Computers in Biology and Medicine 171 (2024) 108126

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Prediction of dementia based on older adults' sleep disturbances using machine learning

Joel Nyholm [a], Ahmad Nauman Ghazi [b,*], Sarah Nauman Ghazi [c], Johan Sanmartin Berglund [c]

[a] Department of Computer Science, Blekinge Institute of Technology, Karlskrona, 37179, Blekinge, Sweden
[b] Department of Software Engineering, Blekinge Institute of Technology, Karlskrona, 37179, Blekinge, Sweden
[c] Department of Health, Blekinge Institute of Technology, Karlskrona, 37179, Blekinge, Sweden

ARTICLE INFO

Keywords:
Dementia
Sleep
Risk factors
Machine learning

ABSTRACT

Background: The most common degenerative condition in older adults is dementia, which can be predicted using a number of indicators and whose progression can be slowed down. One of the indicators of an increased risk of dementia is sleep disturbances. This study aims to examine if machine learning can predict dementia and which sleep disturbance factors impact dementia.

Methods: This study uses five machine learning algorithms (gradient boosting, logistic regression, gaussian naive Bayes, random forest and support vector machine) and data on the older population (60+) in Sweden from the Swedish National Study on Ageing and Care — Blekinge (n = 4175). Each algorithm uses 10-fold stratified cross-validation to obtain the results, which consist of the Brier score for checking accuracy and the feature importance for examining the factors which impact dementia. The algorithms use 16 features which are on personal and sleep disturbance factors.

Results: Logistic regression found an association between dementia and sleep disturbances. However, it is slight for the features in the study. Gradient boosting was the most accurate algorithm with 92.9% accuracy, 0.926 f1-score, 0.974 ROC AUC and 0.056 Brier score. The significant factors were different in each machine learning algorithm. If the person sleeps more than two hours during the day, their sex, education level, age, waking up during the night and if the person snores are the variables that most consistently have the highest feature importance in all algorithms.

Conclusion: There is an association between sleep disturbances and dementia, which machine learning algorithms can predict. Furthermore, the risk factors for dementia are different across the algorithms, but sleep disturbances can predict dementia.

1. Introduction & background

Sleep is a state of reduced mental and physical activity and is a vital part of everyday activities. When sleep is insufficient, several issues appear, such as impaired learning and increased risk of stress-related diseases such as mood disorders and cardiovascular diseases [1–3].

The older population, adults 60+, have additional problems with sleep and sleep disturbances. They are more prone to sleep disturbances, such as insomnia [4]. These problems combined can lead to and increase the risks for other diseases. A common one among older adults is dementia.

Sleep deficit and bad sleeping habits are risk factors for dementia [5–8]. However, sleep is not the only risk factor for dementia. Education, age and one's sex are other factors that contribute to the risk level of a given individual [9–11].

The National Board of Health and Welfare in Sweden reports that around 130,000 to 150,000 individuals suffer from dementia, and this number is expected to double by 2050 [12]. Comparably, other high-income regions, such as Western Europe and North America, have a higher prevalence rate of dementia-related disease (see Fig. 1). Furthermore, this paper only utilizes data on the older population that resides in Blekinge, Sweden.

As data is available on the older population, we can make inferences from several aspects of their lives. Machine learning (ML) is one method that utilizes the data to find patterns. ML is a program that uses experience to learn and predict based on its experiences [13]. These programs can find patterns in the data and thus explain or find associations.

Several studies in the reviewed literature (see Section 2) have examined sleep and dementia. Several of these use ML to study the

* Corresponding author.
E-mail address: nauman.ghazi@bth.se (A.N. Ghazi).

https://doi.org/10.1016/j.compbiomed.2024.108126
Received 1 November 2023; Received in revised form 14 December 2023; Accepted 6 February 2024
Available online 9 February 2024

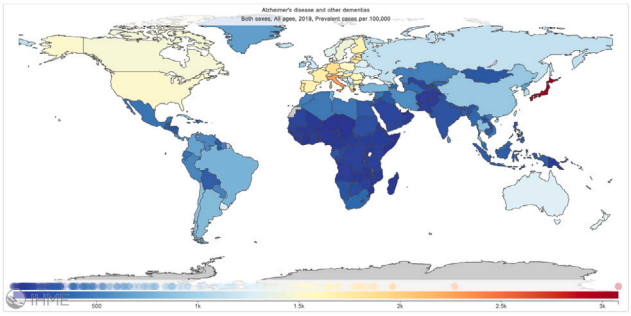

Fig. 1. Worldwide prevalence of dementia [14] The colour grade as indicated in the legend, varies from blue (low prevalence) to red (high prevalence).

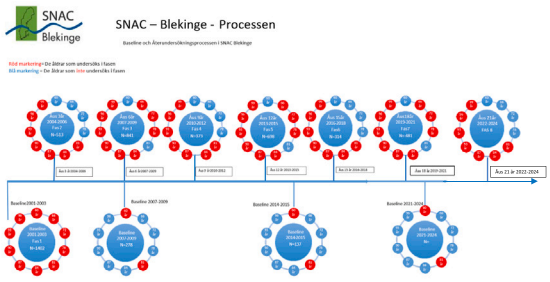

Fig. 2. SNAC-B study's process, [29].

association between sleep and dementia. However, they use other features, methodological approaches and other sleep aspects, such as REM sleep. Therefore using machine learning with different features on other sleep aspects is beneficial.

Because of the risks, the high prevalence rate, data availability and the research gap, investigating ML prediction of dementia based on sleep is worthwhile. Can machine learning predict dementia and examine which sleep disturbance factors indicate dementia? These are the main issues at hand.

1.1. Dementia

Dementia is a term that is described as a group of symptoms, such as cognitive disabilities and memory loss, that occur due to some structural changes in the brain [15,16]. Dementia is prevalent in the following diseases, namely (sorted in order of prevalence), Alzheimer's disease (50%–75%), vascular dementia (20%), dementia with Lewy bodies (5%) and frontotemporal dementia (5%) [17,18]. A few prevalent symptoms of dementia are memory loss (often short-term memory), confusion, repetition when speaking, personality change and poor judgment [16,19,20].

Dementia occurs because of changes in some brain regions that lead to neurons and their subsequent connections stopping working. Although the exact reason for these changes remains unknown, several factors have shown indications to decrease the risk and prevalence of dementia, such as leading a healthy lifestyle [19].

Studies have shown that lifestyle factors have decreased the risk and prevalence of dementia. However, a review of the current research concludes that there is inconclusive evidence of these lifestyles [21]. The silver lining is that some results are "*encouraging but inconclusive*" and could show a causational relationship with dementia. There are many recommendations for further research, such as social engagement interventions, depression treatment, dietary interventions and sleep quality interventions [21].

At present, dementia remains incurable, but some treatments can help alleviate its symptoms. One such symptom is sleep disturbances, which can be managed through either pharmacological or non-pharmacological means [22]. As Dodson and Zee [23] mention, starting with non-pharmacological approaches, such as establishing good sleep hygiene, avoiding bright light in the evening, and increasing light exposure in the morning, is a reasonable approach. If these methods prove ineffective, only then pharmacological interventions, like melatonin, may be appropriate.

1.2. Sleep

Sleep is crucial for everyone, without it we experience mental fatigue, impaired learning, and increased risk of stress-related diseases such as mood disorders and cardiovascular diseases [1–3]. The number of hours a given person varies, but adults, both young and old, generally need seven or more hours per day [24,25]. This study categorizes sleeping into three categories, sleep disturbances, sleep quality, and sleep behaviours.

First of all, sleep disturbances include several issues concerning sleep, namely, initiation and maintenance of sleep, excessive sleep, dysfunctional sleep-wake schedule, and other issues that relate to sleep [26]. In summary, this category encompasses all issues associated with an increase in sleep disturbances.

Sleep quality means how adequate the sleep was. It is a metric that involves all factors of the sleep experience. The quality depends on the individual's experience and satisfaction, sleep latency and efficiency are two aspects that can measure sleep quality [27].

Lastly, behaviours are how one has acted during the day in factors that influence sleep. For example, has the individual exercised or begun to sleep at a suitable time? This study's primary focus is on features of sleep disturbances.

1.3. Dataset

1.3.1. SNAC-B

A primary part of this thesis is the data. Reliable and trustworthy data from patients is needed to conclude anything of importance. Therefore, the data comes from the Swedish National Study on Aging and Care in Blekinge (SNAC-B) [28].

SNAC is a long-term longitudinal study that started in 2001 and is still ongoing (2022). The study is national and happens in four areas of Sweden: Blekinge, Kungsholmen, Nordanstig, and Skåne.

The participants in the study are aged 60 to 99 (see Fig. 2). The study consists of two parts, the baseline and the re-examinations. The baseline is where new participants enter the study. This process happens every six years and introduces older adults aged 60 and 81, except for the first baseline, where new participants are aged 60–99. The re-examination occurs every three years. However, only people aged 81 and older do the re-examinations every third year. The younger adults (<81) do the re-examination every sixth year until they become 81 years old.

The study bases its results on samples from older adults. Collecting the data happens via questionnaires, surveys, interviews, and clinical examinations, where the aim is to sample information on social conditions, health, diseases, and functional capacity.

This study primarily focuses on the data of participants' sleep disturbances. It is important to note that the data used in this paper is a compounded version of SNAC-B's longitudinal data, which does not consider the original longitudinal approach.

1.3.2. Measurements

Classifying the participant's cognition is vital. Therefore, we use the Mini-Mental State Examination combined with the Clock-test score. Both tests are robust when screening for dementia [30,31] and are significantly improved when using both in combination [32].

Table 1
The sleep and personal-related variables utilized in the experiment and analysis, from the SNAC-B dataset.

ID	Description	Type	Values
Personal features			
Sex	The participants' sex	Nominal	Male/Female
AGE	The participants' age	Discrete	60–99
A1	The participants' civil status	Nominal	1–4
C51_Education	The participants' highest completed education	Ordinal	1–8
Dementia	If a given participant is in the dementia class	Nominal	Yes/no
Sleep features			
D31	Trouble falling asleep	Nominal	Yes/no
D32	Taking/addiction to sleeping medicine	Nominal	Yes/no
D33	Waking up during the night	Nominal	Yes/no
D34	Difficulty falling asleep because of stress/mood	Nominal	Yes/no
D35	Difficult sleeping because of pain/itching	Nominal	Yes/no
D36	Trouble falling asleep after waking during the night	Nominal	Yes/no
D37	Waking up early	Nominal	Yes/no
D38	Feeling tired and sleeping more than 2 h during daytime	Nominal	Yes/no
D39	Frequency of taking sleep medicine	Ordinal	1–5
D40	Hours of sleep per night	Continuous	0–24
E36	Snoring	Nominal	Yes/no
E135	Sleep quality	Ordinal	0–6

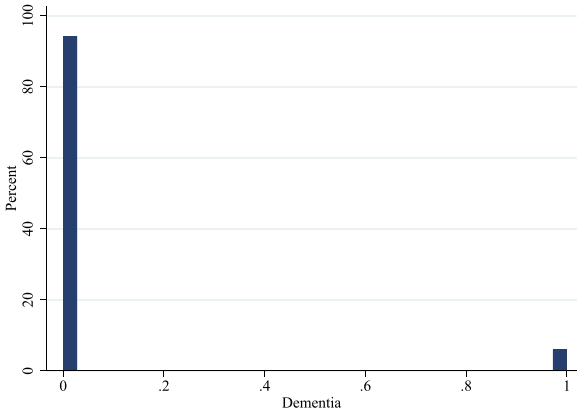

Fig. 3. Histogram, dementia class distribution.

The threshold for being classified as having dementia is to have an MMSE score lower than 24 out of 30 [33] and a Clock-test score of less than 8 of the maximum score of 10 [34].

Of the 4175 samples in the dataset, 247 participants met the requirements to be in the dementia class. However, this creates an imbalanced dataset which needs resolving (see Section 1.3.3). Furthermore, the experiment uses 16 features for measuring sleep disturbances, general sleep quality, and personal parameters (see Table 1).

1.3.3. Data characteristics

The original dataset has 3208 participants. Each participant can have one or several samples. Therefore there are 5033 samples to use before dataset cleaning.

The dataset cleaning consists of removing invalid data samples. The invalid ones do not have an MMSE and Clock-test score. Without these test scores, the study cannot label each participant's cognition. There was a total of 858 invalid samples.

After the dataset cleaning, there were 1821 participants from the SNAC study and 4175 samples (see Table 2).

However, two issues appear with the features: missing values and an imbalanced dataset. KNN imputation solves the issues with missing values, where it fills the data cells with generated values. The other problem is the data imbalance between the two groups, the healthy group and the dementia group (see Fig. 3). ADASYN solves this issue by using oversampling, which creates new samples based on the current data. This process will even out the class distributions.

Table 2
SNAC-B participants.

Age	Participants	Percent
60	444	10.63
66	577	13.82
72	603	14.44
78	468	11.21
79	2	0.05
80	44	1.05
81	488	11.69
83	20	0.48
84	524	12.55
86	26	0.62
87	416	9.96
89	2	0.05
90	294	7.04
92	11	0.26
93	151	3.62
95	6	0.14
96	74	1.77
98	3	0.07
99	19	0.46
102	3	0.07

1.4. Data pre-processing

Machine learning needs data, which is where it learns from. The data quality affects the final results [35]. Because of this, pre-processing of the data is vital. With it, we can mitigate several issues, and it may drastically change the final prediction positively or negatively [36].

1.4.1. K-nearest neighbour imputation

K-nearest neighbour (KNN) imputation is an algorithm for handling missing data in a dataset. Imputation is a technique to handle missing data by assigning values to the missing fields. Several studies [37–39] have confirmed that KNN imputation is a robust and accurate imputation algorithm. The algorithm functions as follows [37]:

1. Determine parameter K, ($K = 5$, in this study)
2. Calculate the Euclidean distance, for each missing data instance

$$d = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

where:
n = Data's dimensions
x = A point
y = A point

3. Calculate the value of the missing cell, which is based on the mean of the nearest neighbours

$$z = \frac{1}{K}\sum_{i=1}^{K} x_i \tag{2}$$

where:
z = Missing data cell
m_i = K-Nearest neighbour's value

1.4.2. ADASYN

The adaptive synthetic (ADASYN) is an algorithm which handles imbalanced datasets. It solves this issue by adaptively generating data based on the distribution of the original dataset [40]. Several studies [41–43] have used ADASYN in medical and other ML studies with good results. The algorithm functions as follows:

1. Calculate the class imbalance

$$d = \frac{m_s}{m_i},\ d \in [0,1] \tag{3}$$

where:
m_s = Minority class examples
m_i = Majority value examples

2. If $d < d_{th}$, then d_{th} is the threshold for the tolerated ratio of class imbalance

   (a) Calculate how many synthetic data points that are to be generated, for the minority class

$$G = (m_i - m_s) \times \beta \tag{4}$$

   where:
   β = Is the desired balance level after generation, $\beta = 1$ means a fully balanced dataset
   m_i = Majority value examples

   (b) For each point in the minority class find the K-nearest neighbours and calculate the ratio

$$r_i = \frac{\Delta_i}{K}, i = [1, \ldots, m_s], r_i \in [0,1] \tag{5}$$

   where:
   Δ_i = KNN for each example that belongs to the majority class

   (c) Normalize r_i, so $\hat{r_i}$ is a density distribution of $(\sum_i \hat{r_i} = 1)$

$$\hat{r_i} = \frac{r_i}{\sum_{i=1}^{m_s} r_i} \tag{6}$$

   (d) Calculate how many data points are to be generated for each minority example x_i

$$g_i = \hat{r_i} \times G \tag{7}$$

   where:
   G = Minority's class total of data which will be generated (see Eq. (4))

   (e) For each minority class example x_i, generate g_i data examples
   Loop from 1 to g_i:

      i. Randomly choose an example x_{zi} from the KNN for example x_i
      ii. Generate the data

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \tag{8}$$

   where:
   $G = \lambda$ is a random number, $\lambda \in [0,1]$

1.4.3. Statistical power

Ensuring credible results require adequate statistical power. The effect size is one part of calculating the power. It is the relational strength between two variables in a given population. As the effect size is unknown, it is preferred to choose a minimal value that does not overestimate the relationship [44], in this case, between sleep disturbances and dementia. Thus it is given a value of $\gamma = 0.05$.

With these assumptions, the data groups minimally need 512 data points per group (see Eq. (9)), with 1024 data points in the experiment. The dataset has 4175 data points, so the statistical power is adequate ($\delta \geq 0.8$).

$$N = 2\left(\frac{\delta}{\gamma}\right)^2 = 2\left(\frac{0.8}{0.05}\right)^2 = 512 \tag{9}$$

where:
N = Number of subjects
δ = Statistical power
γ = The effect size

2. Related works

The use of technology to enhance different fields is commonplace, and medicine is no exception. Several works for predicting or diagnosing dementia with machine learning already exist. Most studies focus on ML that uses binary or multivariate classification to decide if a patient has a high predicted risk of dementia. Another commonality is the frequent use of supervised ML algorithms. The most common algorithms in the reviewed literature (in sorted order) are support vector machine (SVM) with thirteen usages, random forest (RF) with eight usages, logistic regression (LR) with seven usages, Gaussian naive Bayes (GNB) with three usages and extreme Gradient Boosting with two usages.

Dementia is a complex disorder that can be affected by how we live during our entire lifetime, and many factors contribute to a high risk of developing dementia [11]. Before anyone develops dementia, they have mild cognitive impairment (MCI), which can lead to further cognitive decline. Many risks exist for this, the topmost being age, depression, anxiety, mental health, sleep disturbances, and exercise [45,46].

For dementia, the risk factors are similar. Age, exercise, education, hypertension, body mass index, diabetes, depression and the participant's sex are some risk factors for dementia [9,47]. Furthermore, research into dementia in men and women shows that some risk elements are different between the sexes. Financial problems, regular physical activity, and Schuster-social support score are risk elements for men, whilst depression, hypertension, and alcohol abstention are risk elements for women [48]. Two additional factors impact both sexes, moderate physical activity and cognitive engagement [48].

The risk factors can also divide themself into three categories, based on one's age, early life (<45 years of age), midlife (45–65 years of age) and later life (>65 years of age) [11]. Early life has one risk factor, less education. Midlife has five risks, hearing loss, traumatic brain injury, hypertension, >21 alcohol units consumed per week and obesity. Lastly, later life has two factors which are smoking and depression. These risks and those in the previous paragraph conclude that dementia is affected by our accumulated lifestyles, which also means one can lower the risk of developing dementia.

There is an abundance of risk factors, which are composed of several other elements. Therefore, it exists many methods of predicting dementia. One common approach is utilizing metrics from the brain such as Electroencephalogram (EEG), resonance imaging (MRI) and Positron emission tomography (PET). EEG is a testing method where several electrodes are placed on the patient's scalp to measure brain activity. This test, combined with ML, has seen encouraging results.

Several studies [49,50] have been able to classify, with various algorithms, dementia in patients. One study has measured the EEG whilst the patients are asleep to classify MCI with promising results [7].

Another method is medical imaging using either MRI or PET. These methods have also been effective at classifying dementia and differentiating between different forms of dementia [51–53]. Further, combining MRI and EEG to train an ML model has promising results. The research indicates that combining both metrics leads to increased accuracy versus using either of the metrics by itself [54].

A final method that relies on biomarkers is classification through genes. They also impact dementia, and identifying dementia in genes in combination with ML is another promising field of dementia classification. Several studies [55–57] have conducted experiments based on this. Gene classification could also be a cheaper alternative to the more expensive medical imaging techniques [57].

Identifying dementia is not only possible through various biomarkers. It also shows itself through a change in behaviour, worse memory and several more symptoms. Therefore, studies have examined speech and eye-tracking and if they can identify dementia.

In the reviewed literature, speech has the most research of the previously described methods. Several studies [58–61] have classified MCI and dementia through the patients' speech patterns. Further, this method is an accessible and low-cost alternative, which can happen without the patients being physically present.

Eye-tracking does not have as much research, but it has positively classified MCI. The eye-tracking happened as the subjects were reading to capture the data, and the ML model finally had an accuracy of a maximum of 86% [62].

As discussed above, identifying dementia can happen using several methods. However, this study focuses on dementia and sleep disturbances. Sleeping disturbances and deprivation are both factors which can promote brain damage, which can lead to dementia [10]. Further, dementia patients have less rapid eye movement (REM) sleep than others, which leads to a lack of this type of sleep [10].

Several studies have examined sleep and its association with dementia [5–7,63–66]. Using REM sleep as a feature to examine dementia has had a good result and showed that it was a high accuracy factor when predicting dementia [6].

Various forms of sleep disturbances is another factor related to dementia. Sleep fragmentation [64], sleep-disordered breathing [63], and sleep disturbances [65,66] have an association with dementia or cognitive decline. A study measured the participants using a wristband with sensors and an unsupervised ML algorithm [5]. The study could classify severe dementia patients with an accuracy of 91% and mild dementia subjects with 87% accuracy. This approach presents an affordable and non-invasive method of additional diagnosis help for dementia.

These studies show that dementia presents itself in various forms, and sleeping is one of them. Sleep and dementia have an association with each other, and continuing research into this is the focus of this study.

3. Research methodology

3.1. Research questions

The thesis has four RQs, four primary and two sub-questions. The sub-questions will answer the main RQ and use the experiment results as its data and motivation for its answers.

Research question 1: What is the association between sleep and dementia in older adults?

Research question 2: To what extent can machine learning predict dementia in older adults based on their sleep disturbances?

Research question 3: What are the sleep disturbance factors that lead to an increased risk of cognitive decline in older adults?

> ***Research question 3a***: To what extent do the algorithms' feature importances conclude the same result?
>
> ***Research question 3b***: What are the primary sleep disturbance features that increase the risk of conceiving dementia?

Research question 4: Which machine learning algorithm has the best mean prediction accuracy when predicting dementia in older adults based on their sleep disturbances?

3.1.1. Research question 1

RQ1*: What is the association between dementia and sleep based on older adults?*

Investigation of this question is vital to know if there is an association between sleep and dementia. Multiple studies (see Section 2) have reported an association between the two factors. Nonetheless, this RQ aims to give credibility or contradict the association based on the SNAC-B data.

Other studies have not used the same combination of sleep features. If an association exist between SNAC-B's sleep features and dementia, it will also give the other RQs more credibility and worth.

3.1.2. Research question 2

RQ2*: To what extent can machine learning predict dementia in older adults based on their sleep disturbances?*

The question's primary concern is the accuracy of the ML prediction. If the results are inaccurate, concluding any reliable results from the models will be difficult. Therefore, this question is necessary for the study. Furthermore, we can also find the most accurate model in the study.

3.1.3. Research question 3

RQ3*: What are the topmost sleep disturbance factors that lead to an increased risk of dementia in older adults?*

To further give insight into if or how sleep influences dementia, we need to know which features have a higher risk. Feature importance is a term that explains this for ML. It demonstrates which features have the highest impact on the model.

Because of the reasons above, it is worthwhile to answer this RQ. The answer can indicate or give insight into the association between sleep disturbances and dementia.

3.1.4. Research question 3a

RQ3a*: To what extent do the algorithms' feature importances conclude the same result?*

The ML algorithms might conclude different results in the feature importances. The algorithms build their models differently, which could result in differences in the final results. If the algorithms have the same conclusion, we can easily determine which factors have the highest impact. However, if they do not, it is worthwhile to see which algorithms do and do not have the same results.

3.1.5. Research question 3b

RQ3b*: What are the primary sleep disturbance features that increase the risk of conceiving dementia?*

This question needs the result from RQ3a (see Section 3.1.4). If that question concludes that the algorithms have the same conclusion in all feature importances, then this question is simple to answer. However, if they do not, the answers become inconclusive. Concluding which algorithms have the correct feature importance and which do not is futile. However, indicating which factors regularly have a high impact is possible.

3.1.6. Research question 4

RQ4*: Which machine learning algorithm has the best mean prediction accuracy when predicting dementia in older adults based on their sleep disturbances?*

All algorithms have differences in how they achieve their task. These differences make them adept in certain areas whilst not in others. Do these differences make a vast difference in the results, and if they do, which is the best-suited algorithm for this task? The result can indicate which ML algorithm is the best suited for prediction in this area.

3.2. Variables

The independent variables are the patients' data, whilst the dependent variable is the mean Brier score (see Eq. (10), [67]) of the ML model for RQ2 (see Section 3.1.2) and the mean permutation feature importance per feature (see Eq. (11), [68]) for RQ3a and RQ3b (see Sections 3.1.4 & 3.1.5).

The Brier score is a metric that assesses the accuracy of predictions. It always gives a value between zero and one, where zero is a perfect prediction. To obtain the final Brier score, we utilize 10-fold stratified cross-validation and calculate the mean of the ten results. The cut-off value in the hypotheses is 0.25, which gives the score when the prediction has a 50% chance of right or wrong ($(0.5-1)^2 = (0.5-0)^2 = 0.25$). There are some issues with this metric. If the dataset is small (< a few hundred) [69] and if the probabilities are very low or high [70,71]. These issues resolve themself because of the sufficiently large dataset ($n = 4175$) and the distribution of all samples. The samples from a population will have normality because of the central limit theorem [72]. Therefore, the low and high probabilities should be small and not have a high impact on the result.

The mean permutation feature importance per feature gives a set of values with the size of the number of features. The result is a set where each entry is the mean of a given feature's feature importance per cross-validation fold. Because this thesis uses 10-fold stratified cross-validation, there will be ten values per feature, which, after calculation, become the mean value at a given position in the set. All features in the set undergo this process to obtain a final set of values.

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2 \tag{10}$$

where:
N = Number of observations
f = Predicted probability of classification
o = Outcome of the event (0 if happened, 1 otherwise)

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \tag{11}$$

where:
i_j = Importance
K = Repetitions
$s = R^2$ regression score
j = A given feature of the dataset

3.3. Performance metrics

Using other performance metrics shows how the algorithms perform in a broader spectrum. Therefore, the thesis utilizes accuracy (see Eq. (12)), f1-score (see Eq. (15)) and ROC AUC for the descriptive statistics (see Section 4.1). Additionally, several studies in the reviewed literature (see Section 2) use the metrics described above. Further, including these metrics facilitates easier comparisons of this thesis with future studies.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

$$F_1 = 2\frac{precision * recall}{precision + recall} \tag{15}$$

where:
TP = True positive
TN = True negative
FP = False positive
FN = False negative

Table 3
Sklearn parameters.

Algorithm	Class name	None-default parameters
GB [73]	GradientBoostingClassifier	random_state = 0
GNB [74]	GaussianNB	
LR [75]	LogisticRegression	solver = "liblinear" random_state=0
RF [76]	RandomForestClassifier	random_state = 0
SVM [77]	svm.SVC	kernel = "linear" probability = True max_iter = 10000000

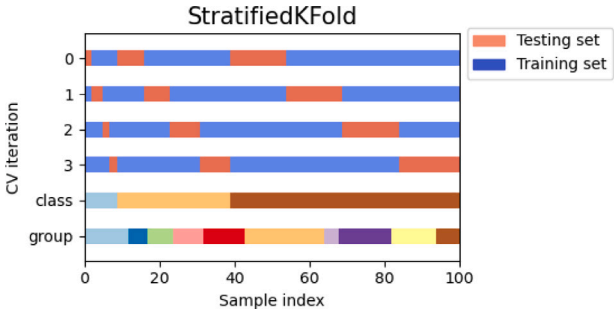

Fig. 4. K-fold stratified cross-validation, [78].

3.4. Experimental instrumentation

The required materials for this experiment are a computer for execution and the ML code (link).

The ML code uses Python 3 as its programming language and scikit-learn as the ML library. RF, SVM, GB, LR and GNB are the chosen algorithms (see Table 3) because of their wide use in related studies with satisfactory results.

All algorithms use the "random_state" to have consistent randomization. LR and SVM have additional parameters to enable correct functionalities. LR's "liblinear" solver uses a one-vs-rest approach. Therefore, it trains different binary classifiers for each class, which is the primary reason behind using this solver.

The change in SVM's "kernel" and "probability" parameters is necessary to collect the probabilities that the Brier score needs. Furthermore, the "max_iter" is set to 10 million because of the converging time.

Further, the model will use 10-fold stratified cross-validation (see Fig. 4) to ensure a higher result validity, which ensures that each set has approximately the same percentage of each class as the original one [78].

3.5. Hypotheses

The thesis has four primary hypotheses (see Sections 3.5.1, 3.5.2, 3.5.8 and 3.5.9). However, RQ2 has five sub-hypotheses (see Sections 3.5.3–3.5.7) to conclude if each algorithm's prediction accuracy is adequate. Each hypothesis has a null- and alternative hypothesis.

3.5.1. Hypothesis 1

This hypothesis answers RQ1 (see Section 3.1.1) and uses multiple regression to investigate the association between sleep disturbances and dementia. After the regression, each factor will perform a t-test to calculate if its coefficient equals zero.

H_0 There is no association between dementia and sleep disturbances in adults aged 60 and higher.

H_1 There is an association between dementia and sleep disturbances in adults aged 60 and higher.

$H_0 : \forall f \in F, f = 0 \qquad H_1 : \forall f \in F, f \neq 0$

F = The set of all sleep disturbances
f = A given sleep disturbance

3.5.2. Hypothesis 2

RQ2 has five hypotheses (see Sections 3.5.3–3.5.7), one for each ML algorithm. Each hypothesis examines if the given algorithm has an equal, worse or better accuracy than a random classifier, where the Brier score is the metric employed to assess the accuracy of each algorithm.

3.5.3. Hypothesis 2a

H_0 Gradient boosting predicts as well as a random classifier.
H_1 Gradient boosting accurately predicts dementia based on older adults' sleep disturbances.
H_2 Gradient boosting does not accurately predict dementia based on older adults' sleep disturbances.

$H_0 : \mu_{GB} = 0.25 \; H_1 : \mu_{GB} < 0.25 \; H_2 : \mu_{GB} > 0.25$

where:
μ_{GB} = Mean Brier score of gradient boosting

3.5.4. Hypothesis 2b

H_0 Logistic regression predicts as well as a random classifier.
H_1 Logistic regression accurately predicts dementia based on older adults' sleep disturbances.
H_2 Logistic regression does not accurately predict dementia based on older adults' sleep disturbances.

$H_0 : \mu_{LR} = 0.25 \quad H_1 : \mu_{LR} < 0.25 \quad H_2 : \mu_{GB} > 0.25$

where:
μ_{LR} = Mean Brier score of logistic regression

3.5.5. Hypothesis 2c

H_0 Gaussian naive Bayes predicts as well as a random classifier.
H_1 Gaussian naive Bayes accurately predicts dementia based on older adults' sleep disturbances.
H_2 Gaussian naive Bayes does not accurately predict dementia based on older adults' sleep disturbances.

$H_0 : \mu_{GNB} = 0.25 \quad H_1 : \mu_{GNB} < 0.25 \quad H_2 : \mu_{GNB} > 0.25$

where:
μ_{GNB} = Mean Brier score of gaussian naive Bayes

3.5.6. Hypothesis 2d

H_0 Random forest predicts as well as a random classifier.
H_1 Random forest accurately predicts dementia based on older adults' sleep disturbances.
H_2 Random forest does not accurately predict dementia based on older adults' sleep disturbances.

$H_0 : \mu_{RF} = 0.25 \; H_1 : \mu_{RF} < 0.25 \; H_2 : \mu_{RF} > 0.25$

where:
μ_{RF} = Mean Brier score of random forest

3.5.7. Hypothesis 2e

H_0 Support vector machine predicts as well as a random classifier.
H_1 Support vector machine accurately predicts dementia based on older adults' sleep disturbances.
H_2 Support vector machine does not accurately predict dementia based on older adults' sleep disturbances.

$H_0 : \mu_{SVM} = 0.25 \quad H_1 : \mu_{SVM} < 0.25 \quad H_2 : \mu_{SVM} > 0.25$

where:
μ_{SVM} = Mean Brier score of support vector machine

3.5.8. Hypothesis 3

This hypothesis answers RQ3a (see Section 3.1.4). The test has two parts, an ANOVA test to see if the feature importance results of all algorithms are equal. If the results are equal, the tests end here. Otherwise, the study uses a Tukey test to see the differences between groups or which ML algorithms have the same results for a given feature.

H_0 The ML algorithms' feature importances have no difference between algorithms
H_1 The ML algorithms' feature importances have differences between algorithms

$H_0 : M_{RF} = M_{SVM} = M_{GB} = M_{GNB} = M_{LR}$

$H_1 : M_{RF} \neq M_{SVM} \neq M_{GB} \neq M_{GNB} \neq M_{LR}$

where:
M_{RF} = Median feature importance per feature for random forest
M_{SVM} = Median feature importance per feature for support vector machine
M_{GB} = Median feature importance per feature for gradient boosting
M_{GNB} = Median feature importance per feature for gaussian naive Bayes
M_{LR} = Median feature importance per feature for logistic regression

3.5.9. Hypothesis 4

This hypothesis answers RQ4 (see Section 3.1.6). This test uses either ANOVA or Kruskal–Wallis, depending on the normality of the data. The purpose is to see if the Brier scores are equal or not. If they are unequal, then one or more algorithms have a significant difference between their scores. Therefore the thesis uses the post-hoc Tukey test to detect and assess the pairwise comparisons between algorithms. Which obtains which algorithm or algorithms have statistically lower Brier scores and, thus, are more accurate than the remaining ones.

H_0 The ML algorithms' Brier score has no difference between algorithms
H_1 The ML algorithms' Brier score has differences between algorithms

$H_0 : \mu_{RF} = \mu_{SVM} = \mu_{GB} = \mu_{GNB} = \mu_{LR}$

$H_1 : \mu_{RF} \neq \mu_{SVM} \neq \mu_{GB} \neq \mu_{GNB} \neq \mu_{LR}$

where:
μ_{RF} = Mean Brier score for random forest
μ_{SVM} = Mean Brier score for support vector machine
μ_{GB} = Mean Brier score for gradient boosting
μ_{GNB} = Mean Brier score for gaussian naive Bayes
μ_{LR} = Mean Brier score for logistic regression

Table 4
Experiment design (hypothesis 2).

Cross-validation fold	Mean brier score for each algorithm
1	✓
2	✓
3	✓
4	✓
5	✓
6	✓
7	✓
8	✓
9	✓
10	✓

Table 5
Experiment design (hypothesis 3).

Cross-validation fold	RF	GB	LR	SVM	GNB
1	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓
3	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓
6	✓	✓	✓	✓	✓
7	✓	✓	✓	✓	✓
8	✓	✓	✓	✓	✓
9	✓	✓	✓	✓	✓
10	✓	✓	✓	✓	✓

3.6. Experiment design

The experiment consists of two sub-experiments which happen concurrently. The first concerns the algorithms' prediction accuracy, measured using the Brier score. This part uses one factor with two treatments (see Table 5). The Brier score of each algorithm is the dependent variable, whilst the patient dataset is the independent variable.

However, as there is only one dataset and using all data is beneficial to both the ML models' training but also for increased validity, we cannot divide the data between the algorithms. Therefore using a completely randomized design, crossover design, or similar designs is wasteful and not done. Thus, all algorithms will use the same data (see Table 4).

The second sub-experiment concerns the feature importance of the five algorithms. One factor with five treatments (see Table 5) is the design for this part. The dependent variable is the set of all mean feature importance for each algorithm, whilst the independent is the patient dataset. As with the previous experiment part, no division of the data happens.

The experiment happens as follows and is completed after these steps:

1. Start the Python program
2. Using 10-fold stratified cross-validation, all three algorithms:
   (a) Trains their model on the training set
   (b) Tests their model on the test set
   (c) Calculates all metrics
3. The program records the results to file

3.7. Analysis procedure

The statistical analysis uses Stata 17 as its primary program and has two steps. The first is descriptive statistics. Here the calculated metrics are presented in tables and figures to understand how the data behave and looks. A Shapiro–Wilks test for normality is taken to examine if the data has a normal distribution.

The second part is hypothesis testing. This thesis has four primary hypotheses, and examining them happens in this phase. RQ1 uses multiple logistic regression to investigate the association between dementia and sleep disturbances. The assumption of linearity of the independent variables, independent observations and multicollinearity is satisfied.

RQ2 uses several t-tests to see if the mean Brier score for each algorithm is equal to the hypothesized value.

RQ3a uses 16 Kruskal–Wallis tests to see if the feature importance medians are statistically equal.

RQ4 uses the ANOVA and the Tukey post-hoc tests to determine if the Brier scores are equal among the algorithms. However, if the null hypothesis is accepted, then the test stops, which concludes that there is not enough evidence to infer that a difference in accuracy exists. In the opposite case, the Tukey test presents the pairwise difference in accuracy across the algorithms.

Lastly, all tests use a significance level of 0.05 ($\alpha = 0.05$).

3.8. Validity threats

3.9. Threats to conclusion validity

The conclusion validity concerns the relationship between the treatment and its outcome, where we want to establish if there is a statistical relationship within the given significance level.

The primary issue with this threat is the reliability of the data. The data in the study is from SNAC-B, which provides data on various factors of older adults. However, one primary issue is missing data values in several features.

Another issue is the dataset imbalance of dementia participants and healthy participants. Both of these issues need resolving, which the KNN-imputation and ADASYN algorithms handle. Both algorithms use the original dataset as the basis for the data generation.

3.10. Threats to internal validity

The internal validity concerns that the relationship between treatment and outcome is casual and not a result of a factor which has not been controlled or is unmeasured. There are two issues in this category, selection and mortality.

The selection threat concerns the natural variation in humans, and the selection process from a population may influence the results. The data in the study comes from stratified sampling, where we partition the population into subcategories or strata and use random sampling within each.

Mortality concerns issues with participants leaving the study due to several reasons. The data from SNAC-B are of older adults. Thus, the participants have a high mortality rate. The issue is that if several participants in a strata leave the study, that stratum may then become underrepresented and could skew the results.

These issues mitigate themselves because SNAC-B uses random stratified sampling, which mitigates under-representative and non-representative populations.

3.11. Threats to construct validity

Construct validity addresses the relationship between theory and observation. If there is a casual relationship, then two things need confirmation. Firstly, the treatment reflects the cause, and secondly, the outcome reflects the effect. Divergent validity is the primary issue in this category.

Divergent validity concerns measuring a construct which relates to another one. This thesis uses personal, sleep behavioural and sleep quality factors. As the experiment uses all of these features simultaneously, the results can conclude a different cause and effect than the real one. However, this study also examines other sleep features which can impact or describe the reason for sleep disturbances. Therefore, there is a combination of the features in the dataset. Additional studies can examine which specific factors in each category affect the cause and effect.

Table 6
Performance metrics results (mean ± std. dev.) across algorithms.

Algorithm	Mean accuracy [%]	Mean exec. time [ms]	Mean f1-score	Mean Brier score	Mean ROC AUC
GB	0.929 ± 0.028	1189.971 ± 15.335	0.926 ± 0.031	0.056 ± 0.020	0.974 ± 0.018
LR	0.708 ± 0.085	38.818 ± 3.900	0.720 ± 0.074	0.194 ± 0.042	0.772 ± 0.107
GNB	0.747 ± 0.089	3.531 ± 0.456	0.753 ± 0.095	0.190 ± 0.070	0.814 ± 0.107
RF	0.890 ± 0.052	867.974 ± 18.656	0.877 ± 0.068	0.087 ± 0.027	0.962 ± 0.022
SVM	0.703 ± 0.083	88 400.330 ± 5003.151	0.725 ± 0.069	0.197 ± 0.038	0.767 ± 0.107

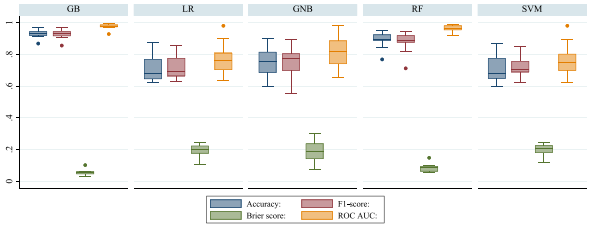

Fig. 5. Box charts for the performance metrics results.

3.12. Threats to external validity

External validity concerns the generalization of the result. One issue appears in this section, the interaction of selection and treatment.

This threat is the effect of having participants from a population that does not represent the genuine population. However, as the data comes from SNAC-B, which uses stratified sampling, this issue resolves itself. The problem in this thesis is more concerned with the high mortality rate, which can underrepresent some strata and, therefore, not accurately represent the population. This issue also resolves itself by recurrently adding new participants to the SNAC-B study.

4. Results & Analysis

4.1. Descriptive statistics

This section presents all the results from the experiment and how the data behaves. GB was the most accurate algorithm (see Table 6). Further, the Shapiro–Wilks test for normality for all performance metrics and feature importances shows if the data has a normal distribution (see Tables 8 & 7, bold cells does not have normality).

Several histograms (see Figs. 6, 7, 8 & 9), a box chart (see Fig. 5), a bar chart (see Fig. 10), ROC curves (see Fig. 11) and confusion matrices (see Fig. 12) show how the experiment's data behaves per algorithm.

The results table (see Table 6) shows the mean accuracy, execution time, F1-score, Brier score and ROC AUC of all algorithms with their respective standard deviations. The results demonstrate that GB and RF perform the best, whilst LR and SVM are the worst-performing algorithms. These results can be seen graphically in the box chart below (see Fig. 5).

Many of the algorithms' feature importances do not follow a normal distribution (see Table 7). Therefore a parametric test is not viable to assess hypothesis 3, and a non-parametric alternative is employed instead.

The majority of the performance metrics have a normal distribution (see Table 8), thus enabling using parametric tests for hypotheses 2a–2e and 4. The distributions can also be seen in the histograms below (see Figs. 6, 7, 8 & 9).

The feature importance between the algorithms varies. Two themes appear LR, GNB, and SVM have more similar feature importances, whilst GB's and RF's importances have similarities. The variation creates an ambiguity between the risk factors. However, we can observe which factors are the topmost from the graph.

Table 7
P-values for Shapiro–Wilks test on feature importance across algorithms (bold cells do not have normality).

Feature	GB	LR	GNB	RF	SVM
A1	0.105	0.988	0.584	0.689	0.285
Age	0.322	**0.012**	**0.001**	0.307	0.210
C51	0.563	0.719	**0.034**	0.221	0.960
D31	0.638	**0.002**	0.265	0.422	**0.002**
D32	**0.007**	0.493	0.385	0.292	0.509
D33	0.626	**0.011**	0.094	0.740	**0.007**
D34	0.997	0.096	0.983	0.867	0.057
D35	0.503	0.190	**0.005**	0.436	0.162
D36	0.140	0.878	0.824	0.714	0.492
D37	0.223	**0.015**	0.300	0.609	0.897
D38	0.301	0.318	0.732	0.202	0.618
D39	0.861	0.970	0.841	0.254	0.681
D40	0.287	**0.008**	0.842	0.812	0.659
E135	0.901	0.865	**0.018**	0.066	**0.026**
E36	0.813	0.702	0.808	0.270	0.883
Sex	**0.041**	**0.003**	0.865	0.426	0.224

Table 8
P-values for Shapiro–Wilks test on performance metrics across algorithms (bold cells do not have normality).

Algorithm	GB	LR	GNB	RF	SVM
Accuracy	0.692	0.998	0.998	0.057	0.444
Exec. Time	0.131	0.529	0.529	0.254	0.436
F1-score	0.465	0.775	0.775	**0.020**	0.671
Brier score	0.199	0.993	0.993	0.171	0.323
ROC AUC	**0.006**	0.881	0.881	0.356	0.569

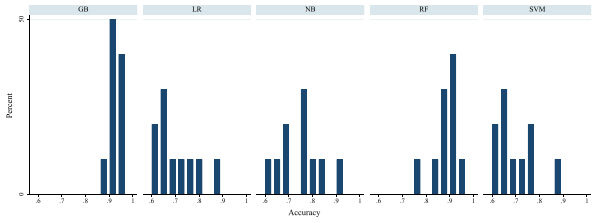

Fig. 6. Histogram of accuracy across algorithms.

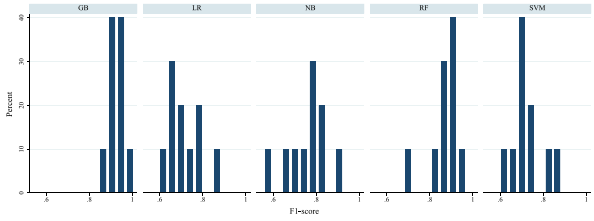

Fig. 7. Histogram of F1-score across algorithms.

As seen in the result table (see Table 6) and box chart (see Fig. 5), GB and RF perform better than the remaining algorithms. This pattern is further seen in the confusion matrices and ROC chart below (see Figs. 12 & 11).

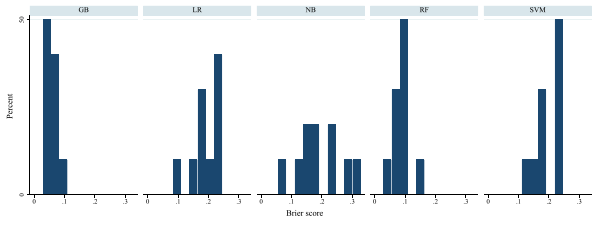

Fig. 8. Histogram of Brier score across algorithms.

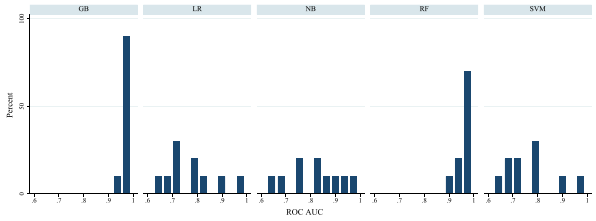

Fig. 9. Histogram of ROC AUC across algorithms.

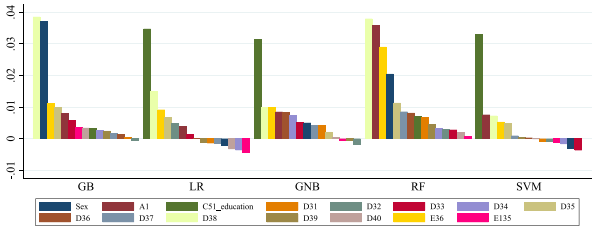

Fig. 10. All feature's feature importances across algorithms.

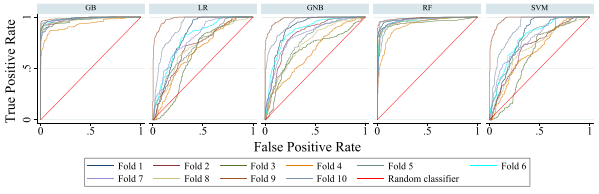

Fig. 11. ROC curves across algorithms with 10-fold cross-validation.

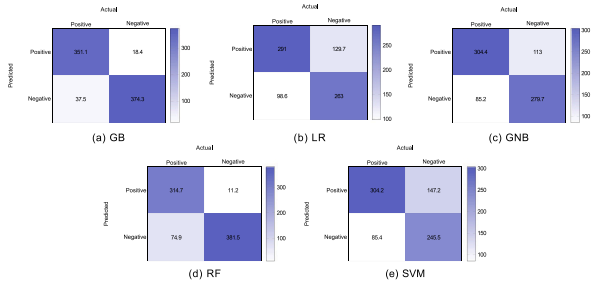

Fig. 12. Confusion matrices across all algorithms (using mean values of the 10-fold cross-validation).

5. Hypotheses testing

5.1. Hypothesis 1

An analysis of the association between sleep disturbances and dementia was conducted using multiple logistic regression. The model has a statistically better fitness than an empty model $\chi^2(12, n = 2779) = 51.19$, $p < 0001$, Pseudo $R^2 = 0.0565$. However, of the 12 variables, only two were statistically significant (see Table 9, bold cells are significant). Namely, if the participant snores ($p < 0.001$) and if they are tired and thus sleep more than two hours during the day ($p < 0.001$). However, as the confounding variables are constant and uncontrolled, they can affect the association.

As the model can find an association between sleep disturbances and dementia, we reject hypothesis 1's null hypothesis and accept the alternative one (see Section 3.5.1).

Table 9
Logistic regression on sleep disturbance features, hypothesis 1.

Feature	Odds ratio	Std. Err.	p-value	95% CI	
D31	0.165	0.305	0.589	[−0.434,	0.763]
D32	−0.030	0.488	0.952	[−0.986,	0.927]
D33	0.291	0.262	0.266	[−0.221,	0.804]
D34	0.134	0.261	0.609	[−0.379,	0.646]
D35	−0.353	0.260	0.175	[−0.863,	0.157]
D36	−0.019	0.298	0.949	[−0.603,	0.565]
D37	−0.164	0.220	0.455	[−0.596,	0.267]
D38	−1.285	0.240	**0.000**	[−1.755,	−0.815]
D39	0.027	0.136	0.842	[−0.240,	0.295]
D40	0.105	0.084	0.209	[−0.059,	0.270]
E135	−0.100	0.103	0.331	[−0.302,	0.102]
E36	0.824	0.229	**0.000**	[0.376,	1.272]

Table 10
t-tests on the Brier score across all algorithms, hypothesis 2.

Algorithm	p-value $\mu \neq 0.25$	p-value $\mu < 0.25$	p-value $\mu > 0.25$
GB	0.000	0.000	1.000
LR	0.002	0.001	0.999
GNB	0.024	0.012	0.988
RF	0.000	0.000	1.000
SVM	0.002	0.001	0.999

5.2. Hypothesis 2

Five one-tailed t-tests were performed (see Table 10), one for hypotheses 2a–2e (see Sections 3.5.3, 3.5.4, 3.5.5, 3.5.6 & 3.5.7). All t-tests reject the null hypothesis that the average Brier score is 0.25. All tests also show that the Brier score is less than 0.25. Therefore, we accept the alternative hypothesis (H_1) for all hypotheses.

5.3. Hypothesis 3

16 Kruskal–Wallis tests, one for each feature in the ML models, show a statistically significant difference in 11 of the 16 features between the five ML algorithms (see Table 11, bold cells does not reject hypothesis 3's H_0). D35, D37, D39, D40 and E135 were the five features that did not meet the required significance level. Because most features differed significantly, we can reject hypothesis 3's null hypothesis and accept the alternative (see Section 3.5.8).

5.4. Hypothesis 4

A one-way ANOVA test determines that the Brier score for each algorithm is significantly different, $F(4, 45) = 25.09, p = 0.000$. Because of the significant difference, we can reject hypothesis 4's null hypothesis and accept the alternative one (see Section 3.5.9). Furthermore, a Tukey post-hoc test shows the pairwise difference in the Brier score of the five algorithms. RF and GB have a statistically lower Brier score than SVM, GNB and LR (see Table 13 and Fig. 13).

Because of the significant difference in the Tukey test, GB and RF perform better than the other algorithms. A two-sample t-test assessed the Brier score between GB and RF. There was a significant result that the algorithms do not have an equal Brier score ($p < 0.0083$). Further, the test concludes that GB has a significantly smaller Brier score than RF ($p < 0.0042$). Therefore, GB has a significantly better accuracy performance than the remaining algorithms (see Table 12).

Table 11
Kruskal–Wallis test on the feature importances, hypothesis 3.

Feature	p-value	$\chi^2(4)$	Rank sum				
			GB	LR	GNB	RF	SVM
Sex	0.000	34.563	431	126	227	357	134
Age	0.000	39.822	77	397	282	133	386
A1	0.000	20.408	216	176	221	438	224
C51	0.000	22.413	115	330	352	160	318
D31	0.006	14.346	205	179	326	371	195
D32	0.037	10.222	220	343	160	315	238
D33	0.044	9.824	321	219	317	270	149
D34	0.002	17.026	267	139	392	281	198
D35	**0.325**	4.654	314	245	193	296	227
D36	0.023	11.356	201	191	350	329	205
D37	**0.197**	6.030	250	181	304	317	224
D38	0.000	24.257	388	198	175	367	147
D39	**0.222**	5.711	276	209	204	338	249
D40	**0.167**	6.469	323	175	243	302	233
E36	0.003	15.796	235	212	229	415	184
E135	**0.704**	2.173	304	210	242	258	262

Table 12
ANOVA test on the Brier scores, hypothesis 4.

	Sum of squares	df	F	p-value
Between groups	0.185	4	25.09	0.000
Within groups	0.083	45		

Table 13
P-values for the post-hoc Tukey test, hypothesis 4.

Algorithm pair	Contrast	Std. err.	t	p-value
LR vs. GB	0.138	0.019	7.210	**0.000**
GNB vs. GB	0.135	0.019	7.010	**0.000**
RF vs. GB	0.031	0.019	1.620	0.490
SVM vs. GB	0.141	0.019	7.350	**0.000**
GNB vs. LR	−0.004	0.019	−0.190	1.000
RF vs. LR	−0.107	0.019	−5.580	**0.000**
SVM vs. LR	0.003	0.019	0.140	1.000
RF vs. GNB	−0.103	0.019	−5.390	**0.000**
SVM vs. GNB	0.006	0.019	0.340	0.997
SVM vs. RF	0.110	0.019	5.730	**0.000**

6. Discussion

Key findings

- An association between sleep disturbance and dementia exists.
- Additional research is needed to discover more fitting sleep disturbance factors
- Identified several sleep disturbance risk factors, which have an association with dementia.
- All five algorithms can predict dementia. However, the results from this study show that gradient boosting (GB) performs best.

6.1. Association of sleep disturbances and dementia

An association between sleep disturbances and dementia exists (see Section 5.1). The regression model is significant but had a minimal effect size (Pseudo R^2 = 0.0565). The result concludes the existence of an association, but all the features are not fitting, which can be issues such as highly correlated and incorrect predictor variables [79]. Other features could find a more fitting association. Nonetheless, the association between sleep and dementia is consistent with current research [5–7,63–66], even though it is minimal.

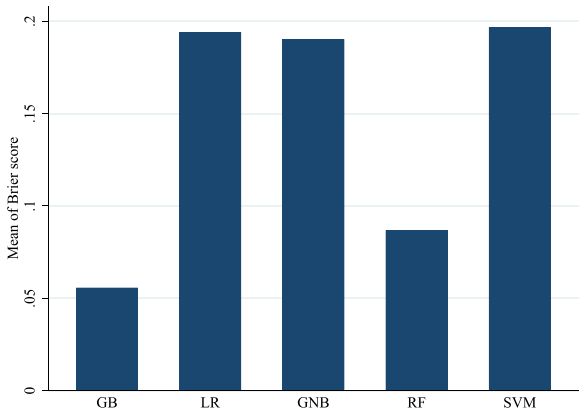

Fig. 13. Pairwise comparison of the Brier score, hypothesis 4.

6.2. Machine learning algorithm's accuracy

Another finding is that all algorithms can predict dementia better than a random classifier based on sleep factors (see Section 5.2). GB is the most accurate ML algorithm in the study (see Table 6 & Section 5.4), with the highest accuracy, F1-score, Brier score and ROC AUC.

The study's results and analysis compare the ML algorithms with a random classifier. All algorithms perform better than one. However, their worth in practice is hard to determine. To be useful in practice, we need further research on which sleep-related features impact dementia.

Furthermore, the algorithms create different models and have different feature importance. Because of the feature importance differences, the results became harder to interpret. Even though the algorithms can predict dementia with their given model, other features could improve it and provide more consistent results across algorithms.

The prediction accuracy of the algorithms in the paper and reviewed research are inconsistent. Most studies with multiple ML algorithms have SVM as the best or close to the best algorithm based on predictive accuracy [7,46,57,61,62]. The results of this study show that SVM is the worst algorithm.

One reason can be the difference in the data and sampling. The studies with SVM have not examined dementia with sleep but dementia with other factors such as EEG. Additionally, other studies without SVM [45,47] show that a version of GB outperforms RF, which is consistent with this study.

Furthermore, the oversampling techniques (see Section 1.4) impact the data and ML algorithm's results, several suggestions exist for using resampling to combat the bias towards the majority class and thus the adverse effects of vast class imbalances [80–82].

Lastly, an issue exists for two algorithms, which is the probability estimation of SVM and GNB. GNB is known to be a poor estimator [83], which can influence its probability estimation and, thus, the results of the feature importance and Brier score. SVM also has the issue with probability estimation [84], which leads to the same possible faults.

6.3. Risk factors for dementia

The feature's feature importance varies vastly (see Fig. 10 & Section 5.3). There are two themes in the results. RF's and GB's feature importances are similar, and LR's GNB's and SVM's feature importances are alike. Which features have the correct feature importance is ambiguous. However, the variables that consistently have the highest feature importance in all algorithms are: If the person sleeps more than two hours during the day, their sex, education level, age, waking up during the night and if the person snores. Therefore we conclude that these are the highest risk factors of the 16 features in the paper.

Current research also has the same results for personal risk factors as this study. As described above, age, sex, and educational level are highly important features, which is consistent with current research [9,11,47].

However, other factors affect sleep. Physical activity and mental health are risk factors for dementia [9,47,48], and they both have a bidirectional association with sleep [85,86]. Because of the bidirectionality, assessing if mental and physical health or sleep is the primary risk becomes arduous. However, if one's sleep quality becomes disturbed, it can be an early sign of issues which can increase one's risk of dementia. As described earlier (see Section 1.1), there are indications that leading a healthy lifestyle can decrease the risk and prevalence of dementia [19].

7. Conclusion

7.1. Summary

This study examines the association between sleep and dementia using several machine-learning algorithms. The algorithms are gradient boosting, logistic regression, Gaussian naive Bayes, random forest and support vector machine. Further, this paper uses data from SNAC-B (Swedish National Study on Aging and Care — Blekinge), a long-term longitudinal study on the older population in Blekinge, Sweden.

A controlled experiment is the chosen method. The primary goal is to obtain each algorithm's feature importance and Brier score, which the hypotheses use. The experiment has 4175 samples, 16 features and uses 10-fold stratified cross-validation to collect the results.

The result concludes that gradient boosting was the most accurate model with 92.9% accuracy, 0.926 f1-score, 0.974 ROC AUC and 0.056 Brier score. It further establishes that there is an association between sleep disturbances and dementia. However, which factors that are significant were different in each machine-learning algorithm. If the person sleeps more than two hours during the day, their sex, education level, age, waking up during the night and if the person snores are the variables that most consistently have the highest feature importance in all algorithms.

7.2. Impact

The results of this study have the same conclusions as the reviewed literature. There is an association between sleep and dementia. However, as in the other studies, the association is small, and which exact sleep factors that influence dementia are hard to determine. Furthermore, this study shows an indication that machine learning can help with the screening of dementia patients.

7.3. Future works

The study's results show an association between sleep disturbance and dementia. However, the association is small and further studies should use other features to examine the association further. Additionally, using data from different countries than Sweden would improve the generalizability of the results.

Using other machine learning algorithms or artificial neural networks to compare their performance accuracy would be beneficial to examine which functions the best in the given field.

Furthermore, studies with concept drift can examine if predictor features change over time. Thus correcting for this can improve learning [87]. Research on concept drift exists in areas such as software quality assurance [88], but examining if this functions well for medical applications would be valuable.

Lastly, the oversampling technique can influence the results, including other would be beneficial, such as cluster-based adaptive data augmentation (CADA) [89], synthetic minority oversampling technique (SMOTE) [90] and other newer techniques, for example, MAHAKIL [91].

CRediT authorship contribution statement

Joel Nyholm: Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Ahmad Nauman Ghazi:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Sarah Nauman Ghazi:** Conceptualization, Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Johan Sanmartin Berglund:** Data curation.

Declaration of competing interest

The authors declare no conflict of interests.

References

[1] S. Herculano-Houzel, Sleep it out, Science 342 (6156) (2013) 316–317, URL http://www.jstor.org/stable/42619900.
[2] P. Meerlo, A. Sgoifo, D. Suchecki, Restricted and disrupted sleep: Effects on autonomic function, neuroendocrine stress systems and stress responsivity, Sleep Med. Rev. 12 (3) (2008) 197–210, http://dx.doi.org/10.1016/j.smrv.2007.07.007, URL https://www.sciencedirect.com/science/article/pii/S1087079207000986.
[3] Center for disease control and prevention, Sleep and chronic disease, 2022, URL https://www.cdc.gov/sleep/about_sleep/chronic_disease.html.
[4] A.E. Ahmed, H. AL-Jahdali, A. Fatani, K. Al-Rouqi, F. AL-Jahdali, A. Al-Harbi, S. Baharoon, Y.Z. Ali, M. Khan, A. Rumayyan, The effects of age and gender on the prevalence of insomnia in a sample of the Saudi population, Ethn. Health 22 (3) (2017) 285–294, http://dx.doi.org/10.1080/13557858.2016.1244624, arXiv:https://doi.org/10.1080/13557858.2016.1244624 PMID: 27846729.
[5] A. Corbi, D. Burgos, Connection between sleeping patterns and cognitive deterioration in women with Alzheimer's disease, Sleep Breath. 26 (1) (2022) 361–371.
[6] H. Byeon, Application of machine learning technique to distinguish parkinson's disease dementia and alzheimer's dementia: Predictive power of parkinson's disease-related non-motor symptoms and neuropsychological profile, J. Personal. Med. 10 (2) (2020) 31.
[7] D. Geng, C. Wang, Z. Fu, Y. Zhang, K. Yang, H. An, Sleep EEG-based approach to detect mild cognitive impairment, Front. Aging Neurosci. 14 (2022) 865558.
[8] D.R. Lee, A.J. Thomas, Sleep in dementia and caregiving – assessment and treatment implications: a review, Int. Psychogeriatr. 23 (2) (2011) 190–201, http://dx.doi.org/10.1017/S1041610210001894.
[9] J.M. Ranson, T. Rittman, S. Hayat, C. Brayne, F. Jessen, K. Blennow, C. van Duijn, F. Barkhof, E. Tang, C.J. Mummery, B.C.M. Stephan, D. Altomare, G.B. Frisoni, F. Ribaldi, J.L. Molinuevo, P. Scheltens, D.J. Llewellyn, M. Abramowicz, M. Berthier, M. Bieler, A. Brioschi, E. Carrera, G. Chételat, C. Csajka, J.-F. Demonet, A. Dodich, B. Dubois, V. Garibotto, J. Georges, S. Hurst, M. Kivipelto, L. McWhirter, R. Milne, C. Minguillón, C. Miniussi, P.M. Nilsson, C. Ritchie, A. Solomon, W. van der Flier, B. Vellas, L. Visser, European Task Force Brain Health and on behalf of the European Task Force for Brain Health Services, Modifiable risk factors for dementia and dementia risk profiling. A user manual for brain health services—part 2 of 6, Alzheimer's Res. Ther. 13 (1) (2021) 1–169.
[10] E.S. Musiek, D.D. Xiong, D.M. Holtzman, Sleep, circadian rhythms, and the pathogenesis of Alzheimer disease, Exp. Mol. Med. 47 (3) (2015) e148.
[11] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, S.G. Costafreda, A. Dias, N. Fox, L.N. Gitlin, R. Howard, H.C. Kales, M. Kivimäki, E.B. Larson, A. Ogunniyi, V. Orgeta, K. Ritchie, K. Rockwood, E.L. Sampson, Q. Samus, L.S. Schneider, G. Selbæ k, L. Teri, N. Mukadam, Dementia prevention, intervention, and care: 2020 report of the Lancet Commission, Lancet (Br. Ed.) 396 (10248) (2020) 413–446.
[12] National Board of Health and Welfare, Nationella Riktlinjer – Utvädering 2018 Vård och Omsorg Vid Demenssjukdom, Socialstyrelsen, 2018.
[13] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning, second ed., in: Adaptive Computation and Machine Learning series, MIT Press, 2018, pp. 1–2, URL https://books.google.se/books?id=dWB9DwAAQBAJ.
[14] Institute for Health Metrics and Evaluation, GBD compare data visualization, 2020, URL http://vizhub.healthdata.org/gbd-compare.
[15] Alzheimer's Association, What is dementia?, 2023, URL https://www.alz.org/alzheimers-dementia/what-is-dementia.
[16] Center for Disease Control and Prevention, About dementia, 2019, URL https://www.cdc.gov/aging/dementia/index.html.
[17] C. Holmes, J. Amin, Dementia, Medicine 48 (11) (2020) 742–745, http://dx.doi.org/10.1016/j.mpmed.2020.08.014, URL https://www.sciencedirect.com/science/article/pii/S1357303920302073.
[18] E.L. Cunningham, B. McGuinness, B. Herron, A.P. Passmore, Dementia, Ulster Med. J. 84 (2) (2015) 79–87, URL https://www-ncbi-nlm-nih-gov.miman.bib.bth.se/pmc/articles/PMC4488926/, PMID: 26170481.

[19] National Institute on Aging, What is dementia? Symptoms, types, and diagnosis, 2022, URL https://www.nia.nih.gov/health/what-is-dementia.
[20] 1177, Demenssjukdomar, 2020, URL https://www.1177.se/sjukdomar--besvar/hjarna-och-nerver/larande-forstaelse-och-minne/demenssjukdomar/#section-115582.
[21] National Academies of Sciences, Engineering, and Medicine, Preventing Cognitive Decline and Dementia: A Way Forward, The National Academies Press, Washington, DC, 2017, http://dx.doi.org/10.17226/24782, URL https://nap.nationalacademies.org/catalog/24782/preventing-cognitive-decline-and-dementia-a-way-forward.
[22] C.S.K. Binish Javed, S.S. Hasan, Pharmacological and non-pharmacological treatment options for sleep disturbances in Alzheimer's disease, Expert Rev. Neurother. 23 (6) (2023) 501–514, http://dx.doi.org/10.1080/14737175.2023.2214316, arXiv:https://doi.org/10.1080/14737175.2023.2214316 PMID: 37267149.
[23] E.R. Dodson, P.C. Zee, Therapeutics for circadian rhythm sleep disorders, Sleep Med. Clin. 5 (4) (2010) 701–715.
[24] Center for disease control and prevention, How much sleep do I need?, 2022, URL https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html.
[25] National health service, Insomnia, 2021, URL https://www.nhs.uk/conditions/insomnia/.
[26] H. Walker, W. Hall, J. Hurst, Clinical Methods: The History, Physical, and Laboratory Examinations, in: NCBI Bookshelf, Butterworths, 1990, pp. 398–403, (Chapter 77).
[27] K.L. Nelson, J.E. Davis, C.F. Corbett, Sleep quality: An evolutionary concept analysis, Nurs. Forum 57 (1) (2022) 144–151, http://dx.doi.org/10.1111/nuf.12659, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/nuf.12659 URL https://onlinelibrary.wiley.com/doi/abs/10.1111/nuf.12659.
[28] Swedish National Data Service, SNAC - Swedish national study on aging and care, 2020, URL https://snd.gu.se/en/catalogue/study/ext0124.
[29] SNAC Blekinge, Redogörelse för verksamheten 2020, 2020, URL https://snac.nu/wp-content/uploads/2021/10/Arsrapport-SNAC-B-2020.pdf.
[30] R. Peters, E. Pinto, Literature review of the Clock Drawing Test as a tool for cognitive screening, Dement. Geriatr. cogn. Disord. 27 (3) (2009) 201–213, http://dx.doi.org/10.1159/000203344, URL https://pubmed.ncbi.nlm.nih.gov/19225234/.
[31] J. Cacho, J. Benito-León, R.G. a García, B. Fernández-Calvo, J.L. Vicente-Villardón, A.J. Mitchell, Does the combination of the MMSE and clock drawing test (mini-clock) improve the detection of mild Alzheimer's disease and mild cognitive impairment? J. Alzheimer's Dis. 22 (2010) 889–896, http://dx.doi.org/10.3233/JAD-2010-101182, URL https://pubmed.ncbi.nlm.nih.gov/20858951/.
[32] I. Aprahamian, J.E. Martinelli, A.L. Neri, M.S. Yassuda, The accuracy of the Clock Drawing Test compared to that of standard screening tests for Alzheimer's disease: results from a study of Brazilian elderly with heterogeneous educational backgrounds, Int. Psychogeriatr. 22 (1) (2010) 64–71, http://dx.doi.org/10.1017/S1041610209991141.
[33] National Board of Health and Welfare, MMSE, MMT (mini mental state examination, mini mental test), 2019, URL https://www.socialstyrelsen.se/kunskapsstod-och-regler/omraden/evidensbaserad-praktik/metodguiden/mmse-mmt-mini-mental-state-examination-mini-mental-test/.
[34] P. Manos, Ten-point clock test sensitivity for Alzheimer's disease in patients with MMSE scores greater than 23, Int. J. Geriatr. Psychiatry 14 (6) (1999) 454–458, URL https://pubmed.ncbi.nlm.nih.gov/20858951/, PMID: 26170481.
[35] S.B. Kotsiantis, D. Kanellopoulos, P.E. Pintelas, Data preprocessing for supervised leaning, Int. J. Comput. Sci. 1 (2006) 111–117.
[36] J. Huang, Y.-F. Li, M. Xie, An empirical analysis of data preprocessing for machine learning-based software cost estimation, Inf. Softw. Technol. 67 (2015) 108–127, http://dx.doi.org/10.1016/j.infsof.2015.07.004, URL https://www.sciencedirect.com/science/article/pii/S0950584915001275.
[37] D.M.P. Murti, U. Pujianto, A.P. Wibawa, M.I. Akbar, K-nearest neighbor (K-NN) based missing data imputation, in: 2019 5th International Conference on Science in Information Technology (ICSITech), 2019, pp. 83–88, http://dx.doi.org/10.1109/ICSITech46713.2019.8987530.
[38] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (6) (2001) 520–525, http://dx.doi.org/10.1093/bioinformatics/17.6.520, arXiv:https://academic.oup.com/bioinformatics/article-pdf/17/6/520/48837104/bioinformatics_17_6_520.pdf.
[39] R. Malarvizhi, A.S. Thanamani, K-nearest neighbor in missing data imputation, Int. J. Eng. Res. Dev. 5 (1) (2012) 5–7.
[40] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322–1328, http://dx.doi.org/10.1109/IJCNN.2008.4633969.
[41] M. Baimakhanbetov, K. Nurumov, U. Ospanova, T. Buldybayev, I. Akoyeva, The effect of the ADASYN method on widespread metrics of machine learning efficiency, Mod. Inf. Technol. IT-Educ. 15 (2) (2019) 290–297, http://dx.doi.org/10.25559/SITITO.15.201902.290-297, URL http://sitito.cs.msu.ru/index.php/SITITO/article/view/518.
[42] N.G. Ramadhan, Comparative analysis of ADASYN-SVM and SMOTE-SVM methods on the detection of type 2 diabetes mellitus, Sci. J. Inform. (Semarang) 8 (2) (2021) 276–282.
[43] G. Ahmed, M.J. Er, M.M.S. Fareed, S. Zikria, S. Mahmood, J. He, M. Asad, S.F. Jilani, M. Aslam, DAD-Net: Classification of Alzheimer's disease using ADASYN oversampling technique and optimized neural network, Mol. (Basel Switz.) 27 (20) (2022) 7085.
[44] J. Miller, J. Daly, M. Wood, M. Roper, A. Brooks, Statistical power and its subcomponents — missing and misunderstood concepts in empirical software engineering research, Inf. Softw. Technol. 39 (4) (1997) 285–295.
[45] L. Liu, B. Yu, M. Han, S. Yuan, N. Wang, Mild cognitive impairment understanding: An empirical study by data-driven approach, BMC Bioinform. 20 (Suppl 15) (2019) 481.
[46] S.C. Mallo, S. Valladares-Rodriguez, D. Facal, C. Lojo-Seoane, M.J. Fernández-Iglesias, A.X. Pereiro, Neuropsychiatric symptoms as predictors of conversion from MCI to dementia: A machine learning approach, Int. Psychogeriatr. 32 (3) (2020) 381–392.
[47] S.O. Danso, Z. Zeng, G. Muniz-Terrera, C.W. Ritchie, Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms, Front. Big Data 4 (2021) 613047.
[48] K.J. Anstey, R. Peters, M.E. Mortby, K.M. Kiely, R. Eramudugolla, N. Cherbuin, M.H. Huque, R.A. Dixon, Association of sex differences in dementia risk factors with sex differences in memory decline in a population-based cohort spanning 20–76 years, Sci. Rep. 11 (1) (2021) 7710.
[49] M. Dauwan, J.J. van der Zande, E. van Dellen, I.E.C. Sommer, P. Scheltens, A.W. Lemstra, C.J. Stam, Random forest to differentiate dementia with Lewy bodies from Alzheimer's disease, Alzheimer's Dement. : Diagn. Assess. Dis. Monit. 4 (1) (2016) 99–106.
[50] R. Cassani, T.H. Falk, F.J. Fraga, M. Cecchi, D.K. Moore, R. Anghinah, Towards automated electroencephalography-based Alzheimer's disease diagnosis using portable low-density devices, Biomed. Signal Process. Control 33 (2017) 261–271.
[51] C. Davatzikos, S.M. Resnick, X. Wu, P. Parmpi, C.M. Clark, Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI, NeuroImage (Orlando Fla.) 41 (4) (2008) 1220–1227.
[52] S. Klöppel, C.M. Stonnington, C. Chu, B. Draganski, R.I. Scahill, J.D. Rohrer, N.C. Fox, C.R. Jack, J. Ashburner, R.S.J. Frackowiak, Automatic classification of MR scans in Alzheimer's disease, Brain (Lond. Engl.: 1878) 131 (3) (2008) 681–689.
[53] A. Katako, P. Shelton, A.L. Goertzen, D. Levin, B. Bybel, M. Aljuaid, H.J. Yoon, D.Y. Kang, S.M. Kim, C.S. Lee, J.H. Ko, Machine learning identified an alzheimer's disease-related FDG-PET pattern which is also expressed in Lewy body dementia and Parkinson's disease dementia, Sci. Rep. 8 (1) (2018).
[54] S.J. Colloby, R.A. Cromarty, L.R. Peraza, K. Johnsen, G. Jóhannesson, L. Bonanni, M. Onofrj, R. Barber, J.T. O'Brien, J.-P. Taylor, Multimodal EEG-MRI in the differential diagnosis of Alzheimer's disease and dementia with Lewy bodies, J. Psychiatr. Res. 78 (2016) 48–55.
[55] C. Park, J. Kim, J. Kim, S. Park, Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles, PLoS One 13 (7) (2018) 1–15, http://dx.doi.org/10.1371/journal.pone.0201056.
[56] X. Huang, H. Liu, X. Li, L. Guan, J. Li, L.C.A.M. Tellier, H. Yang, J. Wang, J. Zhang, Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning, BMC Neurol. 18 (1) (2018) 5.
[57] L. Xu, G. Liang, C. Liao, G.-D. Chen, C.-C. Chang, An efficient classifier for Alzheimer's disease genes identification, Molecules (Basel Switz.) 23 (12) (2018) 3140.
[58] A. Shimoda, Y. Li, H. Hayashi, N. Kondo, Dementia risks identified by vocal features via telephone conversations: A novel machine learning prediction model, PLoS One 16 (7) (2021) e0253988.
[59] J. Weiner, M. Engelbart, T. Schultz, Manual and automatic transcriptions in dementia detection from speech, in: Proc. Interspeech 2017, 2017, pp. 3117–3121, http://dx.doi.org/10.21437/Interspeech.2017-112.
[60] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P.H. Robert, R. David, Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease, Alzheimer's Dement. : Diagn. Assess. Dis. Monit. 1 (1) (2015) 112–124.
[61] K. Lopez-de Ipina, U. Martinez-de Lizarduy, P.M. Calvo, B. Beitia, J. Garcia-Melero, M. Ecay-Torres, A. Estanga, M. Faundez-Zanuy, Analysis of disfluencies for automatic detection of mild cognitive impartment: a deep learning approach, in: 2017 International Conference and Workshop on Bioinspired Intelligence, IWOBI, 2017, pp. 1–4, http://dx.doi.org/10.1109/IWOBI.2017.7985526.
[62] K.C. Fraser, K.L. Fors, D. Kokkinakis, A. Nordlund, An analysis of eye-movements during reading for the detection of mild cognitive impairment, in: Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1016–1026.
[63] K. Yaffe, A.M. Laffan, S.L. Harrison, S. Redline, A.P. Spira, K.E. Ensrud, S. Ancoli-Israel, K.L. Stone, Sleep-disordered breathing, hypoxia, and risk of mild cognitive impairment and dementia in older women, JAMA 306 (6) (2011) 613–619, http://dx.doi.org/10.1001/jama.2011.1115, arXiv:https://jamanetwork.com/journals/jama/articlepdf/1104205/joc15090_613_619.pdf.

[64] A.S.P. Lim, M. Kowgier, L. Yu, A.S. Buchman, D.A. Bennett, Sleep fragmentation and the risk of incident Alzheimer's disease and cognitive decline in older persons, Sleep 36 (7) (2013) 1027–1032, http://dx.doi.org/10.5665/sleep.2802, arXiv:https://academic.oup.com/sleep/article-pdf/36/7/1027/26661156/aasm.36.7.1027.pdf.

[65] A. Tsapanou, G.S. Vlachos, S. Cosentino, Y. Gu, J.J. Manly, A.M. Brickman, N. Schupf, M.E. Zimmerman, M. Yannakoulia, M.H. Kosmidis, E. Dardiotis, G. Hadjigeorgiou, P. Sakka, Y. Stern, N. Scarmeas, R. Mayeux, Sleep and subjective cognitive decline in cognitively healthy elderly: Results from two cohorts, J. Sleep Res. 28 (5) (2019) e12759, http://dx.doi.org/10.1111/jsr.12759, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12759 URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12759.

[66] A. Behrens, P. Anderberg, J.S. Berglund, Sleep disturbance predicts worse cognitive performance in subsequent years: A longitudinal population-based cohort study, Arch. Gerontol. Geriatr. 106 (2023) 104899, http://dx.doi.org/10.1016/j.archger.2022.104899, URL https://www.sciencedirect.com/science/article/pii/S0167494322002862.

[67] S. Glen, Brier score, 2023, URL https://www.statisticshowto.com/brier-score/.

[68] scikit-learn developers, Permutation feature importance, 2023, URL https://scikit-learn.org/stable/modules/permutation_importance.html#permutation-importance.

[69] A.A. Bradley, S.S. Schwartz, T. Hashino, Sampling uncertainty and confidence intervals for the brier score and brier skill score, Weather Forecast. 23 (5) (2008) 992–1006, http://dx.doi.org/10.1175/2007WAF2007049.1, URL https://journals.ametsoc.org/view/journals/wefo/23/5/2007waf2007049_1.xml.

[70] R. Benedetti, Scoring rules for forecast verification, Mon. Weather Rev. 138 (1) (2010) 203–211, http://dx.doi.org/10.1175/2009MWR2945.1, URL https://journals.ametsoc.org/view/journals/mwre/138/1/2009mwr2945.1.xml.

[71] S. Jewson, The problem with the Brier score, 2004, arXiv:physics/0401046.

[72] K.S. Gyu, K.J. Hae, Central limit theorem: the cornerstone of modern statistics, Korean J. Anesthesiol. 70 (2) (2017) 144–156, http://dx.doi.org/10.4097/kjae.2017.70.2.144, arXiv:http://ekja.org/journal/view.php?number=8274 URL http://ekja.org/journal/view.php?number=8274.

[73] scikit-learn developers, Sklearn.ensemble.GradientBoostingClassifier, 2023, URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html.

[74] scikit-learn developers, Sklearn.naive_bayes.GaussianNB, 2023, URL https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html.

[75] scikit-learn developers, Sklearn.linear_model.LogisticRegression, 2023, URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[76] scikit-learn developers, Sklearn.ensemble.RandomForestClassifier, 2023, URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[77] scikit-learn developers, Sklearn.svm.SVC, 2023, URL https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[78] scikit-learn developers, Cross-validation iterators with stratification based on class labels, 2023, URL https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation-iterators-with-stratification-based-on-class-labels.

[79] P. Ranganathan, C.S. Pramesh, R. Aggarwal, Common pitfalls in statistical analysis: Logistic regression, Perspect. Clin. Res. 8 (3) (2017) 148–151, URL https://pubmed.ncbi.nlm.nih.gov/20858951/, PMID: 28828311.

[80] A. Javeed, A.L.D. Moraes, J. Sanmartin Berglund, P. Anderberg, An intelligent learning system for unbiased prediction of dementia based on autoencoder and adaboost ensemble learning, Life 12 (7) (2022) http://dx.doi.org/10.3390/life12071097, open access.

[81] K.E. Bennin, A. Tahir, S.G. MacDonell, J. Börstler, An empirical study on the effectiveness of data resampling approaches for cross-project software defect prediction, IET Softw. 16 (2) (2022) 185–199, http://dx.doi.org/10.1049/sfw2.12052, open access.

[82] A. Viloria, O.B. Pineda Lezama, N. Mercado-Caruzo, Unbalanced data processing using oversampling: Machine learning, Procedia Comput. Sci. 175 (2020) 108–113, http://dx.doi.org/10.1016/j.procs.2020.07.018, URL https://www.sciencedirect.com/science/article/pii/S1877050920316975, The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC), The 15th International Conference on Future Networks and Communications (FNC), The 10th International Conference on Sustainable Energy Information Technology.

[83] scikit-learn developers, 1.9. Naive Bayes, 2023, URL https://scikit-learn.org/stable/modules/naive_bayes.html.

[84] scikit-learn developers, 1.4.1.2. Scores and probabilities, 2023, URL https://scikit-learn.org/stable/modules/svm.html#scores-and-probabilities.

[85] J.A. Nota, C. Chu, C. Beard, T. Björgvinsson, Temporal relations among sleep, depression symptoms, and anxiety symptoms during intensive cognitive–behavioral treatment., J. Consult. Clin. Psychol. 88 (11) (2020) 971–982, URL http://miman.bib.bth.se/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2020-82285-001&site=ehost-live&scope=site.

[86] B.-H. Huang, M. Hamer, M.J. Duncan, P.A. Cistulli, E. Stamatakis, The bidirectional association between sleep and physical activity: A 6.9 years longitudinal analysis of 38,601 UK Biobank participants, Prev. Med. 143 (2021) 106315, http://dx.doi.org/10.1016/j.ypmed.2020.106315, URL https://www.sciencedirect.com/science/article/pii/S009174352030339X.

[87] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, Learning under concept drift: A review, IEEE Trans. Knowl. Data Eng. 31 (12) (2019) 2346–2363, http://dx.doi.org/10.1109/TKDE.2018.2876857.

[88] K.E. Bennin, N.b. Ali, J. Börstler, X. Yu, Revisiting the impact of concept drift on just-in-time quality assurance, in: 2020 IEEE 20th International Conference on Software Quality, Reliability and Security, QRS, 2020, pp. 53–59, http://dx.doi.org/10.1109/QRS51102.2020.00020.

[89] S.K. Dasari, A. Cheddad, J. Palmquist, L. Lundberg, Clustering-based adaptive data augmentation for class-imbalance in machine learning (CADA) : Additive manufacturing use-case, Neural Comput. Appl. (2022) http://dx.doi.org/10.1007/s00521-022-07347-6, open access.

[90] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2011) 321–357.

[91] K.E. Bennin, J. Keung, P. Phannachitta, A. Monden, S. Mensah, [Journal first] MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction, in: 2018 IEEE/ACM 40th International Conference on Software Engineering, ICSE, 2018, p. 699, http://dx.doi.org/10.1145/3180155.3182520.