Mitigating dataset bias in Alzheimer's Disease diagnostic prediction

Wendy Zhang
Queen's University
zhang.wendy@queensu.ca

Noah Lin
Queen's University
noah.lin@queensu.ca

Salma Elsayed
Queen's University
salma.elsayed@queensu.ca

Shaheen Khan
Queen's University
20sak19@queensu.ca

***Abstract*—In recent years, artificial intelligence (AI) is being increasingly explored for enhancing the efficiency of diagnosis and prediction of Alzheimer's disease (AD) through the analysis of neuroimaging data such as magnetic resonance imaging (MRI) [1]. However, many predictive models are developed using a limited number of datasets, raising concerns regarding dataset bias, representativeness, and the generalizability of model performance across diverse populations. This project examines the current landscape of AI applications in AD diagnosis through a narrative review. Simultaneously, a prototype machine learning model was developed using MRI image classification to predict AD, providing experimental results for a fine-tuned ResNet-34 architecture. Despite demonstrating high accuracy levels, the findings highlight the ethical challenges associated with deploying AI in real-world environments. To address these findings, we propose an ethical framework grounded in the principles of justice, non-maleficence, transparency, and accountability, aimed to promote equity and responsible development in AL-based diagnostic tools for AD.**

I. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder affecting over 30 million individuals globally, with more than 41 million dementia cases remaining undiagnosed [2],[3]. As populations age, the need for early and accurate detection has intensified, particularly at the stage of mild cognitive impairment (MCI), where intervention may slow disease progression.

Advances in artificial intelligence (AI), particularly deep learning applied to neuroimaging, have significantly improved predictive modeling in AD. Models trained on multimodal data, including MRI, PET, cerebrospinal fluid (CSF) biomarkers, and clinical variables, have demonstrated high classification accuracy. Multimodal deep learning approaches further enhance performance by integrating complementary data sources, outperforming single-modality models [4].

However, these advances raise important methodological and ethical concerns. Many models rely heavily on benchmark datasets such as Alzheimer's Disease Neuroimaging Initiative (ADNI), a longitudinal multimodal dataset widely used for training and evaluation [6]. This reliance introduces concerns regarding demographic representation, external validity, and real-world generalizability [6]. As a result, models may perform inconsistently across populations and risk reinforcing existing health inequities [7].

Despite rapid progress in model development, issues of dataset representation and equity remain underexplored. If training data lack diversity or reflect narrow cohorts, predictive systems may reproduce structural biases and limit clinical reliability. Examining dataset composition and development practices is therefore essential to ensure equitable and generalizable AI-based AD diagnosis.

A. Motivation and Problem Statement

MRI is a primary data source in AI-based Alzheimer's disease (AD) diagnosis, providing detailed structural information that enables detection of features such as hippocampal atrophy [8]. Machine learning (ML) models can identify subtle neurodegenerative patterns that may not be evident through traditional diagnostic approaches.

However, neuroimaging markers vary across racial and ethnic groups, while minoritized populations remain underrepresented in AD research [9],[10]. U.S.-based studies report samples composed predominantly of White participants ($\tilde{8}$8–89%), with Black/African American and Hispanic/Latino individuals comprising only 7% and 3%, respectively [9]. These imbalances limit generalizability and increase the risk of biased model performance [10].

Given that MRI datasets underpin model training and benchmarking, examining their composition and use is critical for identifying sources of inequity. This study therefore evaluates whether high reported accuracies can be replicated using publicly accessible datasets and standard modeling approaches, assessing whether performance reflects generalizable methods or controlled research conditions.

Importantly, overall accuracy may obscure disparities across demographic groups, masking differences in error rates and limiting the detection of inequitable outcomes. Addressing dataset inequities is therefore essential to ensure that AI diagnostic systems are equitable, clinically reliable, and do not reinforce existing healthcare disparities.

B. Contributions

Through a narrative review, this project examines the current landscape of artificial intelligence applications in Alzheimer's disease diagnosis, with particular attention to the datasets and modeling approaches commonly used in predictive systems. Simultaneously, a prototype machine learning model will be developed using MRI image classification to predict stages of Alzheimer's disease, allowing comparison between experimental model performance and trends identified in the litera-

ture. Finally, insights from both the literature review and the technical model will be used to develop an ethical framework aimed at promoting equitable AI use in Alzheimer's disease diagnosis.

II. Methodology

A. Literature Search

A literature search (Figure 2) was conducted in MEDLINE (Ovid), PubMed, and Web of Science in January 2026 to identify studies examining the application of artificial intelligence to magnetic resonance imaging (MRI) for Alzheimer's disease. A total of 207 records were identified through database searches, including MEDLINE ($n = 95$), PubMed ($n = 79$), and Web of Science ($n = 33$). After removal of 87 duplicate records, 120 studies remained for screening. Titles and abstracts were screened to assess relevance to artificial intelligence applications for Alzheimer's disease diagnosis using magnetic resonance imaging (MRI). Nine records were excluded during screening for the following reasons: not related to Alzheimer's disease ($n = 3$), not MRI-related ($n = 3$), not artificial intelligence-related ($n = 1$), or otherwise irrelevant to the research objectives ($n = 2$). A total of 111 studies were included in the final narrative review.

In MEDLINE (Ovid), the search combined the exploded subject heading *Alzheimer Disease* with the subheadings *classification* and *diagnostic imaging*, alongside keyword searches for *artificial intelligence* and *magnetic resonance imaging* using the multi-purpose (.mp.) field across titles, abstracts, and subject headings. Boolean operators were used to combine search concepts.

In PubMed, equivalent searches were performed using Medical Subject Headings (MeSH) for *Alzheimer Disease* with the subheadings *classification* and *diagnostic imaging*, combined with title and abstract keywords for *artificial intelligence* and *magnetic resonance imaging*.

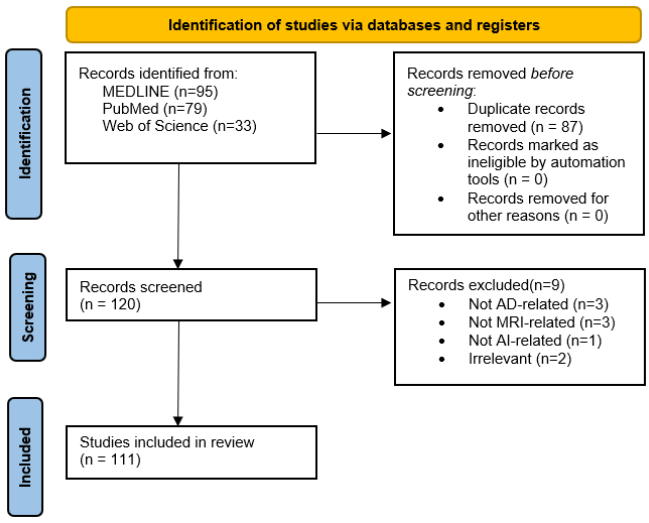

Fig. 1. PRISMA flowchart of the literature search.

B. Technical Development

1) Dataset Used: Datasets were selected based on accessibility and data quality. While commonly used Alzheimer's datasets such as ADNI and OASIS offer comprehensive imaging data, their access requirements involve extensive application procedures. Therefore, publicly available datasets were prioritized to enable timely model development.

Links to datasets:

https://huggingface.co/datasets/Falah/Alzheimer_MRI

https://huggingface.co/datasets/SilpaCS/Alzheimer

The evaluation set comprised 7,680 MRI images drawn from two independent sources: the held-out test split of Falah/Alzheimer_MRI (1,280 images), on which the model was not trained, and the entirety of SilpaCS/Alzheimer (6,400 images), a separate dataset not used during model development.

The Falah dataset is an MRI image classification dataset consisting of 6,400 images distributed across an 80/20 train–test split. As a result, only 1,280 of the total images were used for evaluation. Each image is annotated with one of four diagnostic labels. The dataset reflects a realistic clinical class imbalance, with Non Demented and Very Mild Demented cases visually dominant in the data viewer, while Moderate Demented represents the rarest and most severe category. Images are stored in Parquet format and are directly compatible with the Hugging Face `datasets` library, making the dataset straightforward to integrate into standard deep learning pipelines. The dataset is published under an Apache 2.0 license and has been used across multiple model architectures, including ResNet, Vision Transformer, and EfficientNet variants, with over 1,373 downloads recorded in its most recent month.

The SilpaCS dataset is a separate and unrelated dataset. Because it was not used during model training, all 6,400 images were treated as fully unseen data and incorporated into the evaluation set. These images follow a near-identical format.

2) Data Proprocessing: Images are processed using standard ResNet-34 preprocessing: images are resized to 224×224 pixels and normalized using ImageNet channel statistics. Training used a batch size of 16 with gradient accumulation over four steps, giving an effective batch size of 64. A fixed random seed of 42 was used, implying deterministic train–test splitting.

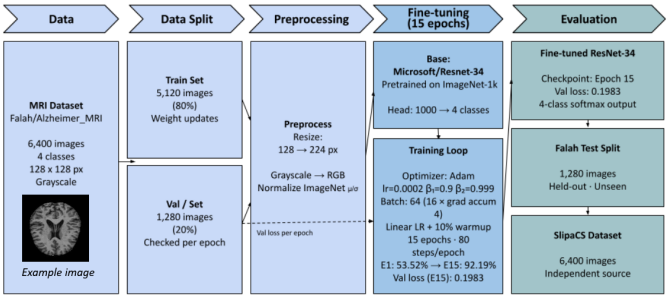

Fig. 2. Pipeline flowchart of the model.

3) Model Architecture: Figure 2 demonstrates the flowchart of the model architecture. Credited to Thamer Smadi, the model is a fine-tuned version of ResNet-34, a 34-layer Residual Network with 21.3 million parameters stored in F32 precision via the Safetensors format. ResNet-34 uses a series of residual blocks with skip connections to mitigate vanishing gradients, organized into four stages of increasing feature map depth (64 $\rightarrow$ 128 $\rightarrow$ 256 $\rightarrow$ 512 channels). The final classification head was replaced and fine-tuned to output logits over four classes corresponding to Alzheimer's severity stages.

Training ran for 15 epochs using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1\times10^{-8}$) with a learning rate of 2×10^{-4} and a linear scheduler with a 10% warmup ratio. The model achieved a final validation accuracy of 92.19% and a validation loss of 0.1983 at epoch 15, trained using `Transformers 4.31.0` and `PyTorch 2.0.1`.

III. Results

A. Review Findings

The Evolution from Feature Engineering to Multimodal Deep Learning: Predictive modeling in Alzheimer's disease (AD) has shifted from feature engineering toward deep learning–based multimodal systems. Early models relied on predefined MRI-derived features, such as hippocampal volume, analyzed using classical machine learning approaches [11],[13]. While effective, these methods were limited by manual feature selection and reduced capacity to capture complex spatial patterns.

The introduction of convolutional neural networks (CNNs) enabled automated feature extraction from neuroimaging data, allowing models to learn distributed representations of neurodegeneration. More recent approaches integrate multimodal data, including MRI, PET, cerebrospinal fluid (CSF) biomarkers, and cognitive assessments, improving performance in early detection and MCI-to-AD prediction [12],[14]. However, these gains introduce challenges related to data requirements and interpretability, limiting clinical feasibility.

Benchmark Dataset Centralization and Structural Dependence: A defining feature of the literature is the centralization of model development around benchmark datasets, particularly the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS). These datasets provide standardized, multimodal data that support model training and cross-study comparison [11]–[13].

However, reliance on a narrow set of datasets introduces structural limitations. Because many studies use overlapping datasets for training and validation, high performance may reflect dataset-specific characteristics rather than generalizable disease patterns [12],[14]. This centralization also drives methodological convergence, with research focusing on incremental architectural improvements evaluated on the same datasets.

Additionally, recruitment biases within ADNI and OASIS raise concerns about representation and equity, as these datasets often overrepresent higher socioeconomic and healthcare-engaged populations [12],[13]. As a result, models trained on these data may not generalize to broader clinical populations. Together, these factors create a structural paradox: while benchmark datasets have accelerated progress, they risk producing models optimized for narrow research cohorts.

The Generalizability–Performance Paradox: A key challenge in AD AI research is the gap between high benchmark performance and limited real-world generalizability. Many models report accuracies exceeding 90% on curated datasets [11],[14], yet these results often fail to translate to clinical settings, where variability in imaging, patient populations, and comorbidities is substantial.

Models may also learn dataset-specific artifacts, such as differences in preprocessing, scanner types, or demographic distributions, which can inflate performance without reflecting true disease pathology [11]. In addition, reliance on internal validation limits insight into real-world robustness, as external validation across independent cohorts remains limited [12],[13].

These limitations highlight that benchmark accuracy alone is insufficient for assessing clinical utility. Improving generalizability requires multi-cohort training, external validation, and more diverse datasets.

Translational and Ethical Constraints in Clinical Integration: Beyond performance, predictive AI systems for AD face key translational and ethical challenges. Deep learning models often lack interpretability, making it difficult for clinicians to understand and trust diagnostic outputs in high-stakes contexts [14].

The integration of AI into clinical workflows also raises concerns around fairness, accountability, and governance. Predictive systems must be evaluated not only for accuracy but also for their impact across populations. Researchers emphasize the need for fairness auditing, subgroup performance reporting, and transparent documentation to ensure equitable deployment [12],[13].

Addressing these challenges requires more representative datasets, explainable AI approaches, and clear regulatory frameworks. Integrating these considerations into model development is essential to ensure that AI systems for AD diagnosis are both clinically reliable and equitable.

B. Technical Findings

The fine-tuned ResNet-34 demonstrates strong and consistent generalisation across a combined evaluation set of 7,680 MRI images drawn from two independent sources (Table 1). Achieving 97.55% accuracy, 98.76% precision, 96.49% recall, and an F1 score of 97.57%, the model maintains near-identical performance to its Falah-only results (98.59% accuracy), indicating that its learned representations transfer robustly to unseen data from a different dataset with different image dimensions and preparation pipelines (Table 2).

Evaluation Set	Score
Falah Only	98.59%
Combined	97.55%
Performance Difference	-1.04%

TABLE I
GENERALISATION PERFORMANCE COMPARISON.

Metric	Score
Accuracy	97.55%
Precision	98.76%
Recall	96.49%
F1	97.57%

TABLE II
PERFORMANCE METRICS OF THE MODEL.

The confusion matrix reinforces this at the per-class level (Figure 3). Non_Demented and Very_Mild_Demented — the two most populous classes with 3,834 and 2,699 true examples respectively — are classified with high fidelity, with only 4 and 120 misclassifications each. Moderate_Demented, historically the most challenging class due to its scarcity in training data, achieves 76 correct classifications out of 79 total examples, a marked improvement in absolute terms over the Falah-only run simply due to the larger sample size surfacing more true positives. The primary confusion pattern remains concentrated along the Mild/Very_Mild and Very_Mild/Non boundary, with 55 Mild_Demented cases predicted as Non_Demented and 119 Very_Mild_Demented cases predicted as Non_Demented — clinically expected given the visual overlap between early-stage and absent dementia on 2D MRI slices.

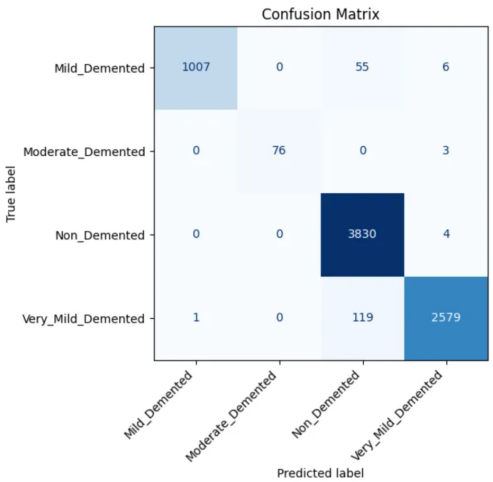

Fig. 3. Figure 3. Confusion matrix of the model.

Overall, the cross-dataset consistency of these results provides meaningful evidence that the model is not merely overfitting to the characteristics of its training distribution but has learned generalized features relevant to Alzheimer's severity staging across independently collected MRI data.

These results demonstrate that relatively accessible datasets and standard machine learning architectures are capable of producing high diagnostic performance. However, this also reinforces a key concern identified in the literature: strong benchmark accuracy does not guarantee equitable or clinically reliable systems. Although the model generalizes well across the evaluation datasets, the absence of demographic data and the presence of class imbalance highlight the same structural limitations discussed throughout the review. Consequently, model performance cannot be assessed across demographic subgroups, making it impossible to determine whether predictive accuracy remains consistent across varying populations.

These findings also emphasize that technical performance alone is insufficient to evaluate the real-world reliability of AI diagnostic systems. Instead, ethical considerations, including dataset representation, transparency, and ongoing monitoring, must be integrated into the development and evaluation process.

IV. ETHICAL FRAMEWORK

To mitigate dataset inequities in artificial intelligence models for Alzheimer's diagnosis, we propose an ethical framework adapted from that of Beauchamp and Childress, grounded in the principles of justice, non-maleficence, transparency, and accountability.

*1) **Dataset Representation and Justice**:* The principle of justice, defined as "fair, equitable, and appropriate treatment of persons," requires that AI diagnostic models perform equitably across populations [15]. Because machine learning models learn from training data, imbalances in dataset composition can lead to systematic differences in performance across demographic groups.

To address this, researchers should report key demographic characteristics, including age, sex, race, ethnicity, and relevant social determinants of health such as socioeconomic status and education. Recruitment processes and eligibility criteria must also be evaluated to identify potential exclusion of underrepresented populations.

Dataset composition should be compared to real-world disease prevalence where possible. When imbalances cannot be corrected through data collection, mitigation strategies such as oversampling, undersampling, and data augmentation should be applied [16].

*2) **Harm Prevention and Non-Maleficence**:* The principle of non-maleficence, defined as the obligation to avoid harm, emphasizes the need to prevent inaccurate or inequitable predictions in AI diagnostic systems [15]. In AD, diagnostic errors may result in delayed intervention, inappropriate treatment, and unequal access to care.

To mitigate these risks, model evaluation should extend beyond overall accuracy to include subgroup-specific error rates, particularly false positives and false negatives across demographic groups. This is especially important for populations at risk of underdiagnosis or misclassification.

Dataset quality must also be assessed for biases in labeling processes, as diagnostic criteria may be influenced by factors such as education. These limitations should be explicitly acknowledged, given their potential impact on model performance and clinical outcomes.

*3) **Transparency in Model Development and Reporting**:* Transparency is essential to preserving trust between patients and healthcare providers, requiring that providers and developers communicate how AI systems function, produce decisions, and process data in a manner understandable to patients. Researchers should provide detailed documentation of dataset sources, inclusion and exclusion criteria, and procedures used

in model training. Additionally, they should report both performance metrics for the general + subgroups, alongside any potential sources of bias. Moreover, interpretability methods should be implemented to explain factors that impact model predictions, as this can enhance clinical understanding, support informed decision-making, and promote trust in AI-assisted diagnostic tools.

*4) **Accountability and Ongoing Monitoring**:* The principle of accountability requires that stakeholders remain responsible for monitoring AI systems post-deployment and implementing mitigation strategies when errors or harms arise [17].

Because model performance is dynamic, continuous evaluation is necessary to detect emerging biases. As models are updated with new data, they may unintentionally reinforce existing inequities, reducing accuracy for certain demographic groups.

Healthcare institutions should therefore implement ongoing auditing procedures to assess subgroup performance, identify disparities, and apply corrective measures. Clear governance structures are also needed to define responsibility for model updates, bias mitigation, and communication of limitations to clinicians and patients.

V. Limitations

This study has limitations related to the literature review scope, dataset selection, and model design.

First, the review was limited to three databases; inclusion of additional sources such as Scopus or Google Scholar may have provided a more comprehensive overview of AI applications in AD diagnosis.

Second, dataset accessibility constrained model development. Widely used datasets such as ADNI and OASIS require extensive application processes, limiting their use within this study's timeframe. As a result, the model was trained solely on the Falah/Alzheimer_MRI dataset, which consists of low-resolution (128×128) images. This may reduce the model's ability to capture fine-grained structural features and limit generalizability to clinical-grade MRI data.

The dataset is also highly imbalanced, with few Moderate Demented samples (15 in the test set), likely contributing to poorer performance for this class. Additionally, the absence of demographic and scanner-related metadata prevents evaluation of subgroup performance, limiting assessment of fairness and clinical applicability.

Several technical limitations further constrain the model. The use of ResNet-34 (21.3M parameters) may limit the ability to capture complex spatial patterns compared to deeper architectures. The model also operates on 2D MRI slices rather than full 3D volumes, discarding spatial context critical for clinical assessment.

Moreover, the model lacks uncertainty estimation, relying on uncalibrated softmax outputs that do not distinguish between low- and high-confidence predictions. Finally, training was conducted on CPU (PyTorch 2.0.1+cpu), which may have limited training duration, batch size, and overall model convergence.

VI. Conclusion

This study examined the current landscape of AI in Alzheimer's disease (AD) diagnosis through a narrative review and the development of a prototype classification model. The literature highlights a clear methodological shift from feature engineering to multimodal deep learning, enabling improved predictive performance while introducing challenges related to interpretability and clinical feasibility.

At the same time, the field is characterized by structural reliance on benchmark datasets such as ADNI and OASIS. While these datasets have accelerated research, their dominance raises concerns about dataset centralization, demographic representation, and limited generalizability. This contributes to a broader generalizability–performance paradox, where high accuracy within controlled datasets does not reliably translate to real-world clinical settings.

The findings also underscore the need for stronger governance and accountability as AI systems move toward clinical integration. The lack of transparency in deep learning models complicates clinical adoption and may limit trust in diagnostic decision-making.

The prototype model developed in this study reinforces these limitations. Despite strong performance, constraints related to dataset composition, class imbalance, and the absence of demographic information highlight that technical accuracy alone is insufficient to ensure equitable and clinically reliable systems.

In response, this study proposes an ethical framework grounded in justice, non-maleficence, transparency, and accountability. By emphasizing dataset representation, subgroup performance evaluation, and continuous monitoring, the framework supports the development of AI systems that are not only accurate, but also equitable and clinically meaningful.

VII. Future Work

Future work should expand both the ethical framework and technical capabilities of AI models for AD diagnosis. The proposed framework should be applied across the full model development lifecycle and extended to other predictive domains, such as multiple sclerosis, where similar ethical concerns arise [18].

From a technical perspective, future models should incorporate full 3D MRI volume processing using architectures such as 3D-CNNs or Vision Transformers to preserve spatial continuity. Integrating multimodal inputs would further align model development with clinical diagnostic workflows.

Addressing dataset limitations is also critical. Class imbalance, particularly for the Moderate Demented category, should be mitigated through techniques such as weighted loss functions, oversampling, or synthetic data augmentation. Additionally, future studies should prioritize datasets with demographic metadata to enable evaluation of performance across age, sex, and ethnicity, which is essential for assessing fairness in clinical AI systems.

Finally, given the progressive nature of AD, future models should leverage longitudinal data to predict disease trajectories

over time rather than static disease stages, improving clinical relevance and decision-making.

VIII. Acknowledgements

The team thanks Shrika Vejandla, the director of design for the QMIND AI ethics node, for her support throughout the project.

References

[1] A. A. Soladoye, N. Aderinto, D. Osho, and D. B. Olawade, "Explainable machine learning models for early Alzheimer's disease detection using multimodal clinical data," *International Journal of Medical Informatics*, vol. 204, p. 106093, Aug. 2025. https://doi.org/10.1016/j.ijmedinf.2025.106093

[2] World Health Organization, "Dementia," Mar. 31, 2025. https://www.who.int/news-room/fact-sheets/detail/dementia

[3] Alzheimer's Disease International, "Over 41 million cases of dementia go undiagnosed across the globe – World Alzheimer Report reveals," Sep. 21, 2021. https://www.alzint.org/news-events/news/over-41-million-cases-of-dementia-go-undiagnosed-across-the-globe-world-alzheimer-report-reveals

[4] D. Agostinho, M. Simões, and M. Castelo-Branco, "Predicting conversion from mild cognitive impairment to Alzheimer's disease: a multimodal approach," *Brain Communications*, vol. 6, no. 4, Jan. 2024. https://doi.org/10.1093/braincomms/fcae208

[5] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, Nov. 2014. https://doi.org/10.1016/j.neuroimage.2014.06.077

[6] L. Winchester *et al.*, "Artificial intelligence for biomarker discovery in Alzheimer's disease and dementia," *Alzheimer's & Dementia*, vol. 19, no. 12, Aug. 2023. https://doi.org/10.1002/alz.13390

[7] C. Birkenbihl, Y. Salimi, D. Domingo-Fernández, S. Lovestone, H. Fröhlich, and M. Hofmann-Apitius, "Evaluating the Alzheimer's disease data landscape," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 6, no. 1, Jan. 2020. https://doi.org/10.1002/trc2.12102

[8] K. A. Johnson, N. C. Fox, R. A. Sperling, and W. E. Klunk, "Brain Imaging in Alzheimer Disease," *Cold Spring Harbor Perspectives in Medicine*, vol. 2, no. 4, Jan. 2012. https://doi.org/10.1101/cshperspect.a006213

[9] M. R. Caunca *et al.*, "Neuroimaging markers of dementia across race/ethnicity and sex/gender using an intersectional approach within the HABS-HD cohort," *Alzheimer's & Dementia*, vol. 21, no. 9, Sep. 2025. https://doi.org/10.1002/alz.70733

[10] A. C. Lim *et al.*, "Quantification of race/ethnicity representation in Alzheimer's disease neuroimaging research in the USA: a systematic review," *Communications Medicine*, vol. 3, no. 1, pp. 1–12, Jul. 2023. https://doi.org/10.1038/s43856-023-00333-6

[11] S. Basanta-Torres, M. Á. Rivas-Fernández, and S. Galdo-Alvarez, "Artificial Intelligence for Alzheimer's disease diagnosis through T1-weighted MRI: A systematic review," *Computers in Biology and Medicine*, vol. 197, Sep. 2025. https://doi.org/10.1016/j.compbiomed.2025.111028

[12] Z. Batool, S. Hu, M. A. Kamal, N. H. Greig, and B. Shen, "Advancing Alzheimer's Diagnosis with AI-Enhanced MRI: A Review of Challenges and Implications," *Current Neuropharmacology*, vol. 23, Jul. 2025. https://doi.org/10.2174/011570159x353595250303064846

[13] K. K. Pandey, A. Mishra, and R. Milan, "Predictive modeling approaches for Alzheimer's disease diagnosis through neuroimaging techniques," *Ageing Research Reviews*, vol. 114, p. 102989, Jan. 2026. https://doi.org/10.1016/j.arr.2025.102989

[14] M. R. Shaikh, A. Jeyabose, and R. V. Arjunan, "Deep learning for Alzheimer's disease: advances in classification, segmentation, subtyping, and explainability," *BioMedical Engineering OnLine*, vol. 24, no. 1, Dec. 2025. https://doi.org/10.1186/s12938-025-01482-6

[15] B. Varkey, "Principles of Clinical Ethics and Their Application to Practice," *Medical Principles and Practice*, vol. 30, no. 1, pp. 17–28, 2021. https://doi.org/10.1159/000509119

[16] J. L. Cross, M. A. Choma, and J. A. Onofrey, "Bias in Medical AI: Implications for Clinical decision-making," *PLOS Digital Health*, vol. 3, no. 11, p. e0000651, Nov. 2024. https://doi.org/10.1371/journal.pdig.0000651

[17] A. J. Gorelik *et al.*, "Ethics of AI in healthcare: a scoping review demonstrating applicability of a foundational framework," *Frontiers in Digital Health*, vol. 7, Sep. 2025. https://doi.org/10.3389/fdgth.2025.1662642

[18] N. Nathoo, B. Zeydan, N. Neyal, C. Chelf, D. T. Okuda, and O. H. Kantarci, "Do magnetic resonance imaging features differ between persons with multiple sclerosis of various races and ethnicities?," *Frontiers in Neurology*, vol. 14, Jun. 2023. https://doi.org/10.3389/fneur.2023.1215774