most individual models barely perform higher than this.

Figure 1 shows the misclassifications made by the ensemble in two ways. The histogram shows how often each case was misclassified, color coded for their diagnosis. This clearly shows a sizable group of cases that consistently defy accurate classification at $m = 30$. These cases were manually reviewed to ensure their labels were accurate. It also shows that the ensemble struggles more with the minority classes, which is to be expected. For example, patients with lvPPA or nfvPPA rarely get misclassified fewer than 15 times. On the other hand, we observe that typical AD cases are usually considered easier, and make up the vast majority of the easiest cases (where $m = 0$).

Table 1: Performance of individual models and ensemble (majority vote)

Model	**Accuracy**	**F1 Score**
Majority Class Baseline	0.6819	-
Logistic Regression	0.5271	0.5808
Random Forest	0.7243	0.6604
Gradient Boosting	0.7314	0.6995
KNN	0.7036	0.6724
Linear SVC	0.6709	0.6877
MLP	0.6787	0.6693
Ensemble (majority vote)	**0.7377**	**0.7119**

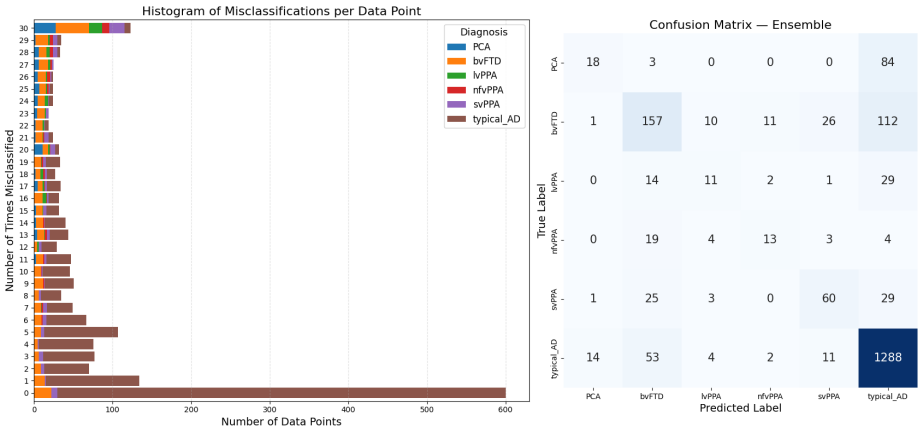
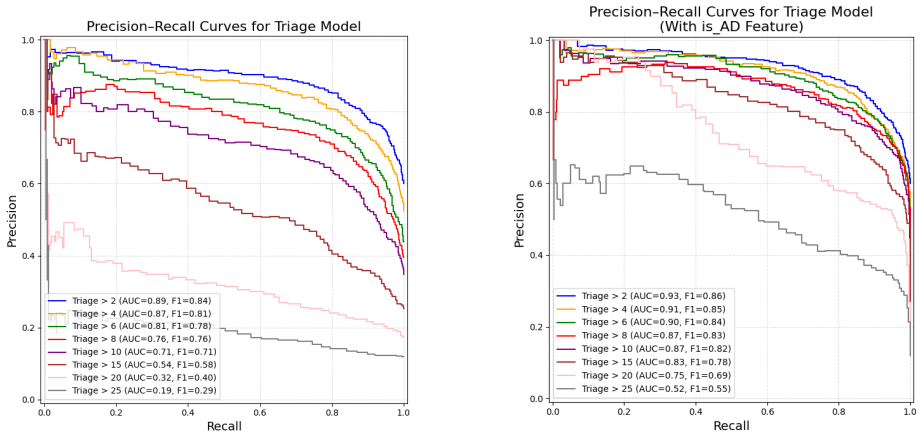


Fig. 1: *Left*: A histogram showing the number of misclassifications per sample by the ensemble (consisting of 30 classifiers). Right: A confusion matrix of the ensemble model, using majority vote.

The confusion matrix shows the kind of mistakes the ensemble makes. For example, it shows that PCA cases are classified as typical AD cases in most cases. The majority class is clearly quite dominant.

3.2 Triage Model

The precision/recall curves (cf. Fig. 2) indicate a similar degree of sensitivity regardless of the inclusion of the *is_AD* variable; a higher precision overall when the *is_AD* variable is included; and a notable performance difference when the easy/difficult threshold requires agreement of more than 15 models.



(a) Without is_AD Feature (b) With is_AD Feature

Fig. 2: Precision–Recall Curves of the Triage Model for Eight Misclassification Thresholds. Lower thresholds have a higher Area Under the Curve (AUC) and better F1 scores.

3.3 Easy Case Diagnostic Model

To demonstrate the potential benefits of triaging, we trained a separate multi-class model exclusively on cases that the triage model predicted as *easy*. This simulates a possible clinical workflow where a simpler model is deployed for confident cases, while harder cases are referred for further analysis.

The results of this model are displayed in Figure 3, for which the raw data can be found in the Appendix. It shows that as the triage model is trained to pass more cases on to the diagnostic model (by setting a higher threshold), coverage increases at the cost of performance. The impact of giving a diagnostic cue is highest for high thresholds. Compared to the ensemble performance reported in Table 1 (74% accuracy, F1 of 0.71), the diagnostic models in the two-stage pipeline perform higher for most thresholds, even without diagnostic cues.

4 Conclusion

In this work, we present a potential strategy for improving classification performance and interpretability in clinically heterogeneous datasets. Our approach centers on the idea that not all errors stem from model limitations — some may instead reflect ambiguity or inconsistency in the data itself. By leveraging an ensemble of diverse classifiers, we were able to identify a subset of patients whose diagnostic label could be predicted reliably and consistently across models. This observation enabled a principled division of the dataset into *easy* and *hard* cases.

Our ensemble-based assessment of difficulty uncovered a latent structure in the dataset that had not been visible through standard modeling techniques. We

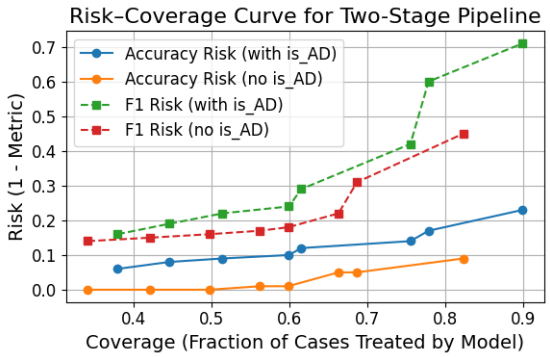


Fig. 3: A risk-coverage graph showing the relationship between the number of included data points (based on how we defined *easy* and *hard* in relation to ensemble votes) and the performance of the model on those selected data points. As coverage increases, so does risk.

used the consistency of ensemble predictions as a lens to probe the quality of data-label relationships across individual samples.

We developed meta-classifiers that, depending on the definition of *easy/hard*, can accurately predict whether a new case is likely to belong to the easy or hard group, based only on demographic and standard clinical features. This classifier serves as a triage mechanism, directing confident cases to a lightweight multi-class model and flagging ambiguous cases for further analysis. This structure mirrors real-world clinical workflows, where certain cases are quickly diagnosed based on clear patterns, while others prompt more intensive testing and expert consultation.

An alternative to our two-stage pipeline would be to use a single classifier that outputs both the predicted diagnosis and an associated confidence score, deferring [12] or rejecting [6] cases based on that confidence. We opted for a dedicated triage model as a separate stage, trained specifically to distinguish between easy and hard cases based on patterns of ensemble misclassification. This approach externalizes the notion of "confidence" as a separate supervised task. In clinical practice, it allows using different types of data for the two stages. For example, we foresee the application of informal clinical observations in the first stage but not in the second stage. Future work will compare standard defer/reject framworks to our two-stage solution.

The final classifiers trained exclusively on predicted-easy cases achieved relatively high F1 scores. While this is to be expected from a model trained on an easier dataset (which is also heavily unbalanced towards AD), this supports the notion that model performance might be improved not only by engineering better algorithms, but by acknowledging and adapting to internal heterogeneity in the data.

This framework is particularly well suited to clinical applications, where label noise, overlapping syndromes, and atypical presentations are common. By explicitly modeling confidence and routing decisions accordingly, we increase transparency and reduce the risk of overconfidence in automated predictions.

More broadly, our method offers a general strategy for incorporating uncertainty estimation, meta-prediction, and structural dataset analysis into classification pipelines. It is model-agnostic, does not require external annotations, and can be applied to a wide variety of real-world datasets where heterogeneity is expected or suspected.

5 Discussion

Neurodegenerative diseases such as AD and FTD and their rare, atypical phenotypes present with overlapping symptoms and have variable progression, resulting in frequent diagnostic uncertainty. These factors contribute to internal heterogeneity and labeling noise, which in turn limit the ceiling of performance achievable by a single unified model.

Our findings show that ensemble disagreement can be used for dataset introspection. By analyzing ensemble disagreement, we identified cases that consistently defied accurate classification across multiple models and cross-validation splits. These "difficult" cases likely stem from a combination of subtle feature patterns, borderline diagnostic categories, or data quality issues. We contrast a single model trained on both clean and noisy examples with a triage classifier to predict the likely difficulty of new samples based on demographic and feature information. This classifier, trained independently of the diagnostic task, effectively separates cases that benefit from automated classification from those that require additional scrutiny. Conceptually, this means splitting a heterogeneous dataset into one more homogeneous subset and one super-heterogeneous one.

The results show that both stages of the pipeline perform better when the threshold for *easy* is set low. While this is understandable from an algorithmic perspective, it is not always clinically desirable — one of the goals of this study is to reduce unnecessary escalations and avoid potentially expensive or invasive follow-up testing. This trade-off reflects a known limitation of standard precision/recall optimization: typical implementations do not account for the relative costs of false positives versus false negatives [19]. However, our pipeline provides clinicians with tools to set their own priorities. By adjusting thresholds, model bias, and class weights, the system can be tuned to favor either more conservative or more aggressive escalation policies. In this way, our framework aligns closely with clinical priorities, where minimizing incorrect diagnoses is often more important than maximizing accuracy.

It is important to note that the accuracy of the second-stage diagnostic model cannot be directly compared to that of the full ensemble model. First, the second-stage model is only evaluated on cases predicted to be *easy*, and we do not assess performance on the *hard* group, which may have significantly different characteristics and lower predictability. Second, in the proposed two-stage pipeline,

overall performance depends not only on the diagnostic accuracy within each group, but also on the triage model's ability to correctly assign cases to the appropriate subgroup. In other words, the final pipeline accuracy is a cumulative outcome of both triage and classification performance, making isolated comparisons potentially misleading.

This triage mechanism does not attempt to "boost" performance in the traditional ensemble sense. In contrast to methods such as XGBoost, which combine multiple weak learners in an additive, error-correcting sequence, our approach routes individual samples through different processing pipelines based on predicted difficulty. Each sample is processed only once, either through the main classification model (for *easy* cases) or deferred for alternative strategies. This behavior is more akin to a conditional mixture-of-experts system, where a gating function selects which model (or subsequent decision step) is responsible for a given input.

The inclusion of a simple binary feature (AD vs. non-AD) further improved performance of the triage classifier. This aligns with clinical practice, where a general impression of whether a case fits a typical Alzheimer's profile or not can guide further decision-making. There are also strong biomarkers for AD which could be included in follow-up research. We have also shown that clinicians can tweak the model parameters to either minimize false negatives or false positives, allowing them to make their own cost-benefit analyses.

In summary, we introduce a diagnostic pipeline that aims to accommodate the inherent heterogeneity found in clinical datasets. By separating confidently classifiable cases from more complex ones, our method allows for tailored handling of different patient profiles. While the current results are encouraging, they are exploratory in nature. Further work is needed to quantify the full benefits of this approach, particularly in real-world clinical settings and when compared directly to conventional models. Nonetheless, this study highlights a promising direction: structuring diagnostic workflows based on case difficulty, with the potential to improve both performance and interpretability.

Further research may focus on different stages of the pipeline. We intend to amplify the performance of the triage model using additional data sources, such as the notes and tests done by the patient's general practitioner. Ideally, this will increase performance enough that the AD / non-AD feature is no longer necessary. Clinical research may focus on the feature differences between the easy and hard group, identifying which neuropsychological tests have the most explanatory power, and a gap analysis to determine which additional tests would be best suited to further diagnose the hard group.

References

[1] Pauline Aalten et al. "The Dutch Parelsnoer Institute–Neurodegenerative diseases; methods, design and baseline results". en. In: *BMC Neurol.* 14.1 (Dec. 2014), p. 254.

[2] Sebastian J Crutch et al. "Consensus classification of posterior cortical atrophy". en. In: *Alzheimers. Dement.* 13.8 (Aug. 2017), pp. 870–884.

[3] Thomas G. Dietterich. "Ensemble Methods in Machine Learning". In: *Multiple Classifier Systems (MCS).* Vol. 1857. Lecture Notes in Computer Science. Springer, 2000, pp. 1–15. DOI: 10.1007/3-540-45014-9_1.

[4] Jay L.P. Fieldhouse et al. "Clinical Phenotypes of Behavioral Variant Frontotemporal Dementia by Age at Onset". In: *Journal of Alzheimer's Disease* 82.1 (2021). PMID: 34024833, pp. 381–390. DOI: 10.3233/JAD-210179. eprint: https://journals.sagepub.com/doi/pdf/10.3233/JAD-210179. URL: https://journals.sagepub.com/doi/abs/10.3233/JAD-210179.

[5] Wiesje M. van der Flier and Philip Scheltens. "Amsterdam Dementia Cohort: Performing Research to Optimize Care". In: *Journal of Alzheimer's Disease* 62.3 (2018), pp. 1091–1111. DOI: 10.3233/JAD-170850.

[6] Yonatan Geifman and Ran El-Yaniv. "SelectiveNet: A Deep Neural Network with an Integrated Reject Option". In: *Proceedings of the 36th International Conference on Machine Learning (ICML).* Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 2151–2159. URL: https://proceedings.mlr.press/v97/geifman19a.html.

[7] Abbott Gifford et al. "Biomarkers in frontotemporal dementia: Current landscape and future directions". In: *Biomarkers in Neuropsychiatry* 8 (2023), p. 100065. ISSN: 2666-1446. DOI: https://doi.org/10.1016/j.bionps.2023.100065. URL: https://www.sciencedirect.com/science/article/pii/S2666144623000059.

[8] M L Gorno-Tempini et al. "Classification of primary progressive aphasia and its variants". en. In: *Neurology* 76.11 (Mar. 2011), pp. 1006–1014.

[9] Haibo He and Edwardo A Garcia. "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[10] Danielle Jones, Victoria S. Pelak, and Emily Rogalski. "Atypical Presentations of Alzheimer Disease". In: *Continuum (Minneapolis, Minn.)* 30.6 (Dec. 2024), pp. 1614–1641. DOI: 10.1212/CON.0000000000001504.

[11] Muriel Deutsch Lezak et al. *Neuropsychological Assessment.* 5th ed. New York, NY: Oxford University Press, 2012.

[12] David Madras, Toniann Pitassi, and Richard Zemel. "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer". In: *Advances in Neural Information Processing Systems (NeurIPS) 31.* 2018, pp. 6527–6538. URL: https://proceedings.neurips.cc/paper/7853-predict-responsibly-improving-fairness-and-accuracy-by-learning-to-defer.pdf.

[13] Guy M. McKhann et al. "The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease".

In: *Alzheimer's & Dementia* 7.3 (2011), pp. 263–269. DOI: 10.1016/j.jalz.2011.03.005.

[14] Amelie Metz et al. "Frontotemporal dementia subtyping using machine learning, multivariate statistics and neuroimaging". In: *Brain Communications* 7.1 (2025). DOI: 10.1093/braincomms/fcaf065.

[15] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[16] Katya Rascovsky et al. "Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia". In: *Brain* 134.9 (2011), pp. 2456–2477. DOI: 10.1093/brain/awr179.

[17] Manash Sarma and Subarna Chatterjee. "Machine Learning-Based Alzheimer's Disease Stage Diagnosis Utilizing Blood Gene Expression and Clinical Data: A Comparative Investigation". In: *Diagnostics* 15.2 (2025). DOI: 10.3390/diagnostics15020211.

[18] Ben Schmand, Piet Eikelenboom, and Willem A. van Gool. "Value of neuropsychological tests, neuroimaging, and biomarkers for diagnosing Alzheimer's disease in younger and older age cohorts". In: *Journal of the American Geriatrics Society* 59.9 (2011), pp. 1705–1710. DOI: 10.1111/j.1532-5415.2011.03539.x.

[19] Gaël Varoquaux. "Evaluating Machine Learning Models and Their Diagnostic Utility". In: *Machine Learning for Biomedical Applications*. National Center for Biotechnology Information (US), 2023. URL: https://www.ncbi.nlm.nih.gov/books/NBK597473/.

[20] Y. Wang and other coauthors. "Understanding machine learning applications in dementia research and clinical practice: a review for biomedical scientists and clinicians". In: *Alzheimer's Research & Therapy* 16 (2024), p. 175. DOI: 10.1186/s13195-024-01540-6.

[21] Zhijian Yang et al. "A deep learning framework identifies dimensional representations of Alzheimer's Disease from brain structure". In: *Nature Communications* 12.1 (2021), p. 7065. DOI: 10.1038/s41467-021-26703-z.

Appendix: Performance of the Two-Stage Pipeline

Table 2: Performance of the two-stage pipeline. The second column reports how many data points were labeled *easy* before training the triage model. The triage (i.e. first stage) F1 score is taken from the PR curves in Figure 2. The diagnostic model (second stage) then trains and evaluates on only those cases classified as *easy* in the first stage—the actual number of cases can vary, as shown in Figure 3, due to inaccuracies in the triage model. We report the accuracy and F1 scores of those models, averaged across five folds, with standard deviation in parentheses. The value on the left excludes the binary `is_AD` feature; the value on the right includes it.

Threshold	#Easy (True)	#Hard	Triage F1 Score	Diagnostic Accuracy	Diagnostic F1 Score
2	804	1208	0.84(+0.01)/0.86(+0.01)	0.94(+0.00)/1.00(+0.00)	0.91(+0.01)/0.99(+0.00)
4	957	1055	0.81(+0.01)/0.85(+0.01)	0.92(+0.01)/1.00(+0.00)	0.88(+0.01)/1.00(+0.00)
6	1131	881	0.78(+0.02)/0.84(+0.02)	0.91(+0.01)/1.00(+0.00)	0.87(+0.01)/1.00(+0.00)
8	1215	797	0.76(+0.02)/0.83(+0.02)	0.90(+0.01)/0.99(+0.00)	0.85(+0.01)/0.98(+0.00)
10	1312	700	0.71(+0.02)/0.82(+0.02)	0.88(+0.01)/0.99(+0.00)	0.83(+0.01)/0.98(+0.00)
15	1504	508	0.58(+0.02)/0.78(+0.03)	0.86(+0.01)/0.95(+0.01)	0.80(+0.01)/0.94(+0.01)
20	1662	350	0.40(+0.03)/0.69(+0.02)	0.83(+0.01)/0.95(+0.01)	0.76(+0.01)/0.93(+0.01)
25	1772	240	0.29(+0.02)/0.55(+0.04)	0.77(+0.01)/0.91(+0.01)	0.70(+0.01)/0.89(+0.01)