International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

Leveraging Big Data Analytics and Machine Learning to Identify Population-Level Risk Factors for Alzheimer's Disease

[1]*Ruchita Das
[1]Department of Clinical Informatics School of Graduate Studies University of Maryland, Baltimore, USA.

Abstract

Alzheimer's disease is a progressive neurodegenerative disease, which has implications for the health of people and is the key to targeted treatment. Interventions are the timely detection of the potential risk factors at the population level. The paper utilized big data analytics for self-reported BFRSS data for evaluating cognitive decline predictors in Alzheimer's disease. Software such as Apache was used to implement machine learning models, such as linear for scaling processed data. These include the regression, random forest, and clustering for identifying the higher-risk groups. The random Forest model displayed moderate predictive power (R 2 = 0.4697, RMSE = 17.35), outperforming Linear Regression. Clustering resulted in the identification of populations with mental health burden and geographic differences. Results emphasize the importance of mental health, socioeconomic status, and regional differences as risk factors for Alzheimer. Big data frameworks for population health analytics and supports targeted public health interventions.

Keywords: Alzheimer's disease, machine learning, Apache Spark, public health, BRFSS, predictive modeling, clustering.

1. Introduction

Alzheimer's is widely recognized as one of the leading causes of cognitive decline among the aging population. It has been a major challenge to the healthcare systems all over the world. The identification of risk factors of the disease at an early stage is of importance in prevention and intervention measures. The prior clinical research mainly depends on small and controlled datasets, hence limiting their ability to capture population-level trends.

With the increasing availability of large-scale health datasets, machine learning and big data analytics provide new opportunities to explore complex relationships between demographic, behavioral, and geographic factors. The Behavioral Risk Factor Surveillance System (BRFSS) offers a comprehensive dataset capturing self-reported health behaviors across the United States.

According to recent studies, there are approximately 6.7–6.9 million Americans who are aged 65 years and older and are surviving Alzheimer's dementia. This figure is projected to more than double to about 13–14 million by 2050–2060, in the absence of effective preventive therapies (Fu et al., 2025; You et al., 2022). AD is currently among the main causes of death in older adults, and the reported death cases are increasing by more than 140% since 2000, in contrast to a reduction in mortality from stroke, heart disease and HIV("2024 Alzheimer's disease facts and figures," 2024).

Numerous studies highlight cardiovascular risk factors, depression, sleep disturbances, metabolic conditions and lifestyle factors as the significant sources of cognitive decline and progression to AD (You et al., 2022). According to some studies, the variances in terms of race, ethnicity, geography and socioeconomic status are described as the Black and Hispanic older adults in the United States have a greater risk of Alzheimer's dementia than non-Hispanic White older adults and the prevalence of dementia and mortality varies across regions, with higher incidence often reported in the Midwest and Southern States (Weuve et al., 2017)

Additionally, the digitization of health information has created outstanding opportunities for the analysis of AD risk and progression using large scale datasets. Now, administrative claims, electronic health records (EHRs), national cohorts, and surveillance systems can explore longitudinal information on diagnoses, medications,

Ruchita Das
Department of Clinical Informatics School of Graduate Studies University of Maryland, Baltimore, USA.
Email: das.ruchita2211@gmail.com

Received: 11-Mar-2026 Revised: 13-April-2026 Accepted: 28-April-2026

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

laboratory values, social factors and health behaviors for a higher number of individuals (Park et al., 2020). Many ML and other methods have been used for predicting the incident AD, model progression from mild cognitive impairment (MCI), and identifying disease sub-phenotypes and trajectories. These methods have provided successful predictive performance, identified novel risk indicators and supported the potential of big data analytics for early detection in public health for the purpose of dementia (You et al., 2022).

Although some significant evidence indicates that AD threatens global health, there is limited access to large-scale, population-level survey data and distributed computing to simultaneously model cognitive impairment risk, anxiety, demographic and regional patterns (You et al., 2022), the currently used Machine Learning techniques are mainly based on clinical or healthcare datasets. They cannot be successfully used for population-based health planning, while the variance-focused research uses traditional methods that do not make full use of multidimensional data (Kramer et al., 2024). These limitations highlight the need for more comprehensive, data-driven approaches to support policymakers and public health professionals in identifying at-risk populations and designing targeted interventions.

Thus, this study aims to identify key predictors of cognitive decline associated with Alzheimer's disease and evaluate the effectiveness of machine learning models in predicting risk patterns. Additionally, clustering techniques are used to segment populations and uncover hidden patterns that may inform targeted interventions.

2. Materials and Methods

2.1 Dataset & Preprocessing

The study utilizes a dataset from the CDC regarding Alzheimer's Disease and Healthy Aging (2015–2022), which comprises 284,142 records across 31 variables. It consists of demographics, health variables and geographical variables. For constructing the analytical environment, software Apache Spark 3.3.0 was used, that helped in the removal of duplicates and missing values in core columns, which were processed in the data cleaning, ultimately leading to a refined dataset of 169,435 records. The most significant variables were demographics (LocationDesc, LocationAbbr, YearStart, YearEnd), health indicators (Class, Topic, Question, Data, Data-Value-Type), statistical (Low-Confidence-Limit, High-Confidence-Limit) variables, stratification (Stratificationcategory1, Stratification1, Stratificationcategory2, Stratification2), geographic (Geolocation) and other identifier (RowId, ClassID, TopicID, QuestionID, LocationID) variables. The dataset recorded health indicators in various arenas, such as mental health, general health, caregiving, cognitive deterioration, screenings and vaccinations, nutritional and physical exercise, and smoking and alcohol consumption.

2.2 Exploratory Data Analysis

Exploratory data analysis of numeric columns, count, mean, standard deviation, and percentile were calculated using the in-built support functions of Spark. To evaluate the relationships between Data_Value, Low confidence limit and High confidence limit, correlation matrices have been produced. Frequency distributions were studied in the case of categorical variables in columns of Class, Topic, LocationDesc, Stratification1 and Stratification2. After converting DataFrames to Pandas, Matplotlib and Seaborn are used for visualization of pie charts, bar plots, box plots, and stacked bar charts. These demonstrated how the health classes were distributed, the geographic differences in the availability of data, stratifying age groups, racial/ethnic representation, and how health topics are related to the range of values of data.

2.3 Feature Engineering and Modeling

Categorical variables (e.g., LocationDesc, Class, Stratification) were transformed using Spark's String Indexer, while rare categories (frequency <10) were consolidated to ensure model stability. VectorAssembler was used for integrating features into a single feature vector. The paper applies three modeling techniques, linear regression for establishing a baseline for variance explanation. Secondly, Random Forest regression (a nonlinear technique) was applied to capture nonlinear interactions with hyperparameter optimization through grid search (maxDepth = 10, numTrees = 50); Thirdly, the application of K-Means allowed the identification of natural subgroups within the population. K was selected based on interpretability and exploratory analysis. The evaluation of clusters was based on the demographics and the health-related characteristics.

2.4 Predictive Modeling

RandomSplit was used with a constant seed to split the dataset into training (80%) and testing (20%) sets to ensure reproducibility. Apache Spark MLlib was used to implement three machine learning techniques. For the

linear regression model, performance was evaluated using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). A random forest regressor was used to capture nonlinear relationships and interactions in the data. Maximum depth (maxDepth) and number of trees (numTrees) were optimized by means of grid search. The best values (maxDepth = 10, numTrees = 50) were determined according to the performance of the model. The identification of the natural subgroups in the data was executed with K-Means clustering with k = 3. The silhouette score was used to assess the performance of clustering. In all models, the assembled feature vector was input, and Data_Value was the target variable.

2.5 Model Evaluation

Model performance was evaluated using standardized metrics. The regression models used RMSE for determining the average prediction error, whereas R^2 was used for measuring the extent of variance explained by the model. For analyzing clusters, the silhouette score measures both cohesion and separation. Model performance metrics were used to compare the different approaches to evaluate the advantages of more complex algorithms and hyperparameter tuning. Random Forest model was used to assess the feature importance. It was assessed using Spark's built-in environment to ensure effective distributed computation and reproducibility.

3. Results

3.1 Descriptive Statistics

The descriptive statistics in Table 1 reflect the processed and cleaned data. As a result, the final dataset contained 169435 records, which spread across 10 main analytical domains. The major variables that were presented in this table are the data value, the lower and upper confidence limits. This offers the basic understanding of the central tendency and variability of the health indicators. Table 1 shows the key metrics for data values, which reveal a mean of 38.74% with a standard deviation of 23.70. The table further showed the difference between the minimum 0 and maximum 100 values, along with the interquartile range (25th percentile at 19.2 and 75th percentile at 55.2). This indicates the significance of variables in health indicators among the surveyed populations. The confidence limits portrayed a consistent distribution, with the average of lower and upper confidence limits being 33.80% and 44.03%, respectively.

Table 1 Descriptive Statistics of Healthcare Indicators

Metric	N	Mean	Std. Deviation	Min	25%	50%	75%	Max
Data_Value	169,435	38.74	23.69	0.0	19.2	34.0	55.2	100.0
Low_Confidence_Limit	169,435	33.80	22.97	0.0	15.3	28.2	47.7	99.6
High_Confidence_Limit	169,435	44.03	24.39	1.3	23.5	40.3	63.1	100.0

For evaluating the internal consistency and interdependency of the numeric variables, the correlation matrix was generated by Spark's statistical function, as shown in Figure 1, which analyses the correlation heatmap, reveals the exceptionally strong positive linear relationships among the metrics.

Particularly, it shows that the Pearson correlation coefficient between data value and the lower confidence limit was 0.98, whereas the correlation coefficient between data value and the higher confidence limit was 0.97. In addition to this, the correlation between lower and upper confidence bounds was observed to be 0.96. These higher coefficients indicate the primary health outcome to be tightly coupled with their statistical or mathematical confidence intervals. The observed trends across demographic and geographic groups are statistically reliable and internally coherent.

3.2 Geographic Patterns

Figure 2 shows the geographic distribution of Alzheimer-related health data frequencies across the United States and its territories. It demonstrates the patterns that reveal the significant regional disparities in health surveillance and disease burden: (1) High-Frequency Regions: The most numerous health indicators were recorded in the Northeast (e.g., New York, Connecticut, Pennsylvania), certain Southern states (e.g., Oklahoma, Texas, Arizona), California, and the District of Columbia: (2) Low-Frequency Regions: Conversely, lower frequencies were observed in US territories (e.g., Guam, Puerto Rico, Virgin Islands), several Western states (e.g., Alaska, Hawaii, Nevada, Utah), as well as West Virginia and Maine. These patterns highlight critical public health concerns such as (1) Surveillance vs. Burden which is observed through

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

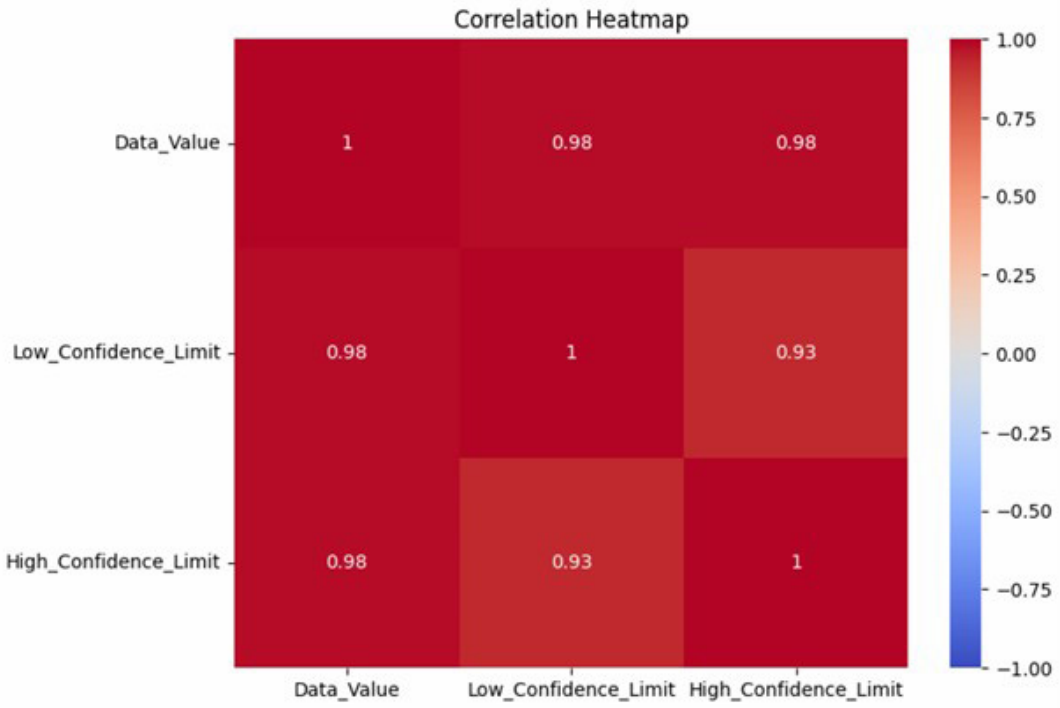

Figure 1 Correlation matrix

the trends that indicate either actual regional differences Alzheimer's or variations/prevalence in the activity levels of regional health surveillance systems; (2) The "Stroke Belt" Connection which shows geographic analysis that identified a high concentration of risk factors in the Southeastern United States. Specifically, states within the "Stroke Belt", including Alabama, Mississippi, Louisiana, and Arkansas, exhibited the highest levels of Alzheimer-associated outcomes, such as mental distress and activity limitations; (3) Socioeconomic and Clinical Drivers were spotted through these Southeastern hotspots for diagnosing patterns of cardiovascular (CV) disease. This suggests that vascular risk factors and limited healthcare access contribute to cognitive aging. While lower prevalence values in Western and Northeastern states (such as Colorado and Utah) are attributed to healthier lifestyle determinants and better healthcare accessibility, (4) It helped in the identification of the targeted intervention through these geographic hotspots which became essential to implement data driven and location specific public health intervention to address the intersection of CV health and cognitive or neurodegenerative decline.

3.3 Demographic Variances

Figure 3 shows the age-based stratification of the health indicator that is analyzed in the study. The relevant demographic variances are centered on the following key findings: First, Highest Burden of Disease: individuals aged 65 years and older exhibit the highest percentage of health indicators compared to other groups. This age sector is considered to face cognitive decline and its related risk factors; Second, Stratification Order defines data which reveals a clear trajectory in the frequency of recorded health indicators, giving first rank to the 65+ age group. The overall group follows it, and finally, the 50-64 years of age group. Third, Public health performance versus risks denotes the relevance to focus on the 65+ demographics. While this group shows the highest rate of screenings and vaccinations for reflecting the effective public health outreach for older adults yet, they also portray the largest average values for critical risk factors, which include high blood pressure, physical activity and subjective cognitive decline. Finally, these results suggest that older adults remain the primary target for data-driven interventions due to their elevated risk profile. Additionally, mental distress appears to be a significant predictor of cognitive decline, potentially contributing beyond the effects of age alone.

Furthermore, the factors were examined for the distribution of health indicators and their role across different population groups. Those factors are race, ethnicity, and gender. The health challenges have been

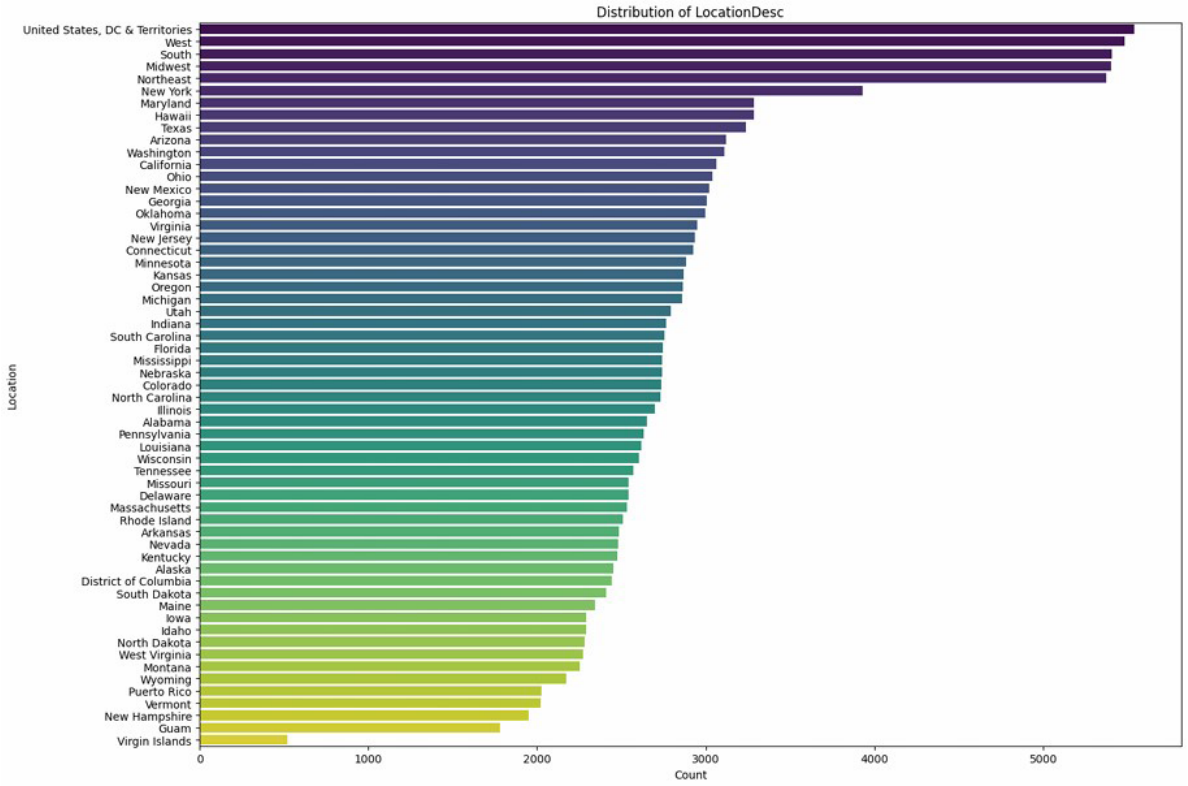

Figure 2 Geographic Distribution

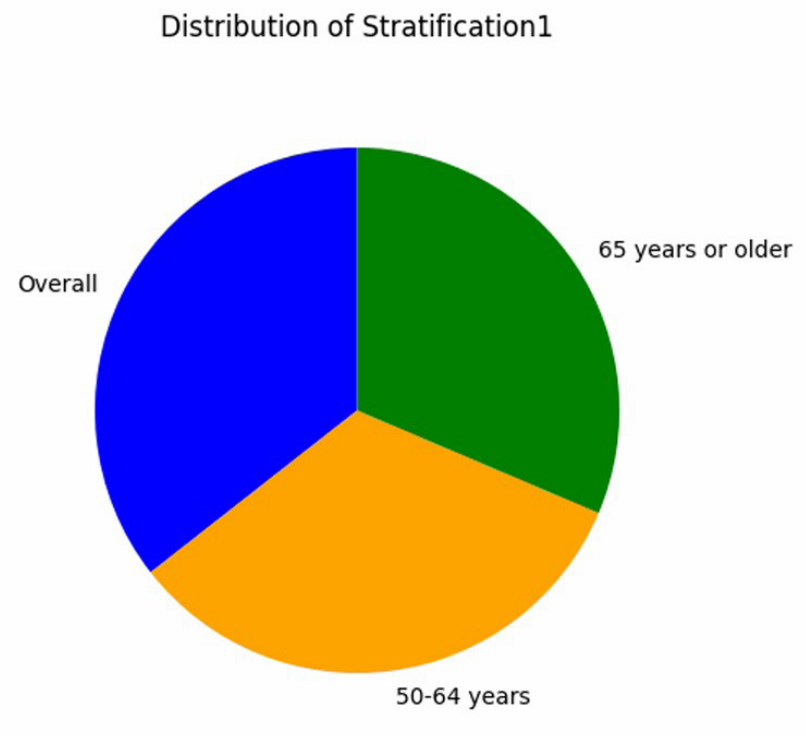

Figure 3 Age Demographic Stratification

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

portrayed and highlighted among these groups, as shown in Figure 4. It also demonstrates a clear hierarchy in the representation of racial and ethnic groups across all age brackets: (1) Highest Representation: White non-Hispanic individuals constitute the largest portion of the dataset; (2) This shows the gradual decrease in the representation of data through observed analysis with other groups Black non-Hispanic, Hispanic, Asian/Pacific Islander, and finally Native American/Alaskan Native individuals.

This suggests that black and Hispanic older people experience a higher risk of the disease compared to the white due to mental distress, functional limitations, and poor-rated health. These also exist across different disparities, which state the social determinant of health and limited or biased health care access to be the main factors.

3.4 Gender Stratification

Females, compared to men, report higher prevalence of disease-related issues, particularly those linked to mental health conditions and cardiovascular (CV) risk factors. On the other hand, males are associated with other factors of smoking and alcohol consumption, which contribute to increased health risks (Figure 4). The analysis helped in the development of culturally sensitive and community-based intervention strategies. Both genders faced different health concerns, allowing public health officials to address better the unique limitations these communities face in the early detection and prevention services of reported diseases.

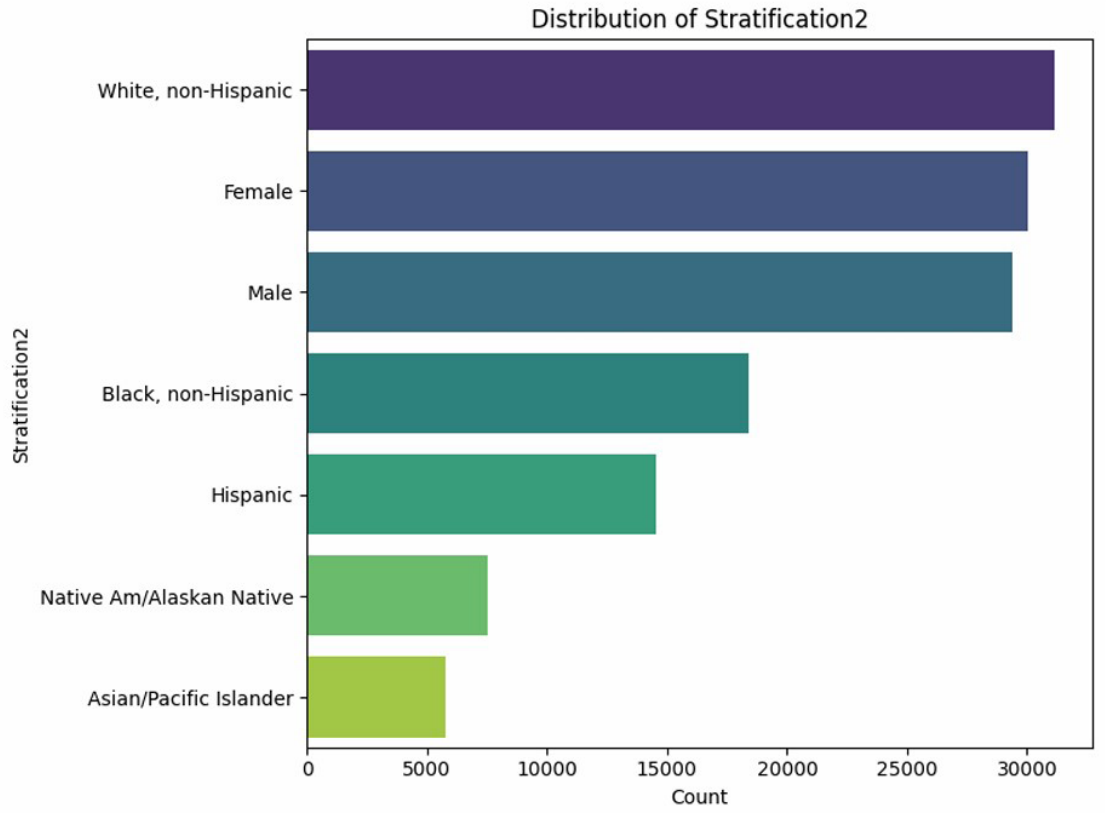

Figure 4 Racial, Ethnic and Gender Demographic Stratification

3.5 Interdependency Analysis

An interdependency analysis between the health categories and geographic locations, which reveals that mental health is the most represented health class across nearly the entire United States of America. Moreover, the data in Figure 5 indicates that health classes related to cognitive decline and caregiving are most dominant in the United States, characterized by a higher elderly population.

The distribution of health classes across specific age groups is shown in Figure 6. It examines the screenings and vaccinations to be significantly more prevalent in the 65+ age group. Figure 6 shows the lower consistency of health indicators related to smoking and alcohol

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

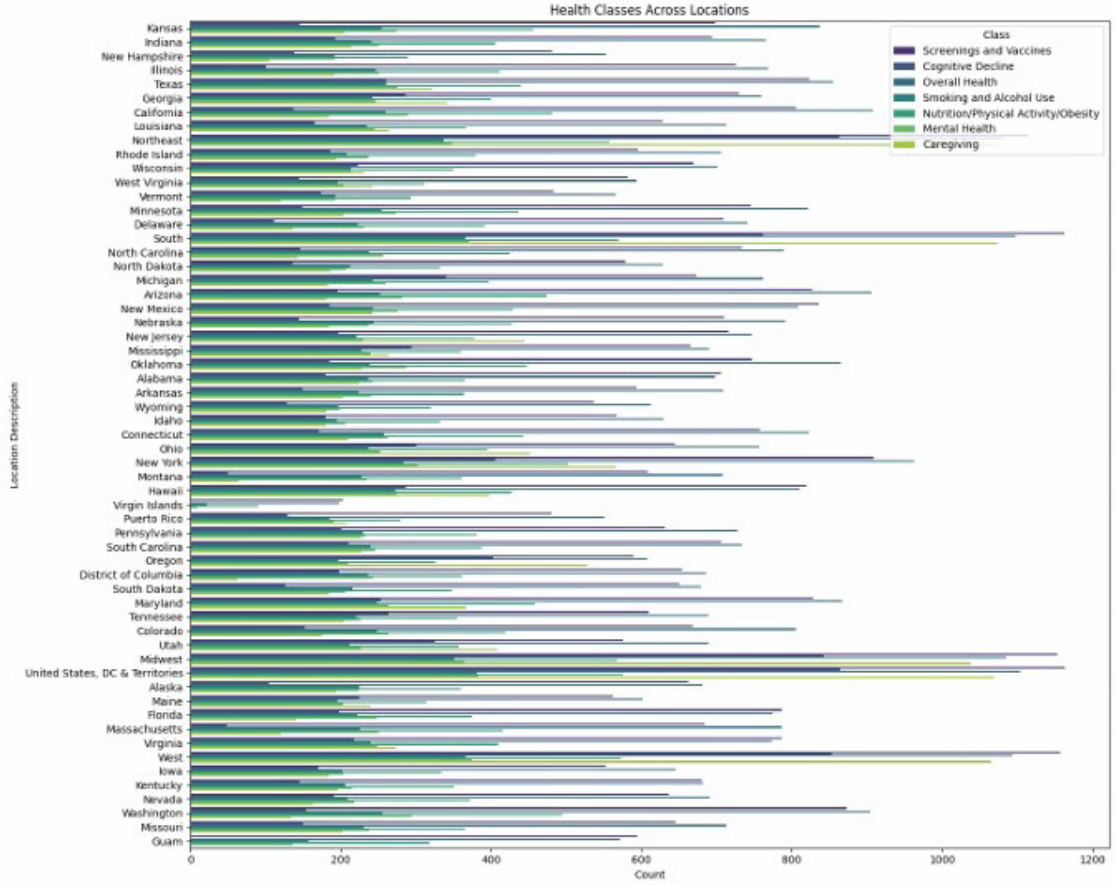

Figure 5 Health Classes vs Location

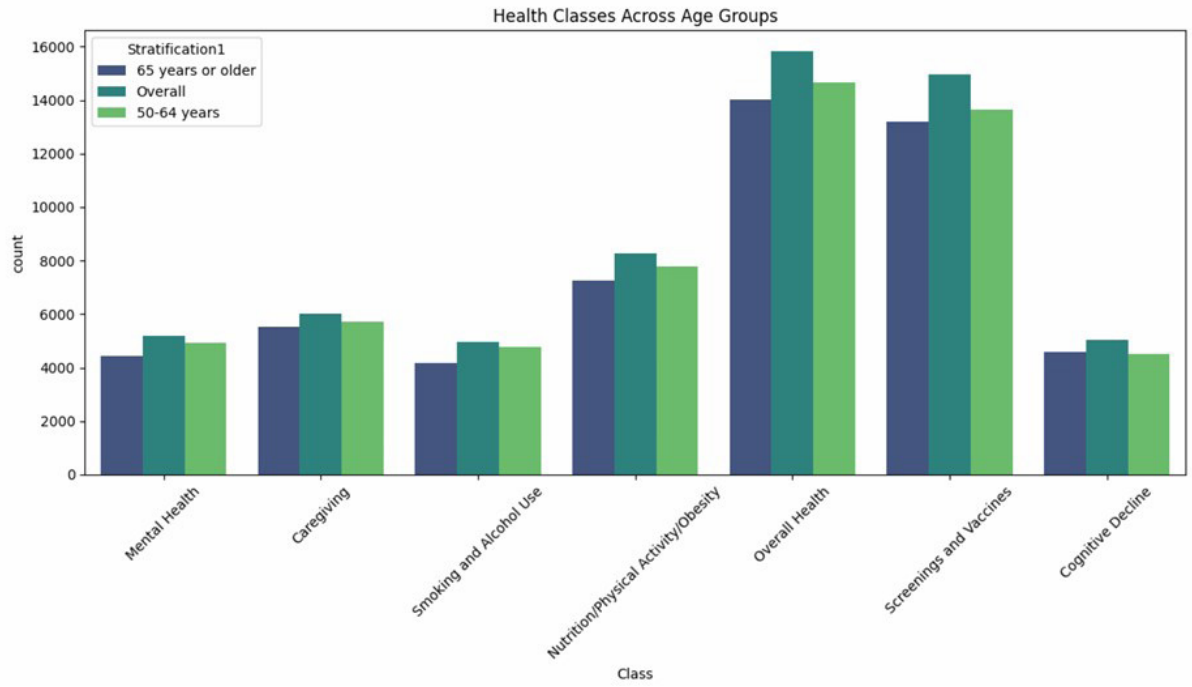

Figure 6 Health Classes by Age

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

consumption across all age groups.

Figure 7 shows the topics Distribution across Health Classes by Data Value Range, which illustrates how health topics are categorized based on their recorded data value ranges (Low, Medium, and High). The data for mental health has been explored, and it can be seen that data points fall within a lower range, whereas the chronic disease-related areas showed more equal distribution. This can be observed by data spread more equally across the low, medium and high ranges.

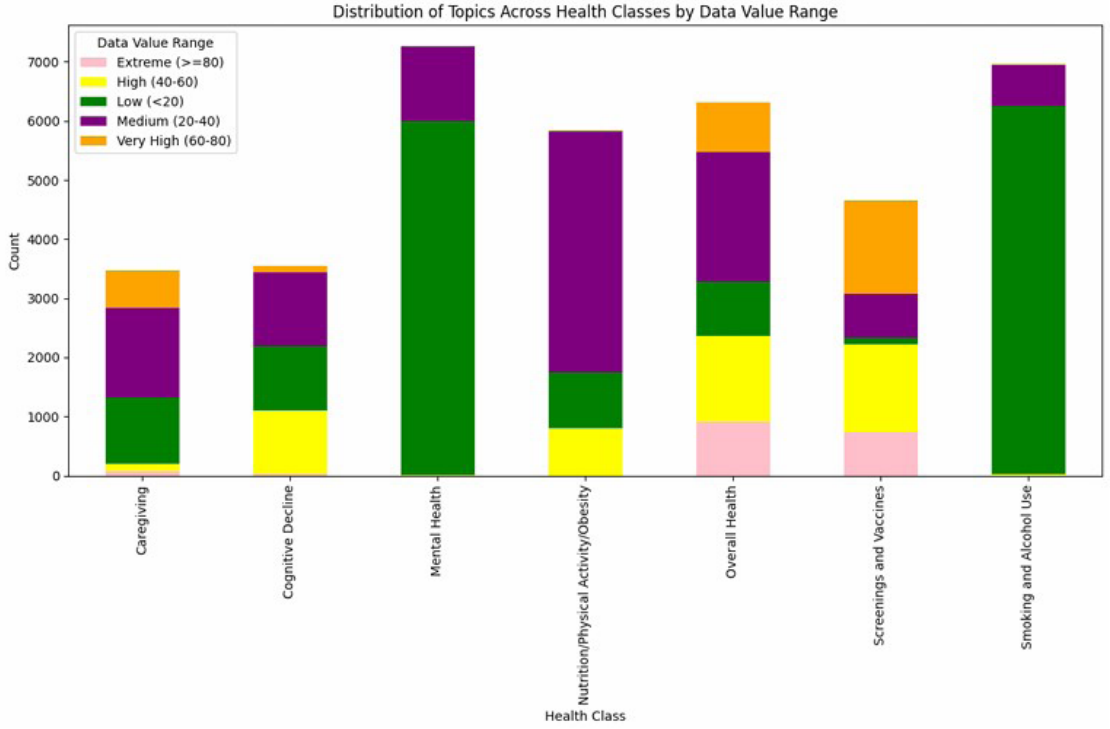

Figure 7 Topics Distribution across Health Classes by Data Value Range

The analysis offers four main critical areas in terms of public health benefits and clinical research. This addresses the demographic disparities for elevated risks among the black and Hispanic populations and the burden faced by them. The discovery of key predictors showed that mental distress is a strong factor in cognitive decline. The targeting of modifiable risk factors helped in providing focused care in the healthcare sector. The scalability of data processing implements the mass scale health analytics for population-level insights.

3.4 Predictive Model Performance

This study applied the predictive model for evaluating the performance of machine learning algorithms to predict future health outcomes based on historical data, for example, the likelihood of cognitive decline. The critical determination was followed by examining the risk factor of reported disease and the interrelationship among different health indicators. It was found that mental distress is a stronger predictor of cognitive decline than conventional predictors like age and ethnicity. Moreover, predictive modeling also gave a quantitative evaluation of the extent to which the variation in health outcomes might be approximated to have been explained by the chosen features, which is why it might be used in early detection and health planning.

To attain this, two main regression methods were applied, and they include Linear Regression and Random Forest Regression. The findings show that the ensemble learning model, Random Forest, was better than the Linear Regression model. Additional improvements in performance were achieved through hyperparameter optimization, as summarized in Table 2.

However, the R^2 meant that there remains much of the unexplained data which is expected given the use of self-reported BRFSS data. This introduced the potential noise, bias and response variability. In addition to this, the

Table 2 Comparison of Models

Model	RMSE	R2
Linear Regression	20.4591	0.2474
Random Forest (initial)	17.5905	0.4544
Random Forest (tuned)	17.3542	0.4697

outcomes are predicted at the population level, which is more complex compared to clinical controlled datasets with precise biomarkers. Despite these barriers, the importance of models remains the same, as they identify broader population trends and risk patterns. Such insights are particularly useful for implementing public health planning and early intervention rather than individual-level prediction.

3.5 K-means clustering

This is the machine learning algorithm for partitioning data into distinct, non-overlapping subgroups. It was utilized for identifying three natural risk profiles within the dataset. It helps in understanding the data better. The analysis of clustering results has been valuable for designing data-driven and targeted interventions aimed at high-risk populations. K = 3 was selected as the ability to complement the regression models. As shown in Figure 8, clustering results reveal distinct patterns in subgroups enabling the public health officials to move beyond the broader generalization for developing interventions for each cluster.

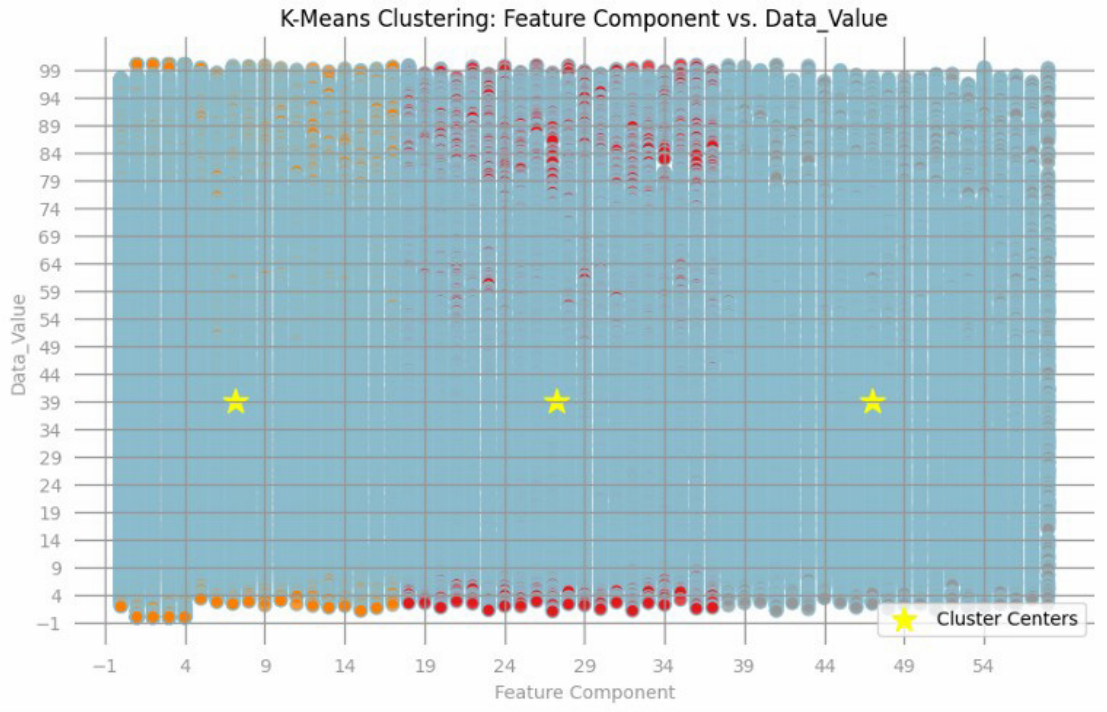

Figure 8. Clustering of Feature Component vs. Data_Value

Figure 8 plots the first component of the feature vector (which includes indexed data for location, health class, and age/racial stratifications) on the x-axis against the Data_Value (the health indicator percentage) on the y-axis. Where color-coded clusters: The data points are divided into three distinct clusters, each represented by a different color. This separation demonstrates that there are significant variations in health indicator patterns across the dataset. The yellow stars in the plot represent the centroids or "cluster centers". These centroids indicate the average feature values and Data_Value for all points within that specific cluster, serving as the mathematical "middle" of each subgroup.

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

3.6 Statistical Significance and Findings

The quality of the clustering model was evaluated using the silhouette score, which was found to be 0.7360.. Figure 1 shows (Largest): 57,297 data points (33.8%), representing the most common risk profile in the dataset. Cluster 0: 43,683 data points (25.8%). Cluster 2: 35,869 data points (21.2%). This means that the cluster is compact and well separated, and the individuals within each group share similar characteristics, although they are different from other groups. The population distributions show the reported characteristics for 169435 records, which were analyzed. Clustering is useful for promoting the interpretability of data for developing targeted public health interventions. They identify the specific high risks in the groups as compared to depending on generalized trends. Consequently, K-Means clustering proves useful for uncovering of hidden patterns in population health data.

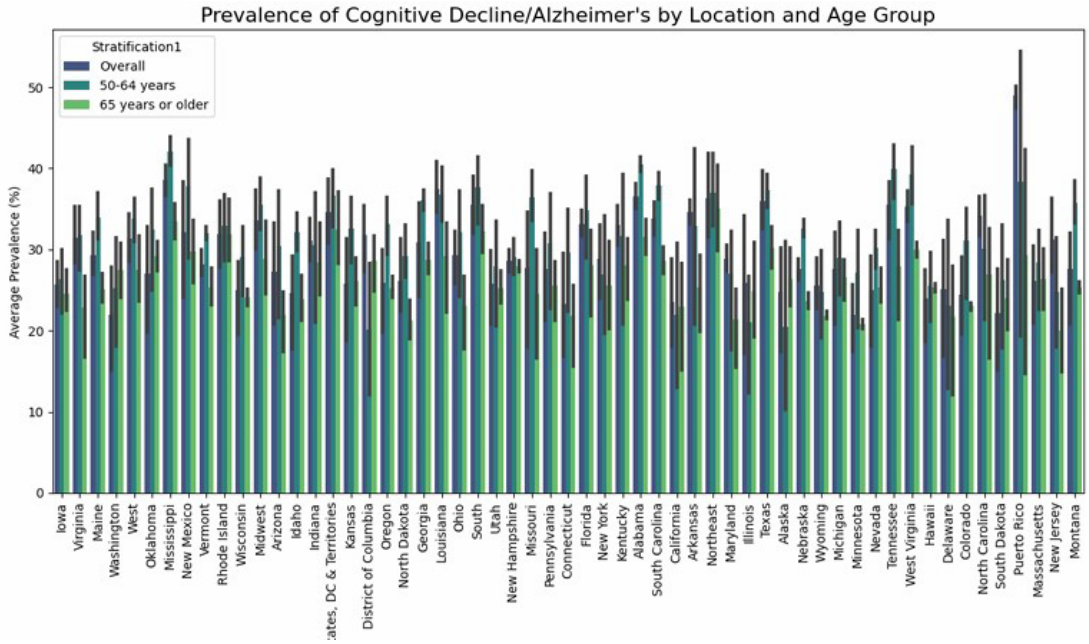

Figure 9. Cognitive Decline Prevalence

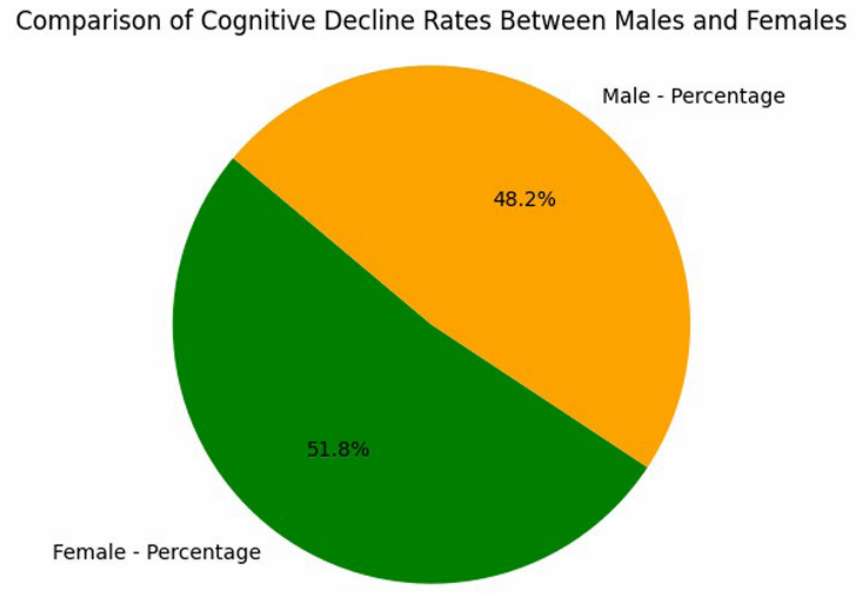

Figure 10. Cognitive Decline Rates in Males Vs Females

3.7 Cognitive Decline Analysis

Figure 9 demonstrates the significant geographic differences in cognitive decline across the United States. The highest prevalence of cognitive decline was reported in the Southeastern region, such as Alabama, Mississippi and Louisiana. Whereas, the states with the lowest prevalence are reported to be Colorado, Utah and Minnesota. The trends and health disparities in cardiovascular (CV) have been corresponding with the geographic pattern shown in Figure 9.

The analysis shows that males (12.1%) and females (12.8%) had almost similar prevalence rates, as shown in Figure 10.

Although females have slightly higher rankings, the results show that gender does not play a significant role in prevalence as compared to other variables such as geography or clinical risk factors. The critical examination of clinical and demographic risk factors was carried out for the reported disease in various age groups. Figure 11 shows the average prevalence of different health topics. This involves the mental health, physical exercise and clinical screening. Individuals aged 65 years and older exhibit the highest average values across health indicators. In contrast to this, the age group of 50 to 64 years and the general population show a two-fold tendency. While, the 65+ group is more engaged in positive health behaviors, including screenings and vaccinations, it also carries a greater burden of chronic risk factors.

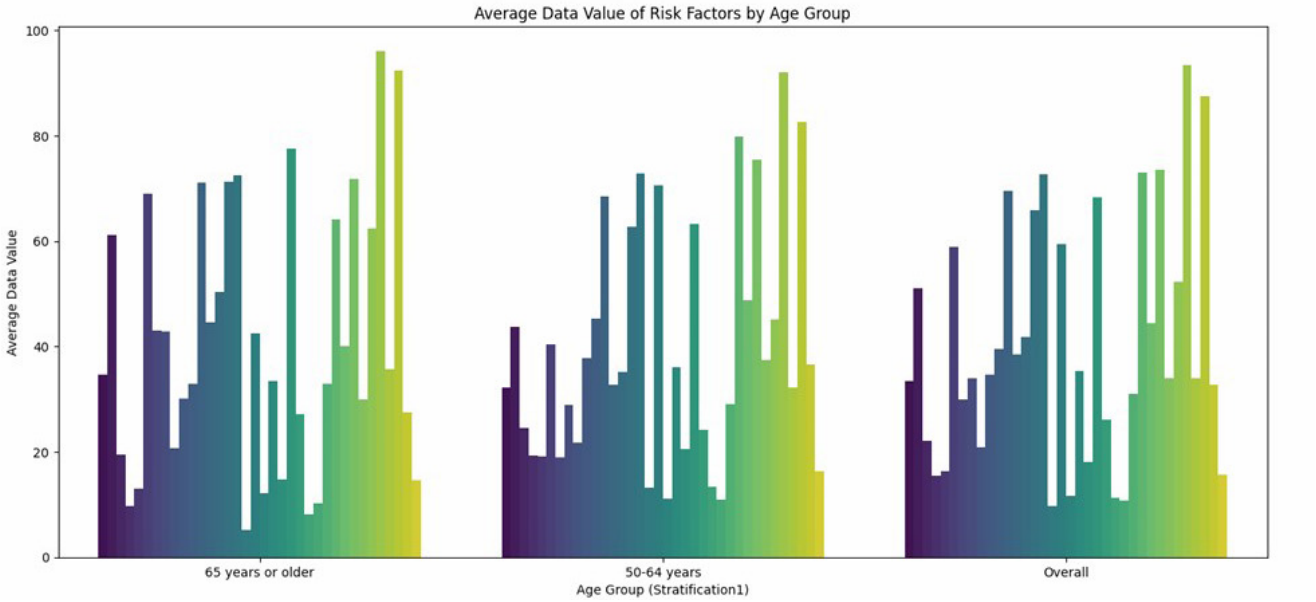

Figure 11. Plotting of Average Data Value of Risk Factors by Age Group

Figure 12 shows the analysis of particular clinical and behavioral issues. Among the high-risk population (65 years and older), the figure reveals significant highs in three categories: high blood pressure, physical inactivity (no leisure-time physical activity), and subjective cognitive decline.

In addition, Table 3 also discusses the prevalence of particular modifiable health behaviors and their strong relationship with cognitive decline. High blood pressure shows the highest prevalence (48.3%), followed by obesity (32.7%) and physical inactivity (28.6%). Other notable risk factors include a history of depression (18.9%), current smoking (15.8%), and subjective cognitive decline (12.4%). These are interesting results that highlight the significant role played by lifestyle and vascular factors that are amenable to change in affecting the risk of Alzheimer's disease, which forms a strong empirical basis to support specific health-based intervention in the population to reduce age-related cognitive impairment.

The summary of critical analysis presented in Table 3 highlights key factors associated with cognitive decline. Individuals aged 65 years and older represent the most affected population group. They have become an important group for reported diseases. They are reported to have high

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

Figure 12. Risk Factors

Table 3. Relationship between Health Behavior and Average Data Value (%)

Health Behavior	**Average Data Value (%)**
Subjective Cognitive Decline	12.4
No leisure-time physical activity	28.6
Obesity	32.7
Current smoking	15.8
High blood pressure	48.3
Lifetime diagnosis of depression	18.9

blood pressure and physical activity. This study offers insight into the direction of preventive healthcare strategies for slow cognitive aging. In particular, discrepancies are observed between higher vaccination rates and persistent clinical risks, for example, obesity and hypertension. This reveals the gap in management of chronic disease and calls for the need for intervention strategies. This also includes the contrast across genders for supporting the monitoring of disparities. However, the same gender reports similar results. These results leverage a data-driven basis to develop culturally sensitive and community-based intervention strategies. This further involves the early detection approaches for reducing the burden of cognitive decline.

4. Discussion

This study provides important insights into Alzheimer's-related risk factors across geographic, demographic, and clinical dimensions, highlighting patterns of cognitive decline among older adults in the United States. States such as Alabama, Mississippi, Louisiana, and Arkansas have the highest prevalence of health outcomes related to Alzheimer's. This includes the mental distress, activity limitation and poor self-rated health. This pattern is consistent with the well-documented "Stroke Belt" region and reflects underlying disparities in cardiovascular (CV) health and healthcare access, may contribute to both cardiovascular and cognitive decline (Sawyer et al., 2023). This helps in emphasizing the needs of targeted public health intervention strategies, which are significant for addressing the intersection of cardiovascular (CV) health and cognitive aging, since higher rates of cognitive impairment are reported in these areas. Whereas, the countries with lower prevalence are Western and Northeastern states, such as Colorado, Minnesota, and Utah. This is most likely due to lifestyle differences, access to healthcare facilities and socioeconomic conditions (Jia et al., 2020).

The black population are more likely to suffer from cardiovascular (CV) factors compared to the white aging population (Gupta, 2021). These inequalities prevailed across geographic regions. This indicates that the systematic factors, for example, unequal preventive care access, social determinants of health and potential genetic differences, play an important role in cognitive outcome. The mandatory and necessary culturally tailored interventions are due to the increased risk among the minority populations. This addresses the specific barrier for early detection and preventive services (Gupta, 2021).

The age group analysis confirms that the population above 65 years has a greater burden of cognitive decline and related risk factors. The highest average data value is reported in this age group for all health indicators (Wooten et al., 2023). Though this age group demonstrates higher representation in screenings and vaccinations, it indicates the successful public health outreach. Problems or barriers like physical inactivity and high blood pressure are common. The risk of disease has increased due to mental stress, which has stood as a powerful predictor of the disease. This aligns with the increased acknowledgement of a bidirectional link between mental health and cognitive function, where chronic stress, depression, and anxiety may accelerate cognitive decline through neuroinflammatory and physiological pathways, while early cognitive impairment may also contribute to psychological distress. The findings show the need for the incorporation of mental health diagnostic tools into standard cognitive evaluations (Wooten et al., 2023).

4.1 Public Health Implications

This study has identified key indicators related to Alzheimer's, demonstrating the clear geographic and demographic variations. These patterns are closely linked to underlying cardiovascular (CV) risk factors and disparities in healthcare access found in the Southeastern Stroke Belt belt states. On the other hand, several western and northeastern states have exhibited lower burdens, may be attributed to higher economic status with better access to preventive care. The healthier lifestyle factors are linked to reduced CV risks (Dubois et al., 2020). These findings suggest that the healthcare access, vascular risk profiles and neighborhood conditions can shape the cognitive health. The critical disparities have been persisting across racial and ethnic lines. Greater mental stress, functional limitations and poorer self-rated health have been reported in older Hispanic and non-Hispanic adults compared to non-Hispanic White populations, even though the geographic location has been adjusted. The evidence that mental health and cognitive decline are strongly related, making depression, anxiety and chronic stress major factors contributing to Alzheimer's through neuro inflammatory and vascular mechanisms (Byers & Yaffe, 2011). Moreover, the machine learning risk models, such as $R^2 = 0.47$, may help in the identification of high-risk populations. Yet external validation is critical for ensuring that they do not exacerbate existing disparities.

4.2 Methodological Considerations and Limitations

This study has considerable amont of limitations. The self-reported data can introduce bias, which can underestimate the true disease burden (Xie & Wang, 2020). Additionally, missing data may introduce selection bias, particularly if the data are not missing at random. The cross-sectional nature of the study does not allow causal inferences. The longitudinal data should be utilized for understanding the disease to identify the relation between risk factors and the disease itself. The lack of genetic markers, biomarkers and clinical information limits the predictive performance of the machine learning algorithm. The dataset-specific nature of the analysis limits the findings of research since it cannot be generalized to other datasets (You et al., 2022).

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

Machine learning algorithms and pipelines provide scalable and efficient approaches for extracting and collecting data for Alzheimer's datasets. In contrast, the hyper parameters of this research have produced modest improvements. The absence of external validation in machine learning predictive algorithms for disease has been a barrier to clinical generalization. The findings indicate that strong clustering and model performance do not necessarily translate into clinical utility, underscoring the importance of biological and clinical validation (Jahan et al., 2023). The future implication should include the embedding of EHR records, which include the genetic data, biomarker, longitudinal tracking and ongoing model refinement. This helps in the alignment with the data and trends of the coming generation to predict and mitigate risk in patient care of reported diseases.

4.3 Future Recommendations

Future recommendations include the research or studies to move beyond the survey-based data for integrating the multiple approaches with the detailed clinical assessments. This includes the neuroimaging, cognitive testing and cerebrospinal fluid biomarkers (Vilkaite et al., 2024).

The combination of data is recommended. This research utilized Spark, which is a distributed computing framework for combining the biological data with demographic and behavioral information. It improves the individualized risk prediction and biomarker discovery while including the genetic markers such as APOE4. This helps in modifiable lifestyle risks (Bi et al., 2020).

The future research should include longitudinal studies or designs for establishing the causal links and evaluation of the modifiable factors that influence the cognitive decline over time, for example, physical activity, smoking and hypertension management (Jia et al., 2020; Welberry et al., 2025). Advanced machine-learning approaches, including deep learning models (CNNs, RNNs, and transformers) (Akter et al., 2022), can be employed for capturing complex and nonlinear patterns, which are often overlooked by traditional models. However, the real-time surveillance systems integrate with geospatial analytics can support continuous monitoring of emerging trends and risk patterns or it can strengthen disease support community-based intervention strategies (Mar et al., 2020).

5. Conclusion

Alzheimer is a leading cause of cognitive decline and poses a significant threat to healthcare systems. Machine learning models have been utilized in this research to analyze the larger-scale population health data for identifying the risk factors of the disease. This study analyzed over 284,000 records from the Behavioral Risk Factor Surveillance System (BRFSS) and identified significant geographic and demographic disparities, including high-risk concentrations in the Southeastern United States, notable racial and ethnic differences and a strong link between mental distress and cognitive decline. Random forest demonstrated moderate performance (R^2 = 0.4697; RMSE = 17.35), while K-Means clustering enabled the identification of distinct population-level risk profiles. This helped in the identification of various risk profiles to enhance interpretability.

The findings include implications for public health needs. They highlighted the need for a potential intervention to be created for targeted location-specific populations. Those that are culturally sensitive and allow the incorporation of the mental health screening into geriatric care. In light of limitations such as the exclusion of longitudinal and self-reported data sets and the lack of genetic and multimodal clinical markers. This study aims to provide a framework to employ the utilization of distributed computing to analyze complex health datasets. Future research should focus on incorporating longitudinal designs and the incorporation of multiple approaches. This also includes the embedding of the latest machine learning models to improve the predictive accuracy. These findings supports early detection and prevention strategies to improve the management of Alzheimer's disease and concludes the potential of big data analytics and machine learning to understand diseases and invent new healthcare approaches.

Declaration

Author contribution: The author has contributed to all aspects of this research work, including in its conceptualization, data collection, analysis, and interpretation of results. The author shares equal responsibility in collecting the research articles, papers, or academic journals to prove their validity, designing research methodologies, and conducting analysis from the CDC and BRFSS datasets.

Conflict of interest: The author does not pose any connection or linkage, such as a financial or commercial relationship, that conflicts with their interests regarding this research work. This work was not influenced externally by

any corporate entities or interests to avoid bias and improve data interpretation.

Funding: This research was not supported by any grants or funds. The author independently supported the finances of data acquisition and analysis, and the preparation of the manuscript.

Ethical approval: The ethical approval or informed consent was not required since the data were publicly available from the CDC and BRFSS centres. Hence, it contains no personal information and is freely accessible to the public.

Consent for publication: The manuscript has not incorporated any personal details of participants, which is why it has not asked for consent for publication from participants.

Availability of data and materials: The data were retrieved and analyzed from Data.gov "Alzheimer's Disease and Healthy Aging Data" at https://catalog.data.gov/dataset/alzheimers-disease-and-healthy-aging-data.

References

2024 Alzheimer's disease facts and figures. (2024). Alzheimer's & Dementia, 20, 3708-3821. https://doi.org/10.1002/alz.13809

Akter, S., Das, D., Haque, R. U., Tonmoy, M. I. Q., Hasan, M. R., Mahjabeen, S., & Ahmed, M. (2022). AD-CovNet: An exploratory analysis using a hybrid deep learning model to handle data imbalance, predict fatality, and risk factors in Alzheimer's patients with COVID-19. Computers in Biology and Medicine, 146, 105657-105657. https://doi.org/10.1016/j.compbiomed.2022.105657

Bi, X.-A., Hu, X., Wu, H., & Wang, Y. (2020). Multimodal Data Analysis of Alzheimer's Disease Based on Clustering Evolutionary Random Forest. IEEE Journal of Biomedical and Health Informatics, 24, 2973-2983. https://doi.org/10.1109/jbhi.2020.2973324

Byers, A. L., & Yaffe, K. (2011). Depression and risk of developing dementia. Nature Reviews Neurology, 7(6), 323-331. https://doi.org/10.1038/nrneurol.2011.60

Dubois, P., St-Pierre, M.-C., Desmarais, C., & Guay, F. (2020). Young Adults With Developmental Language Disorder: A Systematic Review of Education, Employment, and Independent Living Outcomes. Journal of Speech, Language, and Hearing Research, 63(11), 3786-3800. https://doi.org/10.1044/2020_JSLHR-20-00127

Fu, M., Sankararaman, S., Pasaniuc, B., Vossel, K., & Chang, T. (2025). Identifying common disease trajectories of Alzheimer's disease with electronic health records. eBioMedicine, 118. https://doi.org/10.1016/j.ebiom.2025.105831

Gupta, S. (2021). Racial and ethnic disparities in subjective cognitive decline: a closer look, United States, 2015–2018. BMC Public Health, 21. https://doi.org/10.1186/s12889-021-11068-1

Jahan, S., Taher, K. A., Kaiser, M., Mahmud, M., Rahman, M. S., Hosen, A., & Ra, I.-H. (2023). Explainable AI-based Alzheimer's prediction and management using multimodal data. PLoS One, 18. https://doi.org/10.1371/journal.pone.0294253

Jia, L., Du, Y., Chu, L., Zhang, Z., Li, F., Lyu,

International Journal of Medical Discoveries
Volume 2, Issue 2
ISSN: 3067-7912

D., Li, Y., Zhu, M., Jiao, H., Song, Y., Shi, Y., Zhang, H., Gong, M., Wei, C., Tang, Y., Fang, B., Guo, D., Wang, F., Zhou, A.-H., . . . Jia, J. (2020). Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China: a cross-sectional study. The Lancet. Public health, 5 12. https://doi.org/10.1016/s2468-2667(20)30185-7

Kramer, M., Cutty, M., Knox, S., Alekseyenko, A., & Mollalo, A. (2024). Rural–urban disparities of Alzheimer's disease and related dementias: A scoping review. Alzheimer's & Dementia: Translational Research & Clinical Interventions, 11. https://doi.org/10.1002/trc2.70047

Lyall, D., Kormilitzin, A., Lancaster, C., Sousa, J., Petermann-Rocha, F., Buckley, P., Harshfield, E., Iveson, M., Madan, C., McArdle, R., Newby, D., Orgeta, V., Tang, E., Tamburin, S., Thakur, L., Lourida, I., Llewellyn, D., & Ranson, J. (2023). Artificial intelligence for dementia—Applied models and digital health. Alzheimer's & Dementia, 19, 5872-5884. https://doi.org/10.1002/alz.13391

Mar, J., Gorostiza, A., Ibarrondo, O., Cernuda, C., Arrospide, A., Iruin, Á., Larrañaga, I., Tainta, M., Ezpeleta, E., & Alberdi, A. (2020). Validation of Random Forest Machine Learning Models to Predict Dementia-Related Neuropsychiatric Symptoms in Real-World Data. Journal of Alzheimer's Disease, 77, 855-864. https://doi.org/10.3233/jad-200345

Park, J. H., Cho, H. E., Kim, J. H., Wall, M., Stern, Y., Lim, H., Yoo, S., Kim, H.-S., & Cha, J. (2020). Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. NPJ Digital Medicine, 3. https://doi.org/10.1038/s41746-020-0256-0

Sawyer, R., Worrall, B., Howard, V., Crowe, M., Howard, G., & Hyacinth, H. (2023). Methods of a Study to Assess the Contribution of Cerebral Small Vessel Disease and Dementia Risk Alleles to Racial Disparities in Vascular Cognitive Impairment and Dementia. Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease, 12. https://doi.org/10.1161/jaha.123.030925

Vilkaite, G., Vogel, J., & Mattsson-Carlgren, N. (2024). Integrating amyloid and tau imaging with proteomics and genomics in Alzheimer's disease. Cell Reports Medicine, 5. https://doi.org/10.1016/j.xcrm.2024.101735

Welberry, H., Jorm, L., Kiely, K., Huque, H., Peters, R., & Anstey, K. (2025). Sex and socioeconomic differences in 15-year prevalence trends for modifiable dementia risk factors in Australia: a cross-sectional, time series analysis—The Lancet. Healthy longevity, 100711. https://doi.org/10.1016/j.lanhl.2025.100711

Weuve, J., Barnes, L., De Leon, C. M., Rajan, K., Beck, T., Aggarwal, N., Hebert, L., Bennett, D., Wilson, R., & Evans, D. (2017). Cognitive Aging in Black and White Americans: Cognition, Cognitive Decline, and Incidence of Alzheimer's Disease Dementia. Epidemiology, 29, 151. https://doi.org/10.1097/ede.0000000000000747

Wooten, K., McGuire, L., Olivari, B., Jackson, E., & Croft, J. (2023). Racial and Ethnic Differences in Subjective Cognitive Decline — United States, 2015–2020. Morbidity and Mortality Weekly Report, 72, 249-255. https://doi.org/10.15585/mmwr.mm7210a1

Xie, D., & Wang, J. (2020). Comparison of self-reports and biomedical measurements on hypertension and diabetes among older adults in China. BMC Public Health, 20. https://doi.org/10.1186/s12889-020-09770-7

You, J., Zhang, Y.-R., Wang, H.-F., Yang, M., Feng, J., Yu, J., & Cheng, W. (2022). Development of a novel dementia risk prediction model in the general population: A large, longitudinal, population-based machine-learning study. eClinicalMedicine, 53. https://doi.org/10.1016/j.eclinm.2022.101665