Investigating Bias and Fairness for Alzheimer's Diagnosis Using the OASIS-2 Dataset

1st Mansi Singhal
Department of Biomedical Engg.
University of Calgary
Calgary, AB, Canada
mansi.singhal@ucalgary.ca

2nd Mariana P. Bento
Department of Biomedical Engg.
University of Calgary
Calgary, AB, Canada
mariana.pinheirobent@ucalgary.ca

3rd Leticia Rittner
School of Electrical and Computer Engg.
University of Campinas
Campinas, Brazil
lrittner@unicamp.br

***Abstract*—Artificial Intelligence has shown great promise in aiding the early identification of neurodegenerative diseases, including Alzheimer's disease. The transition of AI into the medical community calls for elevated accuracy, fairness, and explainability in clinical decision-making. The current study evaluates the predictive performance, fairness, and explainability, via SHAP (SHapley Additive Explanations), of five machine learning algorithms: logistic regression, SVM, random forest, AdaBoost, and XGBoost, on the OASIS-2 tabular dataset to estimate Alzheimer's diagnosis. Fairness metrics included equal opportunity difference, average odds difference, statistical parity difference, disparate impact, and theil index. The results empower XGBoost as the most predictive and fair answer to the problem of Alzheimer's diagnosis, while also requiring the fewest resources in terms of explainable features. The results show the importance of fairness-conscious and explainable artificial intelligence technology in medicine to ensure broad ambits of equitable and clinically meaningful predictions. All code and experiments are publicly available at: https://github.com/MansiSinghal502/BMEN691**

***Index Terms*—Alzheimer's Disease, Machine Learning, Fairness, Explainable AI, Imbalanced data**

I. Introduction

Artificial Intelligence (AI) systems are vulnerable to biases that may be intentionally or unintentionally introduced dur ing data collection, sampling, and model development and measurement, as these systems are highly dependent on the data that is fed into them [1]. Particularly in the healthcare industry, this may result in wrong forecasts and a higher chance of diagnostic errors, which may cause irreversible harm [2]. Data bias and algorithmic prejudice are unavoidable, despite the best efforts to prevent them [3]. It is critical to identify and address bias, particularly in delicate fields like healthcare [4].

The motivation of this study was to understand the decision-making process of machine learning (ML) models that may be influenced by bias in data, which disproportionately affects minority groups. In this study we were finding algorithmic fairness [5] using the OASIS-2 dataset [6].

The following research questions guide the scope of this project:

- RQ1: Which ML model will show to be effective for Alzheimer's Disease (AD) diagnosis?
- RQ2: How do fairness metrics evaluate AI models used for AD diagnosis or classification?
- RQ3: How do the decision-making processes of different AI models contribute to explaining AD diagnosis?
- RQ4: What are the challenges and open research directions in diagnosing AD?

The questions order follows from general techniques (RQ1) to specific considerations for AD evaluation and explanation (RQ2 and RQ3) and, finally, future directions (RQ4) [7].

II. Related Work

This section explains about various existing techniques and tools to predict the early signs of AD and distinguish between a demented and a nondemented person. The authors Yudong Zhanget al., in their work "Classification of Alzheimer Disease Based on Structural Magnetic Resonance Imaging by Kernel Support Vector Machine Decision Tree" [8], illustrate the working of an SVM based decision tree (DT) model, that is capable of detecting the presence of AD in the OASIS 1 – Cross sectional data. In the work "Prediction of Alzheimer's Disease using Oasis Dataset" by Chandni Naidu et al. [9], the authors predict the presence of AD in subjects based on the CDR in the OASIS 1 dataset. Deepika Bansal et al., in their work "Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia" [10], provide comparative analysis on the algorithms utilized, to predict dementia for the OASIS 1 dataset from the OASIS 2 trained prediction models. This paper proposes a modified capsule network (MCapNet) for early AD prediction using the OASIS dataset, achieving 92.39% accuracy while emphasizing fairness, interpretability, and computational efficiency compared to traditional deep learning models [7]. This study proposes and compares multiple machine learning models including XGBoost, Random Forest, SVM, ANN, and Naïve Bayes—for predicting dementia using OASIS MRI and demographic data, with XGBoost achieving the highest accuracy of 97.87% and demonstrating reliable classification of dementia stages [11].

A. Justification of proposed study

There were a number of works that employ ML algorithms such as logistic regression and random forest, which report high accuracy. However, most of the literature does not examine how such models operate in terms of decision-making and explainability. Instead, we focus on evaluating whether these

AI models are fair and how their underlying decision-making process is performed and interpreted.

III. Materials and Methods

The focus of this paper was to run the OASIS-2 dataset on different ML models and to understand the fairness and explainability of those models. The flowchart outlining the study of this project shown in Fig. 1.

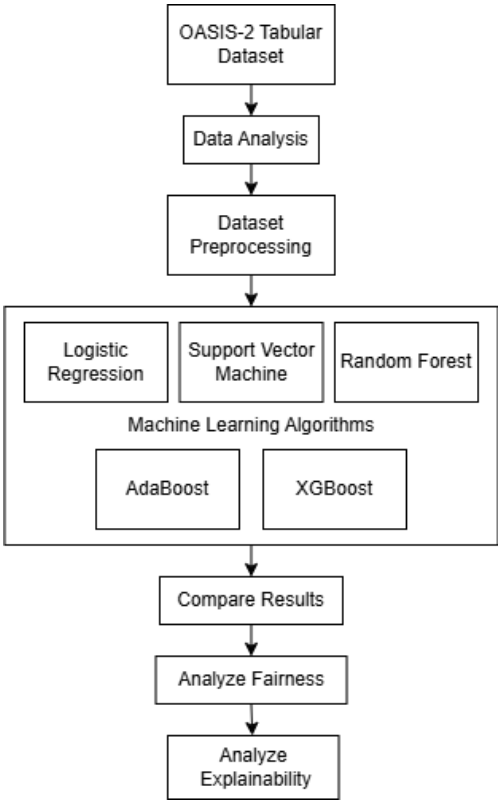

Fig. 1. Flowchart highlighting fairness and explainability in dementia prediction.

A. Dataset Description

The data used for preparation of the paper was taken from Open Access Series of Imaging Studies subset-2 (OASIS-2) [12] which was a longitudinal dataset containing T1-weighted magnetic resonance imaging (MRI) acquisitions of 150 subjects with and without dementia between 2 to 5 independent visits, as well as demographic and clinical information about each subject from each visit. The subjects sampled in the OASIS-2 dataset were a subset of those who participated in OASIS-1 [13], a dataset of cross-sectional MRI images in adults of varying ages with and without dementia. OASIS-2 contained a selection of these individuals who were right-handed, between 60 and 96 years of age, with and without dementia. In this paper, we specifically used the tabular (demographic and clinical) portion of the OASIS-2 dataset. Feature details of the OASIS-2 Dataset are given in Table I

TABLE I
Description of Oasis-2 Data demographic, clinical, and derived imaging measures

Measure	Format	Definition
Age	Integer	Age at time of image acquisition (years)
Sex (M/F)	Category	Sex [Male (M) or Female (F)]
Education (EDUC)	Integer	Years of education
SES	Integer	Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (highest status) to 5 (lowest status)
MMSE	Integer	Mini-Mental State Examination score (range is from 0 [worst] to 30 [best])
CDR	Float	Clinical Dementia Rating. (0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2. = moderate AD)
ASF	Float	Atlas scaling factor (unitless). Computed scaling factor that transforms native-space brain and skull to the atlas target (i.e. the determinant of the transform matrix)
eTIV	Float	Estimated total intracranial volume (cm^3)
nWBV	Float	Normalized whole brain volume, expressed as a percent of all voxels in the atlas-masked image that are labeled as gray or white matter by the automated tissue segmentation process

B. Dataset Exploration

The OASIS-2 dataset presents several important considerations related to bias and data imbalance. Over successive visits, participant attendance declines, resulting in a longitudinal imbalance, particularly in later visits (3 to 5). Additionally, the male-to-female ratio becomes inconsistent across visits, and all participants are right-handed, a design choice that limits generalizability. There was a significant class imbalance, with the nondemented group substantially outnumbering both the demented and converted groups across all visits. Regarding missing data and duplicates, the dataset was relatively clean, containing only 19 missing SES values and 2 missing MMSE scores, affecting a total of 8 unique individuals. No duplicate records were identified. The missing values could either be removed or imputed using group-specific means, as SES and MMSE values significantly differ across dementia groups. In terms of statistical relationships, a very strong correlation was observed between MR delay and visit number (r = 0.92), and between eTIV and ASF (r = -0.99), the latter due to mathematical dependency. Additionally, education and SES are moderately negatively correlated (r = -0.72), as were MMSE and CDR (r = -0.69), reflecting the expected clinical pattern where lower cognitive performance MMSE aligns with higher dementia severity CDR. Other feature correlations were weak, suggesting relative independence among most variables.

C. Dataset Preprocessing

In the preprocessing phase of the OASIS-2 dataset, several critical steps were taken to ensure data quality and consistency for machine learning analysis. Since subjects had multiple visits and duplicate Subject IDs, only the last visit for each subject was retained by sorting based on Subject ID and Visit, reducing the dataset from 373 to 150 records and ensuring

a one-to-one mapping between rows and individuals. The Hand column was dropped because all participants were right-handed, rendering the feature non-informative and potentially bias-inducing. To address missing values, rows with null entries in crucial features such as SES and MMSE were removed, resulting in a fully clean dataset for modeling. The CDR variable, which originally ranged from 0 to 3, was binarized assigning 0 to nondemented individuals and 1 to those with any level of dementia and used to reclassify the group column accordingly [14]. This created a balanced binary classification problem with 73 nondemented and 69 demented subjects. Additionally, a correlation matrix of the numeric features revealed that the ASF feature indicates high correlation with eTIV and is eliminated. Similarly SES also been eliminated because of high negative correlation with EDUC. Lastly, Visit and MR Delay both were removed from feature selection as they both have high positive correlation and we already taking the last visit of each Subject IDs. Preprocessing was performed using a ColumnTransformer, where numerical features were normalized using MinMaxScaler and the categorical feature (M/F) was encoded using 'OneHotEncoder'. Importantly, the scaler was fitted exclusively on the training data and applied to the validation and test sets to avoid data leakage. The preprocessed feature sets were then converted to structured DataFrames, facilitating inspection and further model development. This preprocessing strategy ensured that the data was clean, standardized, and appropriately partitioned for reliable ML model training and evaluation. The models were trained and tested using 5 fold cross validation. The final features, considered as input for the models, were sex, age, education, mini-mental state examination score, and clinical dementia rating. normalized whole brain volume, estimated total intercranial volume.

D. Machine Learning Models

The dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve class distribution. This study employs five widely recognized ML classifiers [15] logistic regression (LR) [16], [17], support vector machine (SVM) [17], [18], random forest (RF) [17], [19], AdaBoost [7], [20], and XGBoost [17], [21] to predict cognitive status using the OASIS-2 longitudinal tabular dataset. Each model was selected for its unique strengths in handling classification problems and its relevance to clinical data analysis [22]. These models were the most used ML classifiers by the researchers.

IV. Results and Discussion

A. Comparison of Models

This study conducted a comprehensive evaluation of five supervised machine learning models logistic regression, SVM, random forest, AdaBoost, and XGBoost for binary classification of dementia using the OASIS-2 dataset. The models test accuracies (Acc) depicted in Table III, as well as detailed performance metrics : precision, recall, F1-score, and class-wise support for both demented and nondemented groups shown in Fig. 2.

Among all, XGBoost obtained the highest test accuracies (Acc) and F1-score for the nondemented class, while random forest showed more balanced generalization across both classes. A persistent challenge across all models was the low recall and F1-score for the demented class. This suggests that the models struggled to detect true positives for dementia, often misclassifying them as nondemented.

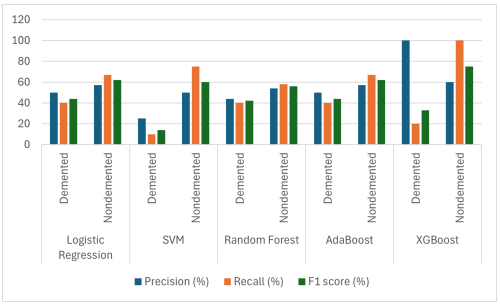

Fig. 2. Precision, Recall and F1 Score of the models.

B. Fairness and Bias Evaluation

Fairness is a powerful protection against biased AI decision-making [23]. Because of social, demographic, or behavioral characteristics, no group should be given preference or marginalization [24]. Group fairness guarantees comparable results for various demographic groupings [25]. All experiments for bias identification are carried out using IBM's AIF360 open-source toolkit [26], which offers a variety of cutting-edge fairness evaluation metrics. For our investigation, we used IBM's AIF360 open source code, which allowed academics to expand on our findings [17].

The fairness assessment of the five ML models was conducted using five key metrics: Equal Opportunity Difference (EOD), Average Odds Difference (AOD), Statistical Parity Difference (SPD), Disparate Impact (DI), and Theil Index (TI) shown in Table II . These metrics evaluate group fairness by comparing prediction outcomes between privileged (male) and unprivileged (female) groups, as well as individual fairness through distributional equity [17].

Table III presents a comparative analysis of five ML models based on various fairness metrics. Among these models, XGBoost showed the best fairness performance, achieving ideal or near-ideal values across all metrics. It records 0 for both EOD and AOD, an SPD close to 0, a DI closest to 1, and the lowest TI, indicating minimal inequality. XGBoost overall fairness score was also the lowest, confirming it as the most balanced model in terms of fairness. In contrast, SVM performs the worst, with the highest values in EOD, AOD, and SP, indicating serious disparities across groups. Its fairness score further highlights its lack of fairness compared to the other models. Random Forest shows moderate fairness, with better metrics than SVM but still higher disparities

than Logistic Regression, AdaBoost, and XGBoost. Logistic Regression and AdaBoost both demonstrate relatively good fairness, with low EOD, AOD, and SP values and fairness scores of 0.3221 each. Notably, XGBoost achieves 0 EOD and AOD, showing no group-level disparity in positive predictions, although it slightly underperforms in overall fairness due to a higher TI. All results of fairness metrics are shown in Fig. 3.

TABLE II
FAIRNESS EVALUATION METRICS

Metric	Ideal Value	Acceptable Threshold	What It Measures
Equal Opportunity Difference (EOD)	0	[-0.1, 0.1]	Disparity in true positive rates (TPR) between privileged and unprivileged groups.
Average Odds Difference (AOD)	0	[-0.1, 0.1]	Average disparity in TPR and false positive rate (FPR) between groups.
Statistical Parity Difference (SPD)	0	[-0.1, 0.1]	Difference in positive prediction rates between unprivileged and privileged groups.
Disparate Impact (DI)	1	[0.8, 1.2]	Ratio of favorable outcomes between unprivileged and privileged groups.
Theil Index (TI)	0	[0, 0.25]	Inequality in prediction distribution across individuals.

TABLE III
FAIRNESS RESULTS FOR VARIOUS MODELS ON THE OASIS-2 DATASET

Metric	Logistic Regression	Support Vector Machine	Random Forest	AdaBoost	XGBoost
Acc	0.5455	0.4545	0.5000	0.5455	0.6364
EOD	0.1250	0.3750	0.2500	0.1250	0.0000
AOD	0.0625	0.2875	0.1250	0.0625	0.0000
SP	0.0513	0.3077	0.1282	0.0513	-0.0342
DI	1.0833	1.4444	1.2380	1.0833	0.9630
TI	0.2596	0.2044	0.3172	0.2596	0.0595
Score	0.3221	1.4142	0.7412	0.3221	0.0712

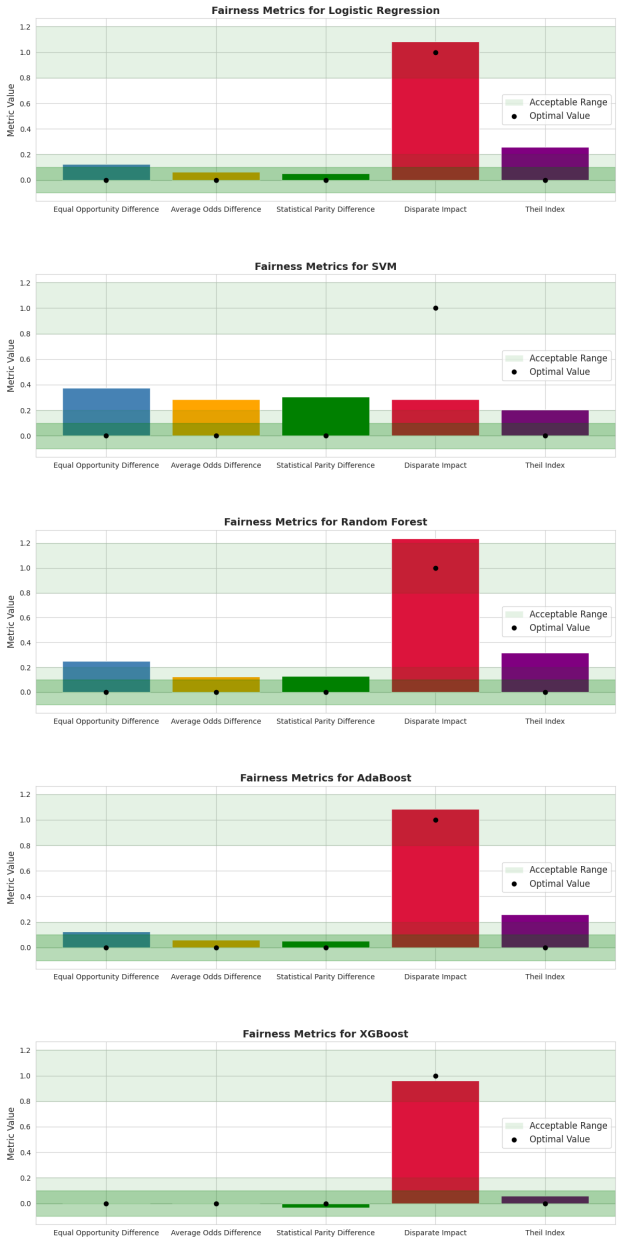

Fig. 3. Comparative Fairness Evaluation of Machine Learning Models Using AIF360 Metrics.

C. Explainability Analysis

A detailed explanation of how each feature contributes to the prediction of AD was carried out using SHAP (SHapley Additive Explanations) [27] analysis on the five ML classifiers [28]. The SHAP waterfall plots provided insight into the contribution of individual features toward the model's output, offering a transparent understanding of model behavior [29], [30]. The SHAP waterfall plot is shown in Fig. 4 for logistic regression, Fig. 5 for SVM, Fig. 6 for random forest, Fig. 7 for AdaBoost and Fig. 8 for XGBoost.

For the logistic regression model, the variables EDUC, Age, and eTIV all positively added to the prediction of AD, while CDR, eTIV, M/F, and age all negatively contributed and increased the likelihood of prediction for AD. For the SVM model, we noticed that there were slight positive SHAP values present. Only EDUC and MMSE had a small positive coefficient, while CDR, M/F, and Age had large negative values, meaning they drastically decreased the predicted probability of diagnosing AD. For the random forest model, AGE, EDUC, and MMSE were all strong positive contributors, and they moved the prediction toward predicting AD diagnosis, while nWBV and the CDR variable were negative contributors. For the AdaBoost model, age was the strongest positive feature for predicting AD, followed by MMSE, EDUC, and eTIV. The CDR and nWBV were the most significant contributors to negatively predicting AD diagnosis. Finally, for the XGBoost model, the results were also similar to that of Logistic

Regression, Random Forest, and AdaBoost models and had the most positive contributions by EDUC, Age, and MMSE, while the negative contributors for XGBoost prediction were nWBV, eTIV, M/F, and CDR.

Overall, AGE and EDUC consistently showed up as key positive drivers in predicting AD across all models, whereas higher nWBV and higher CDR values tended to reduce the likelihood of AD prediction. However, SHAP visualizations clearly demonstrated which features each model relied on the most for its decisions, improving transparency and interpretability in AD diagnosis using ML.

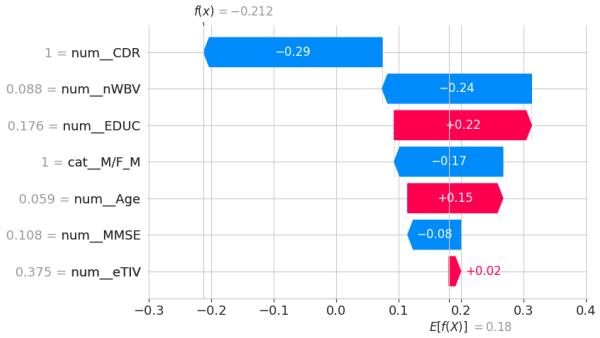

Fig. 4. SHAP Summary of Logistic Regression.

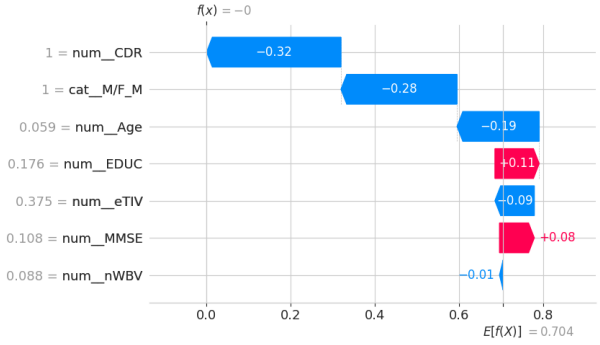

Fig. 5. SHAP Summary of Support Vector Machine (SVM).

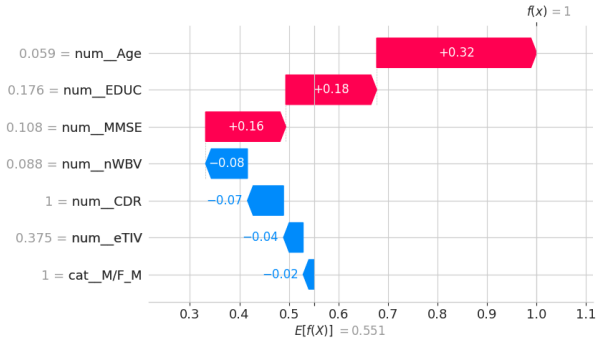

Fig. 6. SHAP Summary of Random Forest.

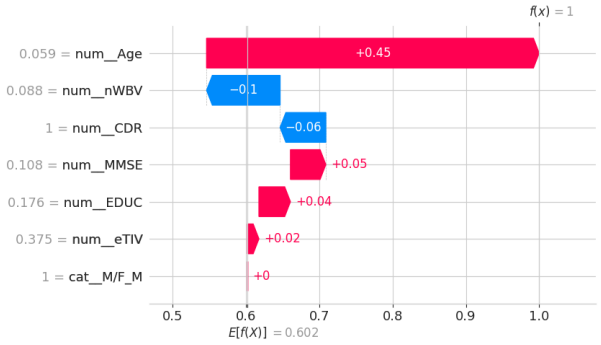

Fig. 7. SHAP Summary: AdaBoost Classifier.

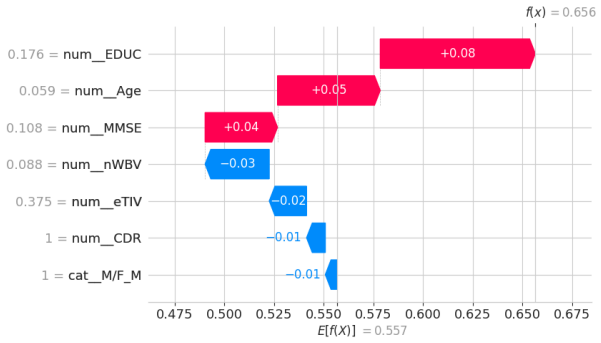

Fig. 8. SHAP Summary: XGBoost Classifier.

V. Conclusions

This study focused mainly on reproducibility using the open-access OASIS-2 dataset. ML models used in this work were preferred, which were mostly used by the researchers in their research paper. Responsible AI is considered in this study through the AIF360 fairness metrics and explainability by SHAP for a better understanding of the model so that no one will trust the model only on accuracy but also look at its decision-making process for better reliability. Results showed that XGBoost had greater accuracy than other models, but after analyzing the fairness and explainability of the model, we now know the reason for that accuracy. These results highlight the significance of incorporating bias mitigation techniques into the ML pipeline. Improving the equity and dependability of prediction models in the healthcare industry requires addressing these biases. Despite promising results, this study has several limitations. The dataset used, OASIS-2, is relatively small and imbalanced, especially in the representation of demented versus nondemented classes and the subjects contain only right handed people. Future research will focus on data augmentation to expand the dataset and will continue work on bias mitigation techniques throughout the ML pipeline.

References

[1] G. Briganti and O. Le Moine, "Artificial intelligence in medicine: today and tomorrow," *Frontiers in medicine*, vol. 7, p. 509744, 2020.

[2] J. Yang, A. A. Soltan, D. W. Eyre, Y. Yang, and D. A. Clifton, "An adversarial training framework for mitigating algorithmic biases in clinical machine learning," *NPJ digital medicine*, vol. 6, no. 1, p. 55, 2023.

[3] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," *Advances in neural information processing systems*, vol. 30, 2017.

[4] F. Anzum, A. Z. Asha, and M. L. Gavrilova, "Biases, fairness, and implications of using ai in social media data mining," in *2022 International Conference on Cyberworlds (CW)*. IEEE, 2022, pp. 251–254.

[5] X. Wang, Y. Zhang, and R. Zhu, "A brief review on algorithmic fairness," *Management System Engineering*, vol. 1, no. 1, p. 7, 2022.

[6] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[7] S. Basheer, S. Bhatia, and S. B. Sakri, "Computational modeling of dementia prediction using deep neural network: analysis on oasis dataset," *IEEE access*, vol. 9, pp. 42 449–42 462, 2021.

[8] Y.-D. Zhang, S. Wang, and Z. Dong, "Classification of alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree," *Progress In Electromagnetics Research*, vol. 144, pp. 171–184, 2014.

[9] C. Naidu, D. Kumar, N. Maheswari, M. Sivagami, and G. Li, "Prediction of alzheimer's disease using oasis dataset," 2019.

[10] D. Bansal, R. Chhikara, K. Khanna, and P. Gupta, "Comparative analysis of various machine learning algorithms for detecting dementia," *Procedia computer science*, vol. 132, pp. 1497–1502, 2018.

[11] S. Vinayak, E. Shahina, and A. Nayeemulla Khan, "Dementia prediction on oasis dataset using supervised and ensemble learning techniques," *Int. J. Eng. Adv. Technol*, vol. 10, no. 1, pp. 244–254, 2020.

[12] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults," *Journal of Cognitive Neuroscience*, vol. 22, no. 12, pp. 2677–2684, Dec. 2010. [Online]. Available: https://direct.mit.edu/jocn/article/22/12/2677/4983/Open-Access-Series-of-Imaging-Studies-Longitudinal

[13] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, Sep. 2007. [Online]. Available: https://direct.mit.edu/jocn/article/19/9/1498/4427/Open-Access-Series-of-Imaging-Studies-OASIS-Cross

[14] K. T. Chui, B. B. Gupta, W. Alhalabi, and F. S. Alzahrani, "An mri scans-based alzheimer's disease detection via convolutional neural network and transfer learning," *Diagnostics*, vol. 12, no. 7, p. 1531, 2022.

[15] E. M. Mohammed, A. M. Fakhrudeen, and O. Y. Alani, "Detection of alzheimer's disease using deep learning models: A systematic literature review," *Informatics in Medicine Unlocked*, vol. 50, p. 101551, 2024.

[16] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[17] F. Dehghani, N. Malik, J. Lin, S. Bayat, and M. Bento, "Fairness in healthcare: Assessing data bias and algorithmic fairness," in *2024 20th International Symposium on Medical Information Processing and Analysis (SIPAIM)*. IEEE, 2024, pp. 1–6.

[18] S. Vishwanathan and M. N. Murty, "Ssvm: a simple svm algorithm," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, vol. 3. IEEE, 2002, pp. 2393–2398.

[19] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[20] G. Rätsch, T. Onoda, and K.-R. Müller, "Soft margins for adaboost," *Machine learning*, vol. 42, pp. 287–320, 2001.

[21] S. Ramraj, N. Uzir, R. Sunil, and S. Banerjee, "Experimenting xgboost algorithm for prediction and classification of different datasets," *International Journal of Control Theory and Applications*, vol. 9, no. 40, pp. 651–662, 2016.

[22] P. Baglat, A. W. Salehi, A. Gupta, and G. Gupta, "Multiple machine learning models for detection of alzheimer's disease using oasis dataset," in *Re-imagining diffusion and adoption of information technology and systems: A continuing conversation: IFIP WG 8.6 international conference on transfer and diffusion of IT, TDIT 2020, Tiruchirappalli, India, December 18–19, 2020, proceedings, part I*. Springer, 2020, pp. 614–622.

[23] J. A. Colquitt and J. B. Rodell, "Measuring justice and fairness," *The Oxford handbook of justice in the workplace*, vol. 1, pp. 187–202, 2015.

[24] J. Yang, A. A. Soltan, D. W. Eyre, and D. A. Clifton, "Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning," *Nature Machine Intelligence*, vol. 5, no. 8, pp. 884–894, 2023.

[25] J. Xu, Y. Xiao, W. H. Wang, Y. Ning, E. A. Shenkman, J. Bian, and F. Wang, "Algorithmic fairness in computational medicine," *EBioMedicine*, vol. 84, 2022.

[26] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[27] SHAP Contributors. (2025, Apr.) Shap waterfall plot — shap latest documentation. SHAP. Accessed: Apr. 10, 2025. [Online]. Available: https://shap.readthedocs.io/en/latest/example_notebooks/apiexamples/plots/waterfall.html

[28] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: assessing shap values as feature selection mechanism," in *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*. Ieee, 2020, pp. 340–347.

[29] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, "Feature selection strategies: a comparative analysis of shap-value and importance-based methods," *Journal of Big Data*, vol. 11, no. 1, p. 44, 2024.

[30] Z. Li, "Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost," *Computers, Environment and Urban Systems*, vol. 96, p. 101845, 2022.

VI. Appendix

Prompt 1. Why am I getting an error while uploading my google colab notebook on github?

ChatGPT response 3: The repo name or path might be wrong, or the repo is private and Colab can't access it.

Prompt 2. Why are my references in IEEE format going out of the margin?

ChatGPT response 2. URLs like the ones in [11], [12], and [23] are not automatically wrapped in most IEEE templates. IEEE format does not justify long links, and they may overflow the column width.