Engineering Data-Driven Ensemble-based Classification Models for the Early Detection of Alzheimer's Disease

Introduction

Alzheimer's Disease (AD) is a significant public health challenge, with rising prevalence imposing considerable socio-economic and personal burdens worldwide. Early diagnosis and understanding of AD progression are crucial for developing effective and preemptive interventions. By examining large existing datasets, predictive models leveraging advanced statistical and machine learning techniques have emerged as powerful tools in assessing AD risk. These models excel in processing extensive datasets, uncovering intricate patterns that traditional methods might overlook. By integrating factors such as sleep sufficiency, mental distress, obesity, geographic distribution, and dietary habits, predictive models offer nuanced insights into AD risk, which can inform interventions and disease outlook in patients. This study presents the development and validation of a predictive model using these factors to forecast AD risk across the U.S. population, drawing on comprehensive census data.

Numerous predictive models have been developed for AD risk assessment, employing methodologies from traditional statistical models to advanced machine learning techniques. For example, one study utilized logistic regression to identify clinical predictors of AD[1], while another explored the use of support vector machines to classify MRI images[2]. Recent advancements include deep learning models like convolutional neural networks (CNNs), which have been demonstrated for predicting AD progression using brain images[3].

Our research aims to contribute to identifying the growing body of literature on predictive modeling in AD, showcasing the potential of big data and machine learning in public health research.

Methodology

Data Synthesis

The data collection began with extracting an Alzheimer's subset from the Behavioral Risk Factor Surveillance System (BRFSS) study, referred to as the Alzheimer's dataset.[4] This longitudinal sub-study, conducted by the U.S. government across all 50 states and the District of Columbia from 2015 to 2022, was gathered by co-authors SA and SA and peer-reviewed by MW. To enhance the predictive model's accuracy, data columns with zero values were excluded, and stratifications were simplified to include only age group and state of residence.

In addition to the Alzheimer's dataset, U.S. Census population estimates from 2015 to 2022 were incorporated to simulate the U.S. population based on current Alzheimer's disease estimates. Population estimates for 2010-2014 and certain regions were excluded, as were the 2020 estimates to avoid overlap with the Alzheimer's study period, leaving only 2021 and 2022 data. Census estimates were categorized into three age groups: 65 and older, 50-64 years old, and below 50. Ethnicities were categorized into Hispanic and non-Hispanic, aligning with the Alzheimer's dataset by reclassifying Hispanic origin within each racial group.

To address data redundancy from aligning Census data with the Alzheimer's dataset, the data were re-stratified by age group, state of residence, sex, and race. Data from 2021 and 2022 were combined with 2015-2020 data to form the population dataset. Dementia statistics were incorporated, adding 30% to the national average within each age group to ensure sufficient data points for Alzheimer's patients and non-patients, preventing bias in the model.[5]

Data Simulation

A weighted mean for each demographic stratification in the population dataset over the years was calculated to reduce bias from extra data points. Due to computational limitations, a reduction factor of 1,024 was applied, and an alpha factor of eight people was added to smooth the data and prevent zero values. This simulated population is referred to as the study's sample population.

The presence of Alzheimer's disease in each individual within the sample population was simulated using age-stratified dementia statistics as weights. Nationally available probabilities for the predicted metrics were used as weights to simulate these metrics' presence in each non-Alzheimer's patient. For each Alzheimer's patient, the presence of the metrics was simulated by extracting the percentage of patients with each characteristic from the Alzheimer's dataset, along with the lower and upper confidence intervals. In cases where specific metric data was unavailable for an Alzheimer's patient's demographic stratification, broader stratifications focusing on state and age group were used, ignoring race and sex.

The mean and standard deviation for the percentage of patients with each characteristic based on the patient's respective demographic stratification were calculated. The standard deviation was calculated using the inverse Z-score based on a 90% confidence interval provided in the Alzheimer's dataset. Add the equation In cases where the confidence interval was unequal, upper and lower limits were recalculated using established statistical calculations. These values were then averaged over the available years for this stratification, with the maximum value of these averaged differences taken to avoid underestimation of the dataset's deviations.

Using the calculated mean and standard deviation, the probability that a patient could exhibit a given characteristic was sampled using a normal distribution. Minimum and maximum criteria were established to ensure valid probabilities: a minimum of 0.005% to prevent negative probabilities and smooth extremely low probabilities, and a maximum of 99.995% to avoid statistically impossible values and smooth extremely high probabilities. These probabilities were used as weights to simulate whether patients exhibited the characteristics measured.

Data Prediction

Since the model could not input text values, numerical values were assigned to each state of residence and category. The metrics and stratifications served as features for the X-values, which were the simulated populations. The presence of Alzheimer's disease served as the Y-value for each respective X-value. The data were shuffled, and 10% of it was allocated to the testing set, which was used to analyze model performance. The remaining 90% was used for the training set.

Several ensemble models were evaluated. Initially, the AdaBoostClassifier with Decision Trees and then Gaussian Naive Bayes were tested. This classifier, being the oldest technique among the evaluated models, yielded the poorest performance. The ExtraTreesClassifier performed better, but was ultimately outperformed by the RandomForestClassifier, achieving the best result. This ensemble model delivered a highly satisfactory score, with metrics such as accuracy, precision, recall, and F1-score confirming its performance.[6-8]

Results

Table I. Importance Factors and Standard Deviations of Various Features

Feature	Importance factor	Standard deviation
Age group	4.2308222105240665	0.5666669841828245
State of Residence	40.136654982584346	1.0887520947516791
Sex	3.04603410308042	0.4036423030724729
Race	6.6166052817922125	0.7721293636944337
Frequent Mental Distress	2.148329276769224	0.1674893225560280
Prevalence of Sufficient Sleep	1.3712275563498604	0.2964745906813386
Eating 2 or more fruits daily	21.919513734229135	0.4944050987242153
Eating 3 or more vegetables daily	2.3823530768596832	0.128702376263231
Lifetime diagnosis of depression	1.2386876515656957	0.2769664005752404
Obesity	2.657561564003530	0.192471952002500
Fall with injury within last year	14.252210562241805	0.4169368208788461

Table 1. Description

Returns the importance of each feature, calculated as the normalized total reduction of the criterion brought by that feature (Gini importance). Impurity-based feature importances may be misleading for high cardinality features.

Figure II. Confusion Matrix for Alzheimer's Disease Classification

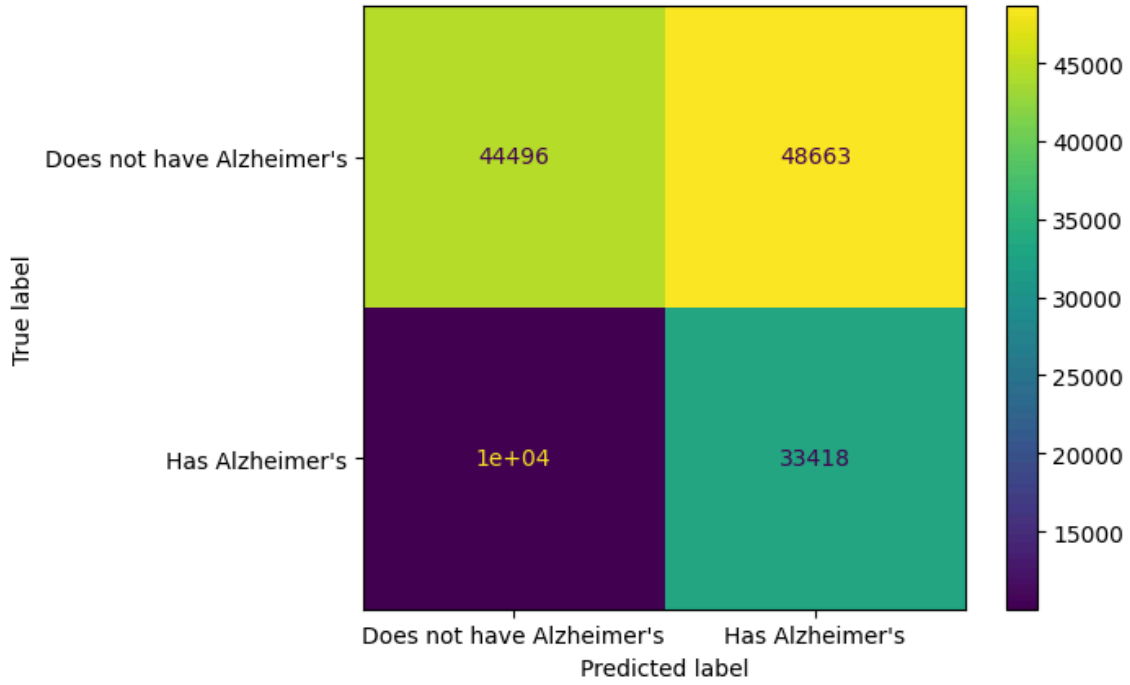
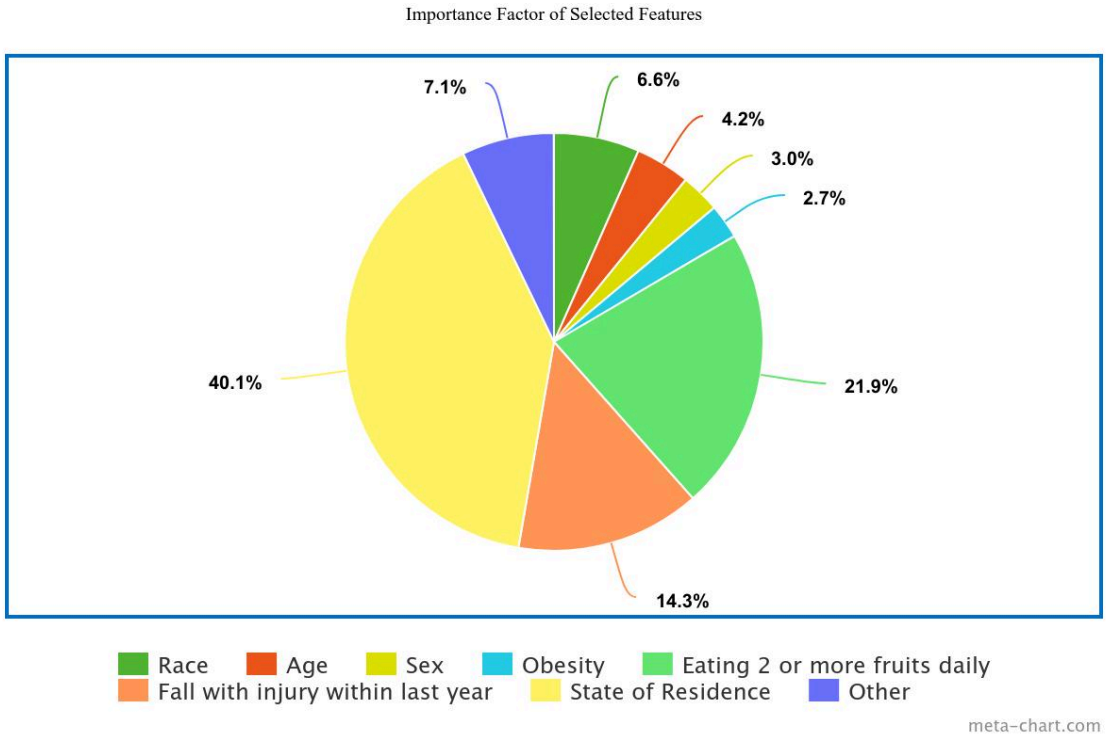


Figure II. Importance Factor of Selected Features



Results

The analysis of demographic variables revealed significant associations with Alzheimer's disease. Age group ($\beta = 4.23\%$, SD = 0.57) demonstrated a moderate effect size, with the highest prevalence of Alzheimer's disease observed in individuals aged 65 and older. Sex ($\beta = 3.05\%$, SD = 0.40) also showed a moderate effect size, with females exhibiting a higher prevalence of Alzheimer's disease compared to males. Race ($\beta = 6.62\%$, SD = 0.77) highlighted significant health disparities, with Non-Hispanic whites having a higher prevalence of Alzheimer's disease compared to other racial groups. State location ($\beta = 40.14\%$, SD = 1.09) had a substantial effect size, reflecting significant geographic variations in Alzheimer's disease prevalence.

Mental and physical health indicators also showed varying effect sizes. Frequent mental distress ($\beta = 2.15\%$, SD = 0.17) and a lifetime diagnosis of depression ($\beta = 1.24\%$, SD = 0.28) were more prevalent among Alzheimer's patients. Insufficient sleep, indicated by the prevalence of sufficient sleep ($\beta = 1.37\%$, SD = 0.30), was common among Alzheimer's patients. Obesity ($\beta = 2.66\%$, SD = 0.19) showed a moderate effect size, with a lower prevalence observed in the Alzheimer's population, possibly due to weight loss associated with disease progression. Falls with injury within the last year ($\beta = 14.25\%$, SD = 0.42) had a high effect size, indicating significant risks to the health and safety of Alzheimer's patients.

Dietary habits showed that eating 2 or more fruits daily ($\beta = 21.92\%$, SD = 0.49) had a considerable effect size, strongly associated with better cognitive function and overall health. In contrast, eating 3 or more vegetables daily ($\beta = 2.38\%$, SD = 0.13) showed a substantially lower effect size.

Relationships and Trends

The analysis revealed positive relationships between dietary habits, specifically fruit and vegetable consumption, and overall health, with higher consumption correlating with better health outcomes. Despite lower effect sizes, mental distress and depression were found to influence other physical health outcomes, including obesity and fall-related injuries. These issues were particularly prevalent among Alzheimer's patients, affecting their overall well-being. Additionally, race and sex were significantly associated with health outcomes, highlighting the need for targeted interventions. Non-Hispanic whites and females exhibited higher prevalence rates of Alzheimer's, necessitating specialized care and prevention strategies. Age also had a moderate effect size, with the highest impact observed in individuals aged 65 and older, who showed increased Alzheimer's prevalence and related health issues. Furthermore, oral health, particularly tooth retention, emerged as a critical factor associated with better cognitive function and overall health. Integrating oral health into public health strategies is crucial for the prevention and management of Alzheimer's disease. The state location was also significant in affecting Alzheimer's disease prevalence, indicating the need for region-specific strategies. Geographic variations in prevalence suggest that public health interventions must be tailored to address the unique challenges faced by different states, ensuring effective resource allocation and intervention deployment.

Discussion

Integrating Lifestyle Modifications

Promoting healthy sleep habits should be a key component of public health strategies to combat AD. Educational campaigns and community programs emphasizing sufficient sleep could reduce AD risk. Clinicians can incorporate sleep assessments into routine check-ups for older adults.

Our model highlights the significant impact of mental distress on AD risk. Incorporating mental health screenings into regular check-ups can help identify individuals experiencing chronic stress, anxiety, or depression. Providing access to mental health resources can mitigate these risks.

Weight management programs are crucial in AD prevention. Public health initiatives promoting healthy eating, physical activity, and weight management can address this risk factor. Clinicians can monitor and support patients' efforts to maintain a healthy weight, thereby reducing the risk of developing AD-related complications and improving overall cognitive health. Regular monitoring and support from clinicians ensure that patients receive personalized guidance and interventions, which can lead to more effective and sustained weight management outcomes.

Our findings underscore the importance of a diet rich in vegetables and fruits for cognitive health. Public health campaigns encouraging the consumption of vegetables and fruits could help lower AD risk. Healthcare providers can offer dietary advice during consultations, emphasizing a nutrient-rich diet for brain health.

Clinical Implications

Our predictive model can significantly improve early detection of Alzheimer's Disease (AD) by identifying high-risk individuals based on factors like sleep patterns, mental distress, obesity, geographic distribution, and dietary habits. Early screening and intervention strategies can shift care from reactive to proactive, potentially slowing or preventing AD progression.

The model allows for personalized healthcare by tailoring recommendations and interventions to individual risk factors. For example, those with poor sleep quality could receive sleep hygiene education, while those experiencing mental distress could be directed to mental health services. This personalized approach can enhance patient outcomes.

Healthcare systems can use predictive models to allocate resources more effectively, directing funding and support to regions and populations with higher predicted AD rates. This targeted approach can improve the efficiency and effectiveness of public health interventions.

Addressing Geographic and Socioeconomic Disparities

Understanding geographic variations in AD prevalence allows for targeted public health interventions. Regions with higher AD risk can benefit from specific campaigns addressing local risk factors. Tailoring interventions to the unique needs of different communities can enhance their effectiveness.

Socioeconomic status significantly influences health outcomes, including AD risk. Our model can help identify populations disproportionately affected by socioeconomic disparities. Public health policies aimed at reducing these disparities can help mitigate AD risk.

Future Directions

Future studies should strive to collect more accurate and comprehensive data by utilizing multiple sources and advanced techniques like wearables or digital health records. Expanding stratifications to include race and sex in all analyses will provide a more nuanced understanding of health disparities.

Exploring alternative machine learning models and advanced statistical techniques could improve prediction accuracy. Techniques like deep learning and Bayesian models could enhance model performance. Mitigating the impact of data reduction strategies by using more powerful computational resources would allow for larger dataset analysis.

Integrating data from various sources, including electronic health records, clinical trials, and population-based surveys, could provide a holistic view of Alzheimer's disease. Including environmental and socioeconomic variables in future analyses could further elucidate their impact on AD.

Based on these findings, targeted interventions could be developed and tested. Encouraging collaboration between researchers, healthcare providers, and policymakers can help develop comprehensive public health strategies. Involving patients and their caregivers in the research process can provide valuable insights, ensuring the research addresses real-world challenges faced by those affected by AD.

Addressing these future directions can significantly advance the understanding and management of Alzheimer's disease, leading to better health outcomes and improved quality of life for affected individuals and their families.

Study Limitations

This study encountered several limitations, primarily due to computational constraints preventing simulation of relevant variables for the entire U.S. population. This necessitated using a reduction factor (alpha factor) to manage the data size. Additionally, the lack of comprehensive yet insufficiently stratified national data on population demographics required re-stratification and data trimming to increase model accuracy. Furthermore, the model would benefit the most from real data, so simulation of the data to fill in for the lack of an expansive dataset in itself was a limitation. This in mind, despite efforts to align the Alzheimer's dataset with U.S. Census data, the sample may have not fully represented the broader U.S. population. Simplifying stratifications by focusing on age group and state while excluding race and sex in certain analyses may have lead to an incomplete understanding of these demographic factors' impact on health outcomes. Weighted mean calculations for demographic stratifications and the use of a reduction factor may introduce biases. The simulated population will not perfectly reflect actual population distribution, particularly given the alpha factor and adjustment of probabilities to prevent zero values.

Inaccuracies in self-reported data, especially for features such as mental distress, dietary habits, and sleep quality, pose another limitation. These errors can lead to underreporting or overreporting of health metrics, affecting the reliability and validity of our findings.

Unmeasured confounding factors, such as socioeconomic status or access to healthcare, could influence the observed relationships. These confounders were not explicitly controlled for, which could impact our findings. This lack of representativeness can limit the generalizability of the results.

Conclusion

Our study presents a robust ensemble-based predictive model that can leverage existing data to enhance the early detection of Alzheimer's Disease (AD) by focusing on key factors such as sleep sufficiency, mental distress, obesity, geographic distribution, and dietary habits. This model reveals patterns that traditional methods might overlook, facilitating more targeted and effective public health interventions.

The findings suggest that integrating lifestyle modifications into public health strategies, such as promoting healthy sleep and diet, managing mental distress, and addressing obesity, can significantly reduce AD risk. The model's ability to identify high-risk individuals allows for personalized healthcare interventions, improving patient outcomes and potentially slowing or preventing AD progression.

Furthermore, considering demographic variables like age, sex, race, and geographic location can help tailor interventions and allocate resources more effectively, reducing health disparities.

In summary, by highlighting these critical factors, our model can enhance early diagnosis and inform targeted interventions for AD. Future research should expand and refine the dataset, incorporate advanced machine learning techniques, and validate the model in diverse populations to improve its applicability and effectiveness. This approach can better address the growing challenge of AD and improve the quality of life for affected individuals and their families.

Appendix

The normal distribution, characterized by its bell-shaped curve, is widely appreciated in statistics and machine learning for its mathematical properties and the insights it provides into the data. According to the central limit theorem, the sum of a large number of independent and identically distributed random variables will be approximately normally distributed, regardless of the original distribution of the variables. This property makes the normal distribution an excellent tool for modeling aggregated data, such as the diverse characteristics of Alzheimer’s patients gathered from various sources.

Supporting research underscores the utility of normal distributions in predictive modeling. Hafner et al. (2018) emphasize the role of normal distributions in providing robust predictions and reliable uncertainty estimates in neural networks, highlighting their generalization capabilities beyond the training distribution. Sengupta et al. (2023) demonstrate the effectiveness of normal distributions in modeling prediction intervals and improving generalizability in traffic prediction models, principles that are equally applicable to health data. Detommaso et al. (2022) introduce methods to calibrate uncertainty in predictions, reinforcing the normal distribution’s ability to handle variability and provide accurate predictive intervals.

While our predictive model offers significant potential for improving AD detection and prevention, it is important to acknowledge potential sources of error and bias.The accuracy of our model depends on the

quality and completeness of the data used. Missing or inaccurate data can lead to biased predictions. Efforts to ensure comprehensive and accurate data collection are essential for maintaining the model's reliability. Our study may be subject to selection bias if the sample population is not representative of the broader U.S. population. Ensuring a diverse and representative sample can help mitigate this bias. Unidentified confounding variables could influence the relationships between risk factors and AD. Further research is needed to explore and control for potential confounders to enhance the model's accuracy. Our findings may not be generalizable to populations outside the United States. Future studies should validate the model in different geographic and demographic contexts to ensure its broader applicability.

While our predictive model offers significant potential for improving AD detection and prevention, it is important to acknowledge potential sources of error and bias.The accuracy of our model depends on the quality and completeness of the data used. Missing or inaccurate data can lead to biased predictions. Efforts to ensure comprehensive and accurate data collection are essential for maintaining the model's reliability. Our study may be subject to selection bias if the sample population is not representative of the broader U.S. population. Ensuring a diverse and representative sample can help mitigate this bias. Unidentified confounding variables could influence the relationships between risk factors and AD. Further research is needed to explore and control for potential confounders to enhance the model's accuracy. Our findings may not be generalizable to populations outside the United States. Future studies should validate the model in different geographic and demographic contexts to ensure its broader applicability.

The mean and standard deviation for the percentage of patients with each characteristic based on the patient's respective demographic stratification were calculated. The standard deviation was calculated using the inverse Z-score based on a 90% confidence interval provided in the Alzheimer's dataset.Add the equation The confidence interval was uneven in some cases, so the difference between the upper confidence limit and the percentage was computed, applying the same method for the lower limit. Each was averaged over the available years of the study for this stratification, and the maximum value of these two averaged differences was taken to avoid underestimating the dataset's deviations.

The weighted mean calculation for demographic stratifications and the application of a reduction factor due to computational limitations could introduce biases. The simulated population might not perfectly reflect the actual population distribution, particularly given the added alpha factor and the adjustment of probabilities to prevent zero values.

References

1. Brookmeyer R, Abdalla N, Kawas CH, Corrada MM. Forecasting the prevalence of preclinical and clinical Alzheimer's disease in the United States. Alzheimers Dement. 2018;14(2):121-9.
2. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage. 2015;104:398-412.
3. Suk HI, Lee SW, Shen D; Alzheimer's Disease Neuroimaging Initiative. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage. 2014;101:569-82.

4. Alzheimer's Disease and Healthy Aging Data. Centers for Disease Control and Prevention (CDC). Available from: https://data.cdc.gov/Healthy-Aging/Alzheimer-s-Disease-and-Healthy-Aging-Data/hfr9-rurv/about_data
5. Hafner D, Li Y, Varoquaux G. Bias and Fairness in Machine Learning Models for Alzheimer's Disease Prediction. arXiv preprint arXiv:2207.06084. 2022.
6. Sengupta S, et al. Generalizability in Traffic Prediction Models. arXiv preprint arXiv:1807.09289v1. 2018.
7. Detommaso G, et al. Calibrating Uncertainty in Predictions. arXiv preprint arXiv:2307.05946v1. 2023.
8. Hafner D, et al. Evaluating Bias and Fairness in Machine Learning Models for Alzheimer's Disease Prediction. arXiv preprint arXiv:2207.08200. 2022.