Predicting Dementia Risk with Machine Learning

Min Seyun Chung Westergaard
2024

1. Abstract

The rising prevalence of dementia poses a significant global health challenge, particularly as populations age. Given the increasing burden of dementia for individuals, families and healthcare systems, there is an urgent need for innovative research to improve prevention, diagnosis, and treatment. Understanding the risk factors and underlying mechanisms of dementia is critical for developing effective strategies to combat this debilitating disease. Early detection is crucial for managing symptoms and potentially slowing progression. By understanding how medical factors influence dementia risk, doctors may be able to identify individuals who could benefit from early intervention and lifestyle changes to help prevent or delay the onset of dementia. Previous research has explored how factors such as high blood pressure, diabetes, and even personality traits can be linked to an increased risk of dementia [3]. However, these studies often rely on self-reported data or focus on a single factor, making it difficult to untangle the complex causes of dementia. The research took a more comprehensive approach by using a dementia patient characteristics dataset with several medical factors. The approach built machine learning models to predict the relationship between patient health parameters such as heart rate, presence of diabetes, and weight and the likelihood of dementia. Logistic regression and random forest algorithms were applied, and the random forest model correctly predicted dementia presence with 52% accuracy, whereas the logistic regression model predicted dementia with 48% accuracy. Future work will tune the models to improve the results and determine which factors were most influential in predicting dementia.

2. Introduction

Dementia, a debilitating decline in cognitive abilities, affects more than 55 million people globally. Early detection is crucial for timely intervention and potentially slowing disease progression. This research explores the potential of utilizing a patient's health information to identify individuals at increased risk of developing dementia before the onset of symptoms. Traditionally, dementia diagnosis relies on cognitive assessments and, sometimes, brain scans. However, these methods often occur after symptoms become evident. Early detection holds immense value. It allows for earlier intervention where mental training and medication management can potentially slow cognitive decline if implemented early. It can also help improve patient care. Early diagnosis enables proactive care planning, empowering patients and families to make informed decisions regarding future care needs. Furthermore, early identification of at-risk individuals facilitates targeted research efforts, leads to improved treatment strategies and potential breakthroughs. Dementia refers to a group of brain diseases that cause memory and thinking to slowly worsen over time impacting memory, thinking, and reasoning. While age is a major risk factor, several other conditions contribute to an increased likelihood of developing dementia. This research focuses on identifying such factors present in a patient's health information that might serve as early indicators of future dementia risk. The type of problem this research delves into is supervised learning and classification. The data that I am working with is numerical data that leverages numerical health data to predict the likelihood of future dementia. By analyzing the data, the model aims to learn patterns and relationships between the various factors and the development of dementia. The research output will be predicting the probability of an individual developing dementia based on their information. This model can serve as a valuable tool for healthcare professionals, aiding in the early identification of at-risk individuals and paving the way for earlier intervention and improved patient outcomes.

3. Background

Research has explored various avenues to address dementia risk. One approach focuses on epidemiology, as researched by van der Flier and Scheltens [1]. This field analyzes disease patterns in populations, offering valuable insights into the prevalence and risk factors associated with dementia. This approach provides a broad understanding of the disease and its societal impact. However, it doesn't predict individual risk.

Another approach investigates specific risk factors associated with dementia, such as those explored in studies examining sociodemographic and diabetes-related factors [2]. This approach identifies specific characteristics, such as age, ethnicity, and diabetes, that correlate with an increased risk of dementia. While valuable, such studies often rely on self-reported data or lack detailed clinical information, potentially missing key factors. Additionally, they may not be readily applicable for individual risk prediction.

Machine Learning (ML) offers a promising approach to enhance dementia risk prediction. Unlike traditional methods, ML algorithms can analyze large datasets of patient information, including medical history, to identify complex patterns and relationships. This allows for the development of predictive models that can estimate an individual's risk of developing dementia based on their specific characteristics. Similar to the studies by van der Flier and Scheltens [1] and [2], it investigates potential, yet under-explored, factors that could contribute to dementia risk prediction. However, this research uses machine learning to develop a model that aims to predict individual risk.

1 = Epidemiology and risk factors of dementiaW M van der Flier,P Scheltens

2 = Risk factors for dementia

4. Dataset

This research leverages a publicly available dataset from Kaggle, titled "Dementia Patient Characteristics Dataset." The dataset includes various numerical health factors potentially associated with dementia risk. These factors include diabetes, blood alcohol level, heart rate, blood oxygen level, body temperature, weight, and MRI scan delay. The variable diabetic indicates if the patient is diabetic. Alcohol level shows the level of alcohol consumption in patients. Blood oxygen level is the oxygen saturation level in the blood, and the MRI delay represents the delay time in an MRI scan, which could suggest a patient's growing condition. Additionally, my data was numerical, meaning my data was labeled by numbers, with 1 being "yes" and 0 being "no", with 1000 samples that were split. We split the data to be 20% testing and 80% training. Fortunately, the data exhibited minimal preprocessing requirements. Initial inspection revealed it to be clean and free of missing values or outliers, eliminating the need for data reduction or transformation techniques. This well-structured dataset provided a valuable foundation for investigating the relationships between various health factors and the risk of developing dementia.

5. Methodology/Models

We compared each factor against dementia status, then used a multitude of visuals to help determine if there were any factors that stood out. We initially created histograms for each health factor, grouped by dementia status (dementia vs. no dementia). While these provided a basic understanding of the data distribution, they didn't reveal any particularly strong individual factor

associations with dementia. Figure 1 is an example of a histogram of alcohol levels and dementia:

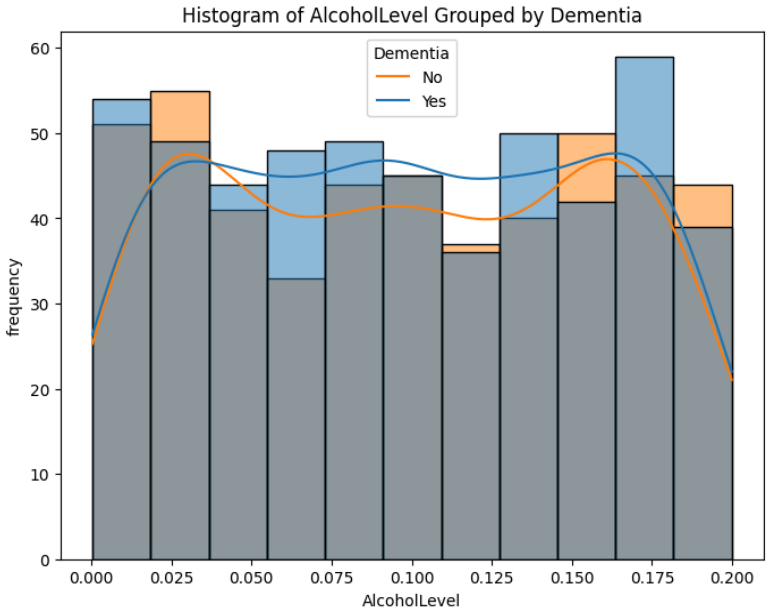

Figure 1. Histogram of alchohol level vs dementia with the orange being no dementia and the blue being yes

We then proceeded to generate boxplots for each health factor vs dementia status. Figure 2 shows a slight elevation in blood oxygen levels among individuals without dementia compared to those with dementia.

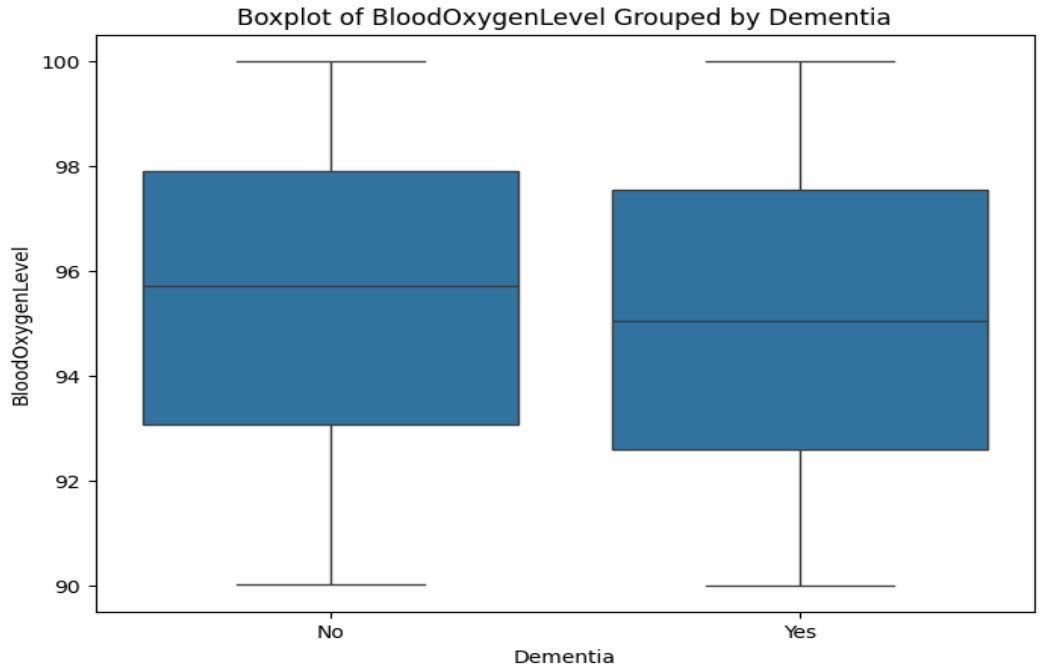

Figure 2. Boxplot of the blood oxygen level of people with and without dementia

Another visual we used was a pairplot which is a collection of scatterplots arranged in a grid, visualizing the relationship between each pair of variables in the data. With this, we noticed

that one of the visuals (Figure 3) showed that people with diabetes are more likely to have dementia. In Figure 3, the blue is 0 and the orange is 1, the blue is visually greater.

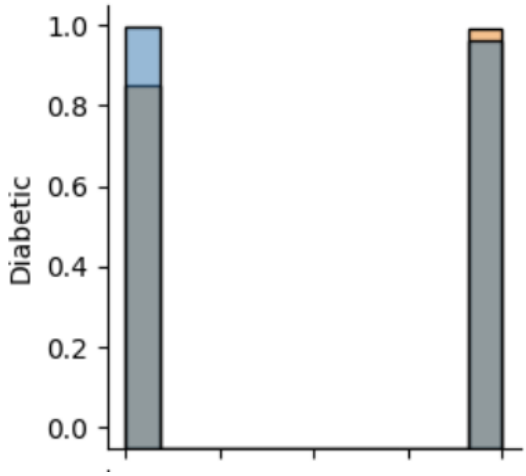

Figure 3. Pairplot comparing dementia patients with and without diabetes. The blue is 0 (no dementia) and the orange is 1 (with dementia)

While not all visualizations yielded significant insights, the exploratory data analysis did identify potential relationships between blood oxygen levels, diabetes, and dementia risk. These findings warrant further investigation using more sophisticated statistical techniques outside this research. We used these factors to run various models, the best being the RandomForest Model.

6. Results and Discussion

Our initial approach to predicting dementia risk employed a logistic regression model. To control training time, we set a maximum of 1000 iterations. The logistic regression model achieved an accuracy of 48% on the original test set.

In an effort to improve the model's accuracy, we subsequently implemented a random forest classifier. RandomForest Models are known for their ability to handle complex relationships between features and the target variable. To optimize the hyperparameters of the RandomForest Model, we utilized GridSearchCV. GridSearchCV performs an exhaustive search over a defined set of hyperparameter values, training the model with each combination and selecting the one that yields the best performance. Following hyperparameter tuning, the random forest classifier was trained on the training data and subsequently used to make predictions on the unseen test set. RandomForestClassifier then trained our training data and then made predictions. Figure 4 shows our accuracy, f1 score, recall, and precision. The best accuracy achieved was 52%.

	precision	recall	f1-score
0	0.49	0.62	0.55
1	0.56	0.43	0.49
accuracy			0.52

Figure 4. An image of the accuracy, f1 score, recall and precision of our AI using a RandomForest model

7. Conclusion

While the RandomForest Model achieved a moderate accuracy of 52% in predicting dementia risk, this research lays a valuable foundation for future exploration. Despite the limitations, the model's ability to identify potential relationships between health factors and dementia highlights the promise of machine learning in this domain. Future work can focus on several key areas to improve the model's performance. Firstly, incorporating additional, potentially relevant health data points could enhance the model's ability to capture the complex interplay of factors contributing to dementia risk. Secondly, exploring feature engineering techniques might extract more informative features from the existing data. This research demonstrates the potential of machine learning to contribute to early dementia risk identification. By continuously refining the model and incorporating new insights, we could capture the signs of dementia early in an accessible and affordable manner, enabling proactive care of patients. This approach has the potential to empower healthcare professionals in their fight against dementia.

8. Acknowledgements

I would like to express my sincere gratitude to InspiritAI for providing me with this invaluable learning experience through this and their summer program. This research would not have been possible without the exceptional guidance and support of my mentor, John Basbagill. His knowledge and assistance throughout the project were instrumental in its development. Additionally, I am deeply grateful for the unwavering support of my mother. Finally, I would like to acknowledge my grandmother, whose experience with dementia initially sparked my interest in this research topic. Her story serves as a reminder of the importance of early detection and technology in healthcare.

9. References

Gilbert Milton 20. “Dementia Patient Characteristics Dataset.” *Kaggle.com*, 2023, www.kaggle.com/datasets/gilbertmilton20/dementia-patient-characteristics-dataset. Accessed 27 June 2024.

McCullagh, Catriona D., et al. “Risk Factors for Dementia.” *Advances in Psychiatric Treatment*, vol. 7, no. 1, Jan. 2001, pp. 24–31, https://doi.org/10.1192/apt.7.1.24.

van der Flier, W M, and P Scheltens. “Epidemiology and Risk Factors of Dementia.” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl_5, 1 Dec. 2005, pp. v2–v7, jnnp.bmj.com/content/jnnp/76/suppl_5/v2.full.pdf, https://doi.org/10.1136/jnnp.2005.082867.

Centers for Disease Control and Prevention. “What Is Dementia?” *Centers for Disease Control and Prevention*, 5 Apr. 2019, www.cdc.gov/aging/dementia/index.html.