scientific reports

Check for updates

OPEN

Pitfalls in using ML to predict cognitive function performance

Gianna Kuhles[1,2]✉, Sami Hamdan[1,2], Stefan Heim[3,4,5], Simon B. Eickhoff[1,2], Kaustubh R. Patil[1,2], Julia A. Camilleri[1,2,6] & Susanne Weis[1,2,6]

Machine learning analyses are widely used for predicting cognitive abilities, yet there are pitfalls that need to be considered during their implementation and interpretation of the results. Hence, the present study aimed at drawing attention to the risks of erroneous conclusions incurred by confounding variables illustrated by a case example predicting executive function (EF) performance by prosodic features. Healthy participants ($n = 231$) performed speech tasks and EF tests. From 264 prosodic features, we predicted EF performance using 66 variables, controlling for confounding effects of age, sex, and education. A reasonable prediction performance was apparently achieved for EF variables of the Trail Making Test. However, in-depth analyses revealed indications of confound leakage, leading to inflated prediction accuracies, due to a strong relationship between confounds and targets. These findings highlight the need to control confounding variables in ML pipelines and caution against potential pitfalls in ML predictions.

Prediction of cognitive performance is a central goal in neuroscience and related areas of research. Predicting cognitive performance is relevant for several reasons. Firstly, it enables the identification of individuals who may be at risk of cognitive decline or neurodegenerative diseases at an early stage[1–6]. This, in turn, allows for preventative measures and early treatment. Secondly, predicting cognitive performance can help us understand the underlying mechanisms of cognitive function and identify potential biomarkers for cognitive abilities[7,8]. Thirdly, it can aid in the development of personalised training programs based on an individual's cognitive capabilities[9].

With the rising number of variables potentially related to cognitive performance, methods for predicting cognitive functions also increase in complexity. Machine learning (ML) offers a way to study individual differences by inspecting many different possible influencing factors. ML is a field of artificial intelligence in which models are trained on data, allowing them to uncover intricate relationships and improve over time. It involves advanced statistical algorithms, which learn patterns from feature-target data with the aim to generalise to previously unseen data[10]. Such methods are of practical use for exploratory research in various fields because unknown, linear, but most importantly non-linear, relationships of a large number of variables can be inspected easily and fast. ML approaches are gaining more importance as they are able to predict the target value of an unseen individual using their features. For instance, when impaired prosodic abilities are related to a disorder, a ML model could be useful for early detection and diagnosis. However, application of ML can be problematic when applied inappropriately, leading to inaccurate results and misleading conclusions.

One of the main challenges in ML relates to preventing models from displaying prediction values that are overly high in comparison to their actual predictive power[10,11]. Barring other reasons, this is usually the case when information that should be kept strictly separate is unintentionally fed into the ML pipeline. This process is referred to as leakage[11,12]. One form of leakage is the incorporation of information from confounding variables through the procedure of confound removal, i.e. confound-leakage[13]. Confound removal refers to the regression of the confounding effect from the data. Regressing out such confounds from the analysis of interest is standard practice in neuroscience and many fields of empirical research[14]. This approach is crucial when the primary goal is to examine the relationship between a feature variable and a target variable without the unwanted influence of a third, potentially biasing factor - commonly referred to as a confound variable. More precisely, confounding variables share variance with both the dependent (target) and the independent (explanatory or predictive) variable. This means that they are associated with both variables in the analysis and can potentially have an impact

[1]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany. [2]Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Jülich, Jülich, Germany. [3]Department of Psychiatry, Psychotherapy and Psychosomatics, Medical Faculty, RWTH Aachen University, Aachen, Germany. [4]Institute for Midwifery Science, Medical Faculty, RWTH Aachen University, Aachen, Germany. [5]Institute of Neuroscience and Medicine, Structural and Functional Organization of the Brain (INM-1), Research Centre Jülich, Jülich, Germany. [6]Julia A. Camilleri and Susanne Weis contributed equally to this work. ✉email: g.kuhles@fz-juelich.de

on the relationship between them. It is desirable to remove the confounding information such that the model's predictions are not influenced by it. A typical example can be found in models trained to predict intelligence, which may yield statistically significant results by relying solely on variance associated with age, rather than capturing genuine cognitive ability. By statistically controlling for confounding variables, one aims to isolate the effect of the predictor of interest, thereby improving the interpretability and validity of the results. However, it is plausible that confound removal procedures might inadvertently introduce confounding information rather than removing it, causing confound leakage[15]. Hamdan and colleagues[13] showed that confound leakage may arise when standard linear confound regression is used in combination with nonlinear machine learning models, meaning that, paradoxically, confound-related variance can be introduced into the features during the confound removal process itself. This risk is amplified when confounds strongly correlate with the target or when using many features, potentially biasing model outcomes[13].

In the following, we demonstrate this issue using a specific example from our research, which aimed to predict cognitive performance based on prosodic variables. As executive functions are crucial cognitive capabilities in everyday human life and constitute a basic requirement for speech and communication[16–18], we focused on predicting executive function performance in this particular application.

The term "executive functions" represents a heterogeneous set of distinguishable processes[19]. According to Ward, executive functions represent complex abilities, with which people optimise their performance in situations that require the organisation of a series of cognitive processes[20]. In spite of the lack of a universal definition of executive function performance and its subordinated domains[21], the grouping of working memory, inhibition, and cognitive flexibility[22,23] is still the most popular[24].

Executive functions are of great relevance in relation to various pathologies, as their impairment can be observed in numerous neurological and psychiatric disorders[25–29]. For this reason, their investigation, both in healthy people and in different patient groups, constitutes a central component of research and diagnostics. Despite great efforts, examination and characterisation of executive functions — and of other domains typically assessed in neuropsychological evaluations — have proven to be extremely difficult[30]. Not only is data acquisition time-consuming and costly, but the results are also dependent on subjective application factors, such as the qualification of the test conductor and the current condition of the person being tested. In addition, the measured performance depends on the individual's motivation.

What we can take advantage of in the context of testing EF is the knowledge about the relationship between executive functions and language: It is assumed that executive functions act as a cognitive control mechanism for the syntactic processing of sentences[31]. Moreover, a large variety of disorders in communication ability are associated with impaired executive functions, including dysarthria, aphasia, language pragmatic disturbances, and verbal reasoning impairments[16]. In addition to the symptoms shown on the linguistic levels of phonetics and phonology, morphology and syntax, semantics and pragmatics, the described aspects of the impaired language function also relate to the level of prosody.

Prosody can be defined as the totality of all acoustically perceptible forms of expression of speech[32]. Since prosody belongs to the realm of the phonetic structures of language and is not tied to the categories of lexeme, morpheme or phoneme, prosodic subfunctions belong to the class of suprasegmentals of language. Although several classifications of prosody have been proposed, four main domains can be distinguished: frequency related parameters, energy/amplitude related parameters, spectral parameters, and temporal parameters[33].

Against the background of current literature regarding the connections between linguistic and cognitive processes, methods can be developed to draw conclusions about underlying cognitive performance with the help of speech variables. In particular, the analysis of prosodic features by spontaneous speech samples provides advantages, as it offers a high external validity as well as time and cost efficiency compared to classical diagnostic procedures[34–36]. This is why procedures for objective speech analysis are gaining increasing popularity and are already in use in clinical diagnostics[37,38].

Studies suggest that prosodic impairments may occur due to immature executive functions[39]. In addition, earlier patient studies have already shown a connection between right-hemispheric frontal brain damage and impairments of prosody[40,41]. Recent studies also demonstrated a relation between suprasegmental disorders, regarding impaired executive functions in Foreign Accent Syndrome[42,43]. Moreover, impaired working memory and impairment in prosody were observed in Parkinson's Disease[44], while reduced performance of fundamental frequency in connection with executive function damage was shown in frontotemporal dementia[45]. Furthermore, a link between prosody and divided attention, working memory and inhibition was shown in Autism Spectrum Disorder[46]. There is also clinical evidence that formant frequencies and Mel Frequency Cepstral Coefficients are associated with depressive disorders and potentially act as a biomarker[47–50]. A relationship between prosodic performance, precisely disfluencies and inhibition in healthy participants was also reported by Engelhardt and colleagues[51].

In summary, a link between deficient executive subfunctions and impaired prosodic skills has been reported in different pathologies[36–38,48]. These associations can be utilised to predict cognitive functions. However, these findings are primarily based on patient studies and a limited selection of variables. Moreover, these studies often relied on manually extracted prosodic features, limiting their replicability and usefulness due to a lack of objectivity[34]. Therefore, our initial aim was to systematically test whether the reported correlations could predict cognitive performance in a healthy sample, using a fully automated feature extraction approach.

Methods

Participants

Participants were recruited at the Forschungszentrum Jülich and through social networks. Testing took place at the Forschungszentrum Jülich (Germany) in 2018. Each test session lasted between 150 and 180 min, depending on the participants' speed and the duration of the instructions. 231 healthy participants without a diagnosis

of neurological or mental impairment were included in the present study (138 female, 93 male). The mean age of the sample at testing time was 35.2 years (standard deviation = 11.1, minimum = 20, maximum = 55). All participants were monolingual German. The sample varies regarding the level of education, ranging from participants who finished secondary school ($n = 8$), professional school/job training ($n = 62$), high school with a university-entrance diploma ($n = 69$), and university with a university degree ($n = 92$). All participants were paid an expense allowance of 50 EUR. The study was approved by the ethics committee of Heinrich Heine University Düsseldorf under the registration number 2,017,064,341. Informed consent was obtained from all participants. All experiments were performed in accordance with relevant named guidelines and regulations. Part of the data used in this study is publicly available upon request, as not all participants consented to data sharing[52].

Design

The test sessions were divided into two parts: Firstly, the executive performance of the participants was assessed. Secondly, spontaneous speech performance was recorded in order to extract prosodic features from speech samples.

The executive function performance was assessed by the computerized test batteries *Vienna Testsystem*[53] and *Psytoolkit*[54], containing common standard tests for measuring executive function performance. In total, 66 variables from 14 different assessments of executive function performance were collected: Trail Making Test (TMT)[55], Raven's Standard Progressive Matrices[56], Wisconsin Card Sorting Test[57], Tower of London[58], and Cued Task Switching[59] are related to cognitive flexibility. Performance of N-back Non-verbal[60], Non-verbal Learning Test[61], and Corsi Block Tapping Test[62] were used in relation to working memory. Inhibition was tested by Stop Signal Task[63], Simon Task[64], and Stroop Test[65]. Divided Attention Test[66], Spatial Attention Test[66], and Mackworth Clock Test[67] were used to measure divided and spatial attention as well as vigilance. An overview of the assessed tests and the exact variables from these are shown in Table 1 (see Appendix A for the descriptions of the tests).

Spontaneous speech was tested based on a collection of three different speech samples per participant. Firstly, the participants were asked to describe the *Cookie Theft Picture*[68] within 90 s in as much detail as possible. Secondly, the participants were asked to talk about what they had watched on television/what kind of book they had read the day before. Thirdly, the participants were asked to describe what their favourite holiday trip would look like if money and time were no limiting factors. For the narrative tasks retelling a story and fictional storytelling, they were asked to talk for five minutes. Participants conducted all tests via a laptop, an external keyboard, and a headset-microphone.

Test	Abbreviation	Variables
COGNITIVE FLEXIBILITY		
Trail Making Test	TMT	Processing time part A, processing time part B, difference part B-A [seconds], quotient B/A, errors part A, errors part B
Raven's Standard Progressive Matrices	SPM	Correct items, processing time
Wisconsin Card Sorting Test	WCST	Number of errors, number of perseveration errors, number of errors (non perseveration), timeouts
Tower of London	TOL	Planning ability, number of correct responses, changed his/her mind, self-correction, choice of wrong pole, choice of blocked pole, choice of impossible position
Cued Task Switching	SWITCH	Number of errors, timeouts, errors of items which are incongruent
WORKING MEMORY		
N-back Non-Verbal	NBN	Correct items, number of commission errors, number of errors, mean reaction tine of correct items [seconds], mean reaction time of errors [seconds]
Non-Verbal Learning Test	NVLT	Sum of correct responses, sum of false responses, sum of difference between correct minus false responses, processing time
Corsi Block Tapping Test	CORSI	Block span, correct items, false items, missed items, sequency errors
INHIBITION		
Stop Signal Task	INHIB	Reaction time [seconds], mean stop signal delay [seconds], stop signal reaction time [seconds], number of commission errors, Number of ommission errors
Simon Task	SIMON	Number of errors in compatible items, Number of errors in incompatible items
Stroop Test	STROOP	Reading interference [seconds], naming interference [seconds], interference-difference [seconds], number of false reactions (reading-baseline), number of false reactions (naming-baseline), number of false reactions (reading-interference), number of false reactions (naming-interference), processing time
ATTENTION / VIGILANCE		
Divided Attention Test	WAF-G	Number of missed items (unimodall visual), number of false alarm (unimodal visual), mean reaction time (unimodal visual) [ms], number of missed items (crossmodal visual/auditive), number of false alarm (crossmodal visual/auditive), mean reaction time (crossmodal) [ms]
Spatial Attention Test	WAF-R	Mean reaction time (unannounced items) [ms], number of missed items (correct announced items), mean reaction time (correct announced items) [ms], number of missed items (wrong announced items), mean reaction time (wrong announced items) [ms], mean reaction time (short SOA) [ms], mean reaction time (long SOA) [ms], number of errors
Mackworth Clock Test	MACK	Number of missed jumps, number of false alarms

Table 1. Assessed executive function variables adapted from Amunts et al[69,70].

Feature extraction

To generate the prosodic features from the audio files collected from the speech tasks, the toolbox OpenSmile (**open**-**S**ource **M**edia **I**nterpretation by **L**arge feature-space **E**xtraction)[71], version 2.1.3, was used to extract the suprasegmental parameters. Although the extraction and analysis of prosodic parameters for research purposes have been done for decades in various fields and is currently a topic of big interest in the context of speech biomarkers in different pathologies[34] a lack of standardisation and thus comparability was observed[71]. The benefit of using the open-source toolbox OpenSmile is its standardised automatic computation of the prosodic features, resulting in a fixed feature set. It offers the extraction of prosodic features within a set that corresponds to the main categories frequency (representing the fundamental frequency), energy/amplitude (representing the intensity), spectral parameters, and temporal parameters (representing the duration). The choice of parameters was guided by the criteria of potentially indexing physiological changes in voice production and its theoretical significance in previous literature[33]. The feature set *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) was chosen, which contains 88 prosodic features. In order to keep the extraction comparable, the first 90 s from each audio file were chosen as input. As there are three audio samples per participant, a total of 264 prosodic features were generated per participant. All features were z-scored, i.e. the mean value was removed, and the variance was scaled to one unit. An overview of the extracted features and their descriptions, as well as the corresponding prosodic category, are shown in Table 2.

Machine learning and statistical analyses

Data management and analysis were performed using Python 3[72]. A ML approach was applied to the data following the machine learning library JuLearn[73]. The 264 extracted prosodic feature variables were specified as

Prosodic feature	Variables	Description
FREQUENCY RELATED PARAMETERS		
F0semitone Mean, standard deviation, percentiles, range, rising slope, falling slope	10	Pitch, logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0)
Jitter Mean, standard deviation	2	Deviations in individual consecutive F0 period lengths
F 1–3 frequency & bandwith Mean, standard deviation	12	Centre frequency of 1., 2., 3. formant, bandwidth of first formants 1, 2, 3
ENERGY / AMPLITUDE RELATED PARAMETERS		
Loudness Mean, standard deviation, percentiles, range, rising slope, falling slope	10	Estimation of perceived signal intensity from an auditory spectrum
Shimmer Mean, standard deviation	2	Difference in peak amplitudes of consecutive F0 periods
Harmonics to noise ratio Mean, standard deviation	2	Relation of energy in harmonic components to energy in noise- like components
SPECTRAL PARAMETERS		
Spectral flux Mean, standard deviation	3	Difference of the spectra of two consecutive frames
Mel frequency cepstral coefficients 1–4 Mean, standard deviation	16	Perceived pitch of the frequency spectrum
Harmonic differences Mean, standard deviation	4	Ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2)/to the energy of the highest harmonic in the third formant range (A3)
Alpha ratio Mean, standard deviation	3	Ratio of summed energy from 50–1000 Hz and 1–5 kHz
Hammerberg Index Mean, standard deviation	3	Ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
Spectral slopes Mean, standard deviation	6	Linear regression slope of the logarithmic power spectrum in the specified bands
F 1–3 energy Mean, standard deviation	6	Formant 1, 2, and 3 relative energy
TEMPORAL PARAMETERS		
Loudness peaks per second	1	Number of volume highlights per second
Voiced segments Mean, standard deviation	3	Amount of continuously voiced regions
Unvoiced segments Mean, standard deviation	2	Amount of the continuously unvoiced regions
Equivalent sound level	1	Sound pressure level which has the same total energy as the actual fluctuating noise

Table 2. Grouped listing of the prosodic features extracted by the toolbox opensmile[71].

features and the 66 executive function variables as targets. The initial goal of our analyses was to predict each of the executive function targets using all of the prosodic features.

Firstly, cross-validation was used to determine the model performance. In cross-validation, the dataset is randomly partitioned into equally sized folds. All folds except for one are used for training the model. The hold-out fold is then used to determine the trained model's performance on unseen data. This process is repeated once for each fold as the validation fold. Then, the average of the validation performances is calculated[74]. Cross-validation was applied with ten folds (Fig. 1). Since all of the prosodic features were used to predict each of the 66 targets, 66 independent cross-validation models were performed.

In order to keep the folds balanced, stratification by target was implemented into the cross-validation pipeline, meaning that the different folds approximately followed the same distribution of the respective target[15]. Stratification can usually improve the success of model training by ensuring that the training and test data have similar distribution which reduces the risk of bias or error in the evaluation of the model. Knowing the influence of different demographic aspects on prosodic performance[75,76] we regressed out the effects of the confounding variables sex, age, and education from the features with a linear regression model. This is standard practice since the goal is to shed light on the relationship between executive functions and prosodic features, independently of factors that are additionally related to the constructs[12,77].

There are several regression models to choose from for usage in machine learning approaches. With his theorem *No Free Lunch* Wolpert postulated that there is no general best machine learning algorithm for all predictive modeling problems such as classification and regression[78]. We chose the Random Forest Regressor as it has already demonstrated to predict executive functions in previous studies[70,79,80] and is commonly used[81,82]. Random Forest is an ensemble estimator that fits a number of decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and to control over-fitting. The decisions made by each tree carry equal weight, while the order of the decisions is random[83].

Following Poldrack et al[84]., accuracy was assessed by the coefficient of determination (R^2)[85], which measures how well the regression predictions approximate the real data points. It can be interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variables. R^2 ranges from 0 to 1, where 1 indicates that the regression model perfectly predicts the data. In cases of negative values, the mean of the data alone fits the results better than the predicted values. Thus, negative values mean a very poor generalisation of the model. For the cross-validation results, the mean of R^2 was calculated over 10 folds.

Secondly, the aim of our study was to investigate which of the many prosodic features were important in connection to all features to train the model successfully. For this purpose, the feature importance was calculated by the impurity-based feature importance of Random Forest, also known as the Gini importance[86,87]. When building a decision tree, features are selected at each node in order to divide the data into subsets that are as "pure" as possible with regard to the target variable. Gini Impurity measures how often a randomly chosen data point within a subset would be incorrectly labeled, reflecting the degree of disorder or „impurity" within the data. In contrast, Gini Importance assesses the overall decrease in node impurity resulting from splits based on a specific feature. It considers the probability of reaching each node and calculates the weighted reduction in impurity. Features with higher Gini importance are considered more important for predicting the target variable[86]. Feature importance was computed for the final estimator, as well as for each fold to estimate the variability of the importance. The sum of all feature importance scores adds up to 1.

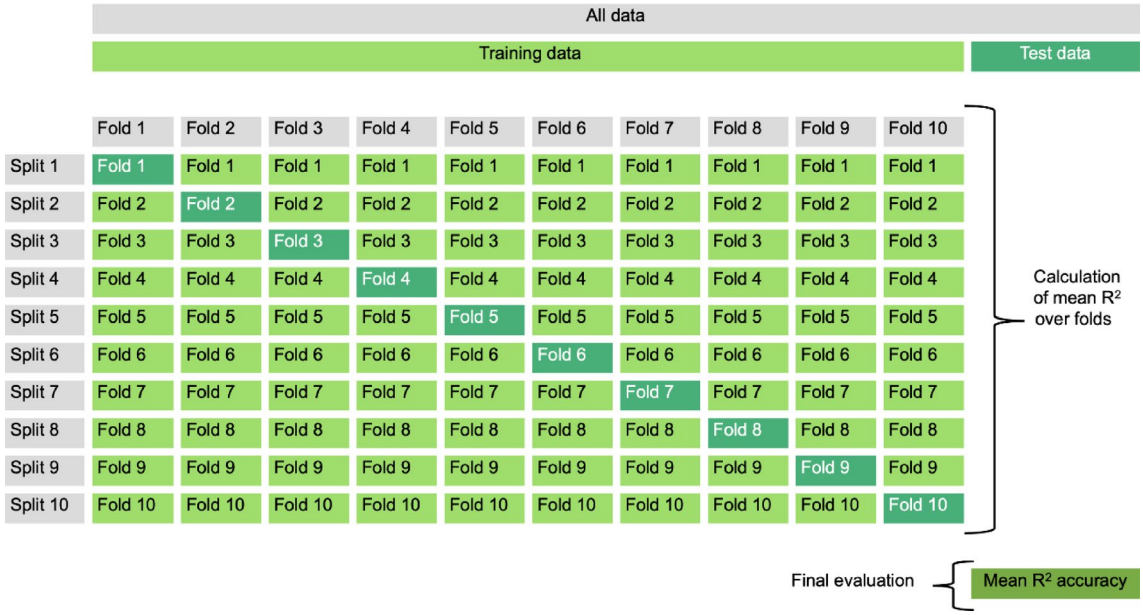

Fig. 1. 10-fold cross-validation design for each executive function target.

Thirdly, detailed analyses were conducted to examine the effects of confound removal and stratification. Here, we used other models such as Random Forest Regressor, ExtraTree Regressor, and Ridge Regression to regress out the confounds from the features in order to compare model performance depending on how the confounds were removed.

Moreover, we employed an approximate permutation test approach, suggested by North and colleagues[88], to disentangle predictive information of the features from that of the confounds. To achieve this, we permuted each feature separately. Here, the association between features and targets is randomised, while the association between confounds and targets remains unchanged. 10-fold cross-validation was performed for each permutation, and R^2 scores for 1000 permutations were used to construct an empirical null distribution, from which p-values were computed as the proportion of permuted R^2 scores greater than or equal to the R^2 score of the original non-permuted data. The threshold value for the two-tailed test was set to $p = 0.05$. Significant p-values indicate that predictive information stems from the features rather than the confounds alone.

Results

In cross-validation, the models were trained to predict each of the EF targets using all of the prosodic features. Regression of the confounding features sex, age, and education, and stratification by target distribution were performed. Evaluation was estimated using the coefficient of determination R^2 averaged over the 10 folds.

Out of 66 executive function targets, 53 variables did not show positive R^2 values, indicating no predictive power for these targets using our modeling approach. 13 executive function targets showed positive R^2 values (Fig. 2). However, only two targets, TMT BTA (processing time part A) and TMT BTB (processing time part B), showed R^2 values > 0.1, representing a reasonable model fit. The described TMT variables belong to the cognitive flexibility domain. An overview of R^2 of all 66 EF targets can be found in the supplements.

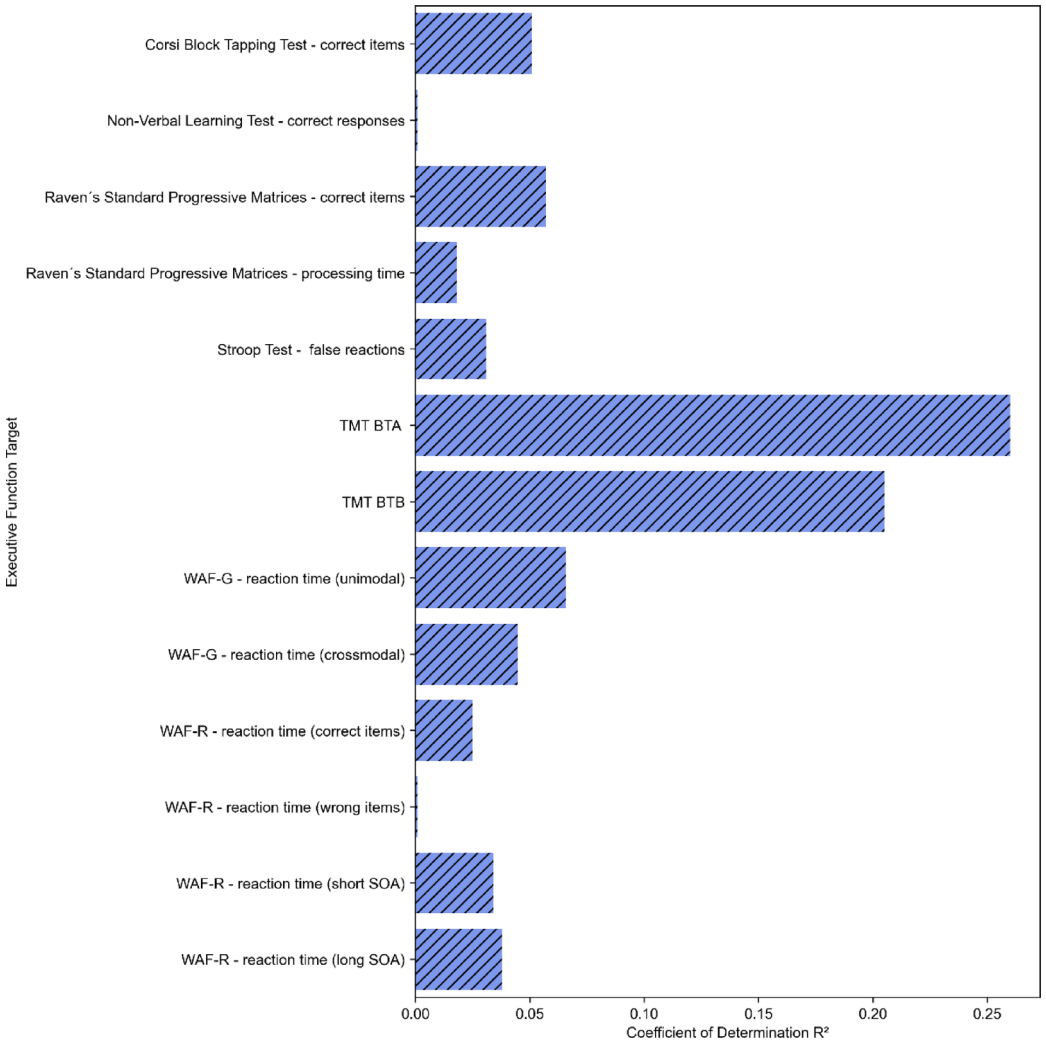

Fig. 2. Prediction of executive function targets by prosodic features. Cross-validation model with confound removal and stratified by target distribution. Only targets with positive R^2 values are displayed. TMT BTA = Trail Making Test - processing time part A, TMT BTB = Trail Making Test - processing time part B.

Feature importance was calculated in order to determine which of the prosodic features were particularly important for successfully predicting the EF targets. Since we observed good prediction performance ($R^2 > 0.1$) for TMT BTA and TMT BTB, we only computed feature importance for these targets. Figures 3 and 4 present the ten most important features predicting the EF targets TMT BTA and TMT BTB (see Appendix B for the feature importance of all prosodic variables). The majority of features identified as most important belong to the spectral prosodic domain. The most frequently appearing prosodic features were the Mel Frequency Cepstral Coefficients.

For the purpose of validation, we contrasted the effects of confound removal and stratification on the prediction performance for the targets TMT BTA and TMT BTB. To begin with, we compared the prediction results with the performance of the cross-validation model without regressing out the confounding variables sex, age, and education. These results indicated a worse prediction compared to the results with confound removal. Results are displayed in Fig. 5. For both TMT targets, prediction performance decreased when not removing the confounding variables. This is true for the stratified set up, as well as for the non-stratified set up. Prediction performance also decreases when not stratifying the cross-validation folds.

To explore the mechanism behind the decrease in prediction performance for the pipeline without confound removal further, and to examine whether it is related to the specific confound removal model used, we exchanged the standard confound removal model Linear Regression with other models, such as Random Forest Regressor, ExtraTree Regressor and Ridge Regression. As demonstrated in Fig. 6, the prediction performance varies depending on the choice of the confound removal model. The pipelines with the confound removal models Linear Regression and Ridge Regression indicate higher R^2 values than the pipelines with the confound removal models Random Forest Regressor and ExtraTree Regressor.

Finally, we evaluated the conditions with different confound removal models by using permutation tests. For the EF target TMT BTA with the cross-validation regressor Random Forest and the confound removal model Random Forest R^2 of 0.057 is significant ($p = 0.001$). For the EF target TMT BTB with the cross-validation

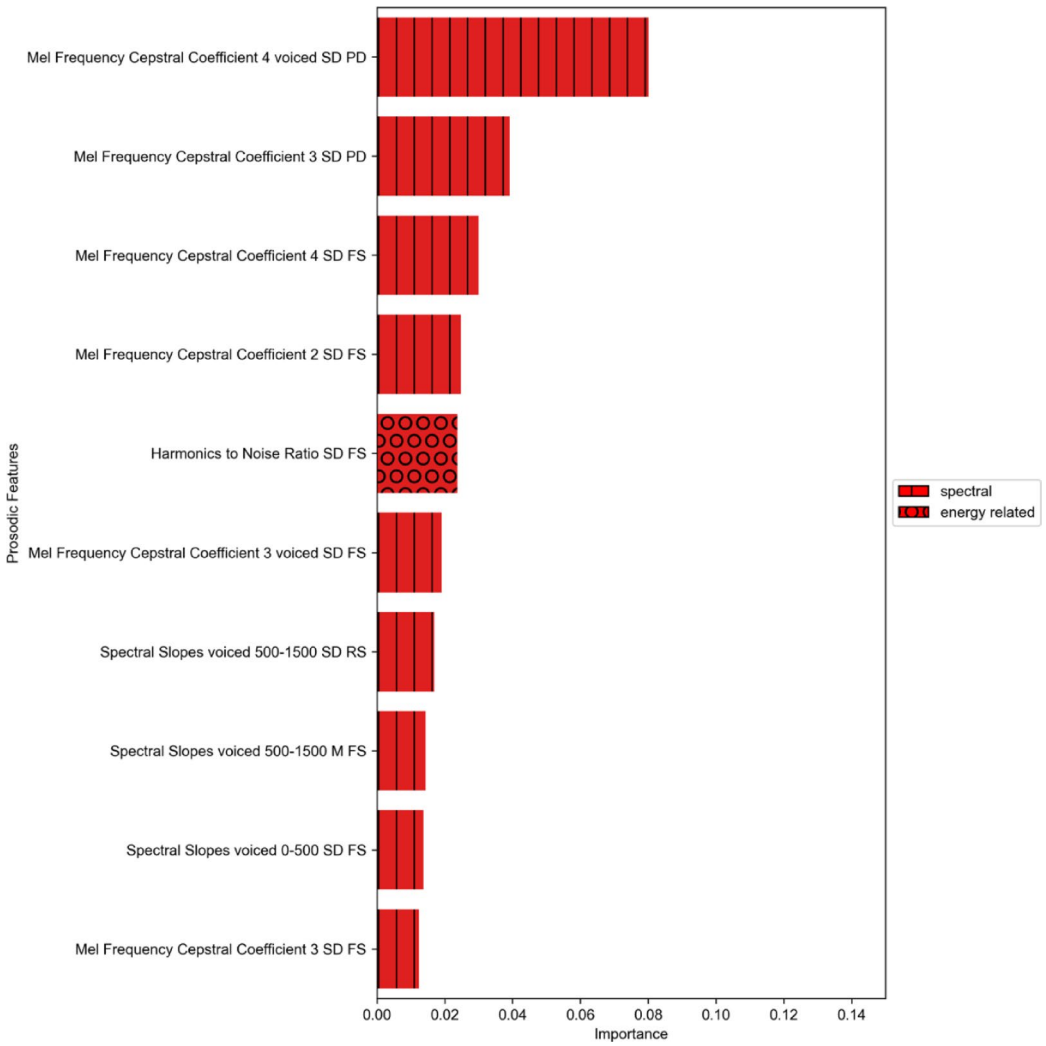

Fig. 3. Feature importance for TMT BTA. TMT BTA = Trail Making Test - processing time part A, SD = standard deviation, M = mean, PD = picture description, RS = retelling a story, FS = fictional storytelling.

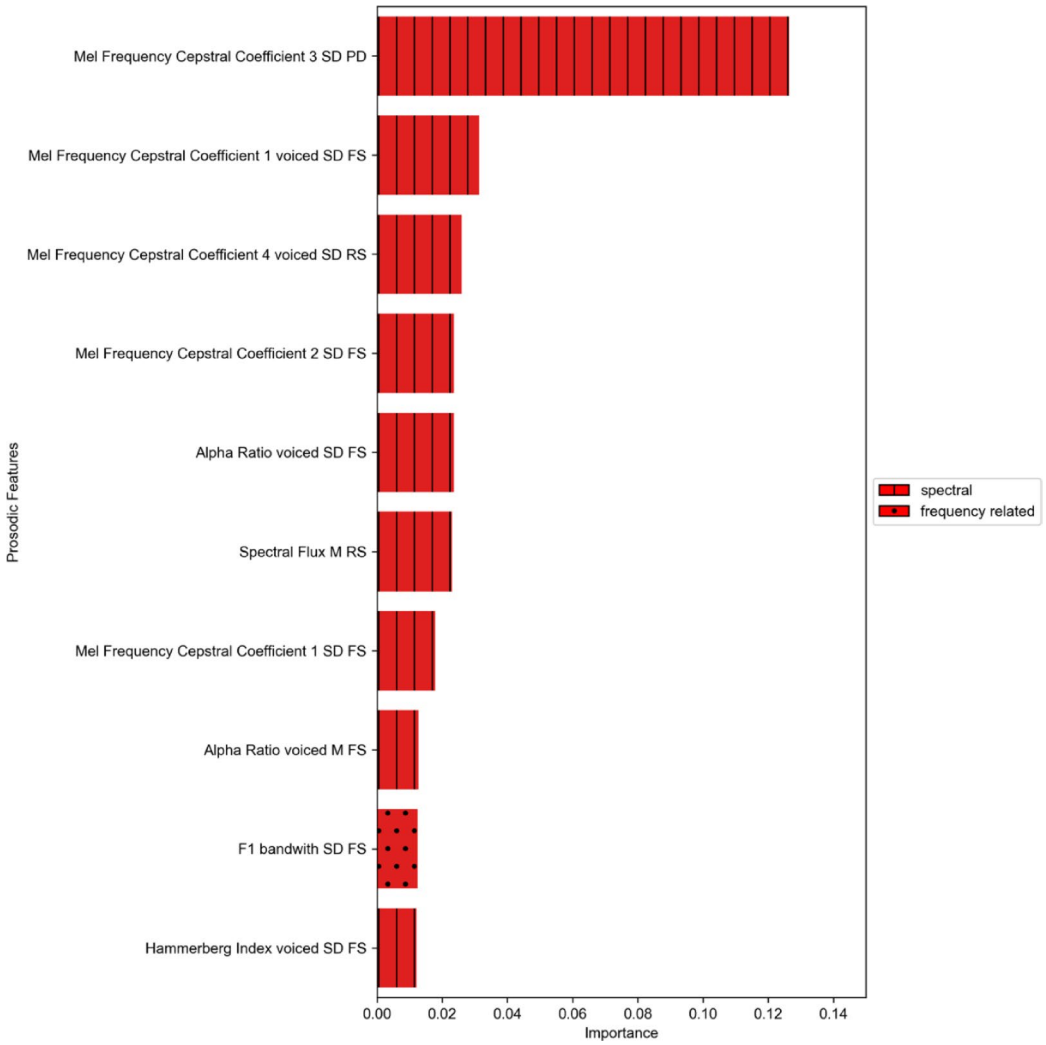

Fig. 4. Feature importance for TMT BTB. TMT BTB = Trail Making Test - processing time part B, SD = standard deviation, M = mean, PD = picture description, RS = retelling a story, FS = fictional storytelling.

regressor Random Forest and the confound removal model Ridge Regression R^2 of 0.196 is significant ($p = 0.032$) such as with the cross-validation regressor Random Forest and the confound removal model Linear Regression R^2 of 0.205 ($p = 0.017$). As shown in Table 3, all other positive prediction performances, measured by R^2 values, are not significant.

To summarise, we initially found a moderate predictive power of TMT BTA and TMT BTB by prosodic features. However, considering all results, there is a decrease in predictive power when not removing the confounding variables sex, age, and education, indicating confound leakage. In addition, the predictive power increases when stratification is performed. Pipelines with different models for removing confounding factors perform differently. Ultimately, two out of 20 models are significant, which suggests that the prediction is at least partly driven by the features in these models.

Discussion

This study is based on an investigation of the relationship between executive functions and prosody through examining whether prosodic features can predict executive functions. In summary, we preliminarily found a moderate predictive power of prosodic features for TMT BTA and TMT BTB. However, considering all results, there is a decrease in predictive power when not removing the confounding variables sex, age, and education, indicating confound leakage for most of the models.

Firstly, we evaluated 66 models, each predicting one executive function variable from the prosodic features. We employed 10-fold cross-validation with stratification by target variable and confound removal of sex, age, and education. The results showed poor or no prediction performance for 64 out of 66 EF targets.

Only the models for the TMT targets TMT BTA and TMT BTB, relating to cognitive flexibility, initially appeared to have a moderately valid predictive performance. Without the additional analyses that we conducted for validation, these results could be interpreted as follows: Our results would have confirmed findings from previous studies on a narrow correlation between executive functions and language in general[18,89], and would

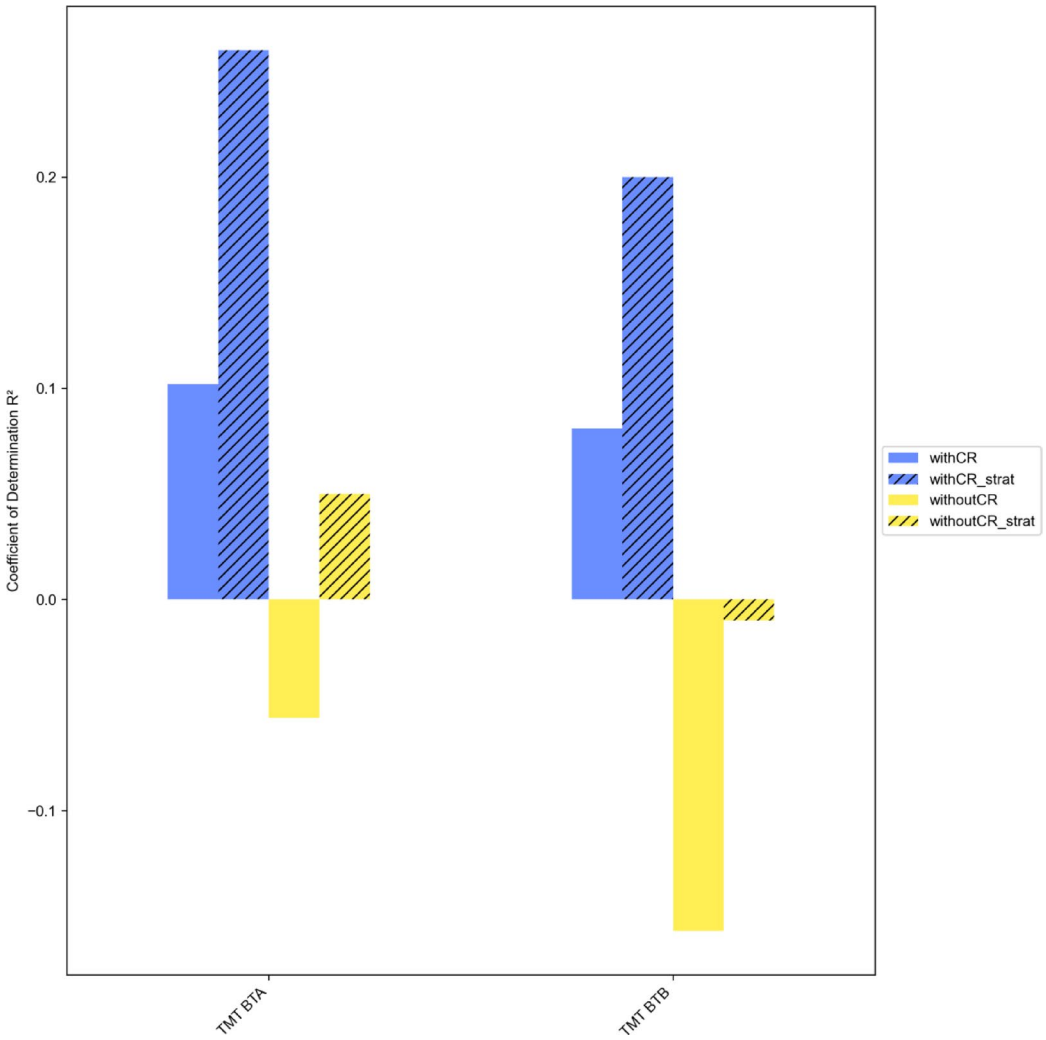

Fig. 5. Prediction of TMT targets in different conditions regarding confound removal and stratification. TMT BTA = Trail Making Test - processing time part A, TMT BTB = Trail Making Test - processing time part B, confounding variables (sex, age, education) and stratification: with CR = with confound removal, strat = stratified, without CR = without confound removal.

have been in line with research conducted in different patient cohorts[44,46,51], reporting connections between cognitive flexibility and prosody[35]. In our study, we would have found these associations in healthy participants. Based on these results, we would have concluded that the strong connection between TMT performance and prosody is likely the key factor driving the superior predictive accuracy observed in the TMT results. Both TMT performance and prosody processing share common cognitive mechanisms, particularly those related to attention, working memory, and cognitive flexibility. Both tasks require sustained and selective attention as well as attentional control: TMT for tracking targets and switching, prosody for detecting and setting vocal cues and phrase boundaries. TMT especially relies on working memory to keep track of sequences and rules, while prosody processing uses working memory to hold and integrate auditory information over time. TMT BTB measures cognitive flexibility and the ability to switch between tasks, which corresponds to the need to switch attention between different prosodic cues or emotional tones in speech. Additionally, both require rapid processing, TMT for visual-motor speed, prosody for timely receptive and productive communication[90–92].

Moreover, both the TMT and prosody processing share brain activation in the prefrontal cortex and parietal structures, meaning TMT performance primarily engages frontoparietal networks associated with executive control, and these same regions are also implicated in prosody processing, particularly in the context of cognitive-linguistic integration. Prosody processing such as coordinating tone, rhythm, and emotion in speech also engages the prefrontal cortex for high-level cognitive processes and executive control, as well as parietal regions for attention and processing of auditory information[93,94]. Additionally, the TMT's test design appears particularly sensitive to subtle individual differences, a characteristic that likely contributes to its superior predictive performance[95]. Previous research has also shown that TMT performance is most predictable from speech features derived from verbal fluency tasks[70]. Consistent with the literature, this study would have shown that features from various prosodic domains are important for the models to learn. This would have

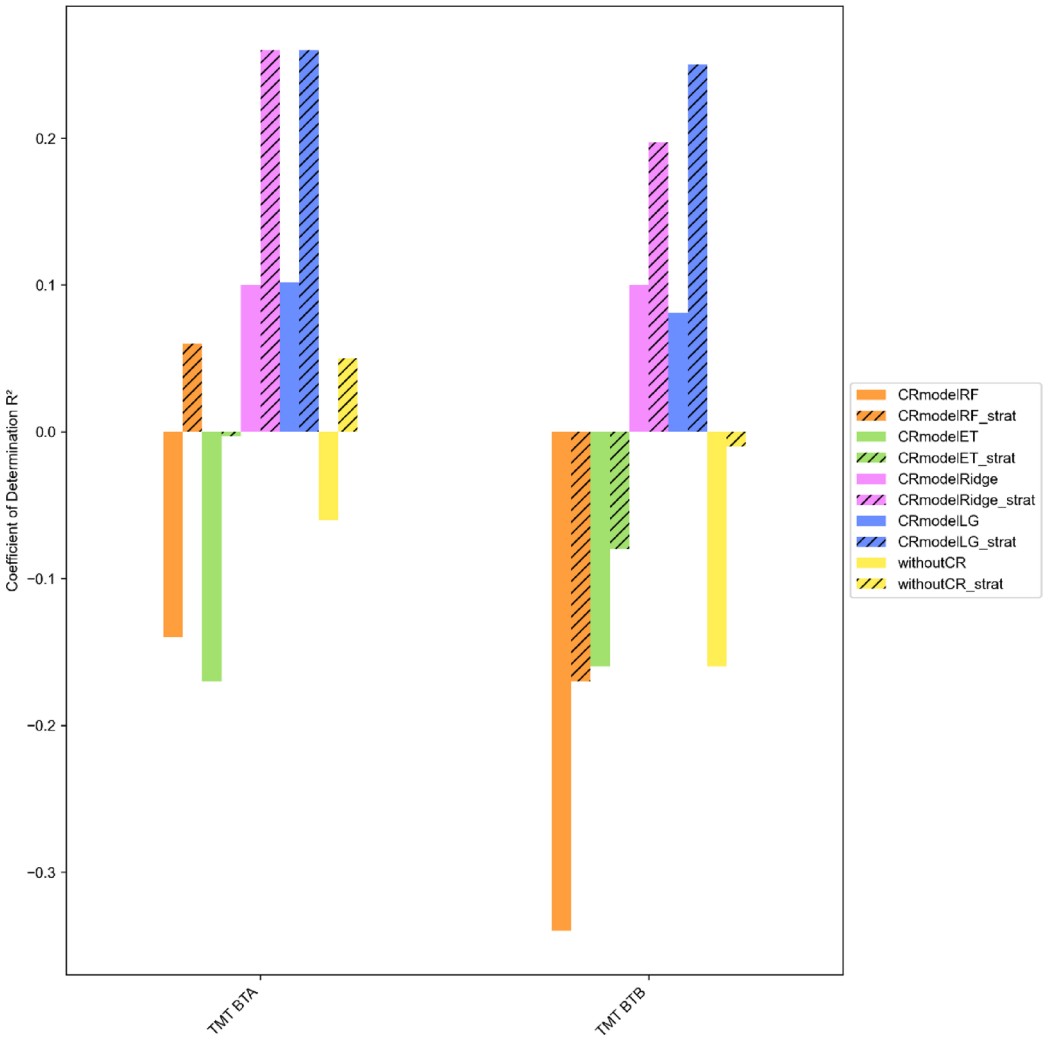

Fig. 6. Prediction of TMT targets in different conditions regarding different confound removal models. TMT BTA = Trail Making Test - processing time part A, TMT BTB = Trail Making Test - processing time part B, confounding variables (sex, age, education) and stratification. CRmodel = Confound removal model, RF = Random Forest Regressor, ET = ExtraTree Regressor, Ridge = Ridge Regression, LG = Linear Regression, withoutCR = without confound removal, strat = stratified.

TMT BTA			TMT BTB		
Condition	R^2	***p*-value**	**Condition**	R^2	***p*-value**
CRmodelRF	−0.142	0.009	CRmodelRF	−0.343	0.161
CRmodelRF_strat	0.057	0.001	CRmodelRF_strat	−0.171	0.069
CRmodelET	−0.172	0.001	CRmodelET	−0.156	0.001
CRmodelET_strat	−0.003	0.005	CRmodelET_strat	−0.082	0.001
CRmodelRidge	0.097	0.691	CRmodelRidge	0.106	0.058
CRmodelRidge_strat	0.262	0.188	CRmodelRidge_strat	0.196	0.032
CRmodelLG	0.102	0.633	CRmodelLG	0.081	0.162
CRmodelLG_strat	0.260	0.200	CRmodelLG_strat	0.205	0.017

Table 3. Comparison of different confound removal models complemented by the p-value. CRmodel = Confound removal model, RF = Random forest Regressor, ET = ExtraTree Regressor, LG = Linear Regression, Ridge = Ridge Regression, withoutCR = without confound removal, strat = stratified.

validated that prosodic features of different kinds are closely related to executive functions, as described in previous studies[96–98]. Furthermore, predominantly spectral prosodic parameters would have shown importance for the model fits, especially the Mel Frequency Cepstral Coefficients, which are already used as a biomarker in depressive disorders[48,50]. As described in Table 2, the Mel Frequency Cepstral Coefficients are defined as the perceived pitch of the frequency spectrum. More precisely, these are coefficients of the Mel scale, which relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear[99]. It therefore would have been deduced from the study that spectral parameters, in particular the Mel Frequency Cepstral Coefficients, are closely related to executive functions. Furthermore, the findings would have confirmed that easy-to-capture spontaneous speech derived from different tasks is suitable for the extraction of prosodic features. In summary, the present research would have raised the possibility that this predictive power of prosodic features could be an important biomarker for executive function impairment or its future decline.

However, given the additional in-depth analyses of the ML pipeline that partly invalidate the initial results, our findings need to be reinterpreted as follows:

We expect models to perform better if the effects of the confounding variables are not excluded, given that this would provide more information for the algorithm to learn. However, the prediction performance decreases for both TMT targets when not removing the confounding variables sex, age, and education. This is not in line with our expectation because in our scenario, the prediction performance should be worse if the confounding variables are removed, as the algorithm can then only learn from the association between confound-free features and the target. Despite the differences in prediction accuracy between the pipelines with and without confound removal being rather small, we deduce that information from these confounds, namely sex, age, and education leaked into the predictions through the confound removal procedure. The inadvertent injection of this information occurs particularly when the confounding variables and the targets show a strong correlation and this is coupled with the use of a high number of features, as explained by Hamdan et al.[13] and Sasse & Nicolaisen-Sobesky et al.[12]. This is indeed the case in our dataset (see Appendix C). There is a strong correlation between the TMT targets and the confounding variables. In addition, we use a high number of features within the cross-validation pipeline, because we wanted to investigate EF and prosody in an exploratory manner. While our dataset was relatively small compared to most ML studies, which typically increases the risk of leakage[100], it represents a reasonable size when compared to studies investigating speech biomarkers[34]. Prior work using larger samples also observed confound leakage[13], which suggests that this is a general issue and not merely a consequence of limited sample size. The results also confirm that these observations occur in both stratified and non-stratified conditions. As expected, it can be shown that stratification by target distribution generally increases the predictive performance. This is in line with Diamantidis et al[101]. and Hastie et al[15]., who show that equally representative cross-validation folds lead to improved predictive power. Additionally, it is demonstrated that stratification can also increase confound leakage. This can be derived from the fact that the difference in predictive power between the pipelines with and without confound removal is even greater in the stratified condition (Fig. 6). Furthermore, the results illustrate that the observed confound leakage is not bound to the use of Linear Regression as the confound removal model but also occurs when other models are employed.

Overall, these observations raise concerns about the trustworthiness of the primary results. Nonetheless, one cannot definitively rule out whether information from the features also influenced the predictive power of the present results. We, therefore, conducted permutation testing for the different cross-validation models. Since the permutation tests for the two TMT targets each identified models that can be interpreted as significant, we speculate that predictive power is partly due to the information contained in the features despite the confounding variables also contributing to the prediction. However, this was only observed in two of 66 EF targets and for these two targets only in specific confound removal models. For this reason, we only conditionally derive the predictive power of prosodic features. Further analyses of this type with other datasets would need to be carried out to verify this.

With this example, we aim to raise awareness that the influence of confounding factors in ML analyses, especially in the prediction of cognitive performance, must be rigorously addressed. The central message of our study is the need for careful quality control when handling potential confounds, as even subtle or unrecognised sources of confound leakage can unintentionally skew results, leading to misleading conclusions. Importantly, such distortions can occur even when standard ML procedures are applied correctly.

When confound leakage happens, information from confounding variables unintentionally leaks into the model, artificially inflating performance. This leads to overly high prediction accuracies[14]. The inappropriate control of confounds can be caused by different factors: On one hand, this can occur if confounding variables are inadvertently retained in the data despite attempts to remove them. This can arise in erroneous cross-validation applications[12]. On the other hand, this can also occur in a correctly implemented ML pipeline, specifically due to leakage stemming from continuous features that deviate from a normal distribution or from unbalanced features with limited precision[13]. In general, a strict separation of training and test set during cross-validation is mandatory, meaning that the confound removal models should be trained on the training data and then applied to both training and test data within each cross-validation fold, to prevent information leaking through[11,73,100]. In addition, we suggest to always compare results with and without confound adjustment as a standard routine. Moreover, analyses should be performed to clarify the relationship between possible confounds and the target variables. We further advise evaluating whether models trained on data with confound removal perform better than models trained on data with completely shuffled features.

By highlighting these methodological challenges, our goal is to encourage more rigorous handling of confounds in future ML-based cognitive research. Paying attention to these factors minimises the risk of confound leakage results, but does not guarantee correctness, as these points cannot claim to be exhaustive.

In conclusion, the present results highlight the pitfalls when conducting ML analyses with the aim of predicting variables of interest including cognitive performance. This example shows which misinterpretations could have been deduced from the initial results. This can be particularly dangerous if the findings match previous studies, as in the case here. This is crucial, as ML studies are becoming increasingly important and widely employed, especially with the accessibility of large amounts of data. In this respect, we caution and recommend that when using ML analyses to predict cognitive performance, quality controls should be performed to prevent false results. This is also true when interpreting ML results of other researchers. This study has contributed to uncovering more insight into a pitfall in ML analysis arising due to confound leakage. As confounding is ubiquitous in social and biological sciences, it should be further deciphered how confound leakage occurs and which contributing factors can be taken into account.

Data availability

Part of the data used in this study is publicly available upon request. Researchers who wish to acquire access to the data are kindly asked to contact Julia A. Camilleri at spexdata@fz-juelich.de, as described in the related publication Camilleri et al.[52].

Code availability

The code used in this study is publicly available at Github and archived with the DOI: https://doi.org/10.5281/zenodo.15301874.

Received: 15 July 2024; Accepted: 13 October 2025
Published online: 29 October 2025

References

1. Karako, K. Predictive deep learning models for cognitive risk using accessible data. *BioSci. Trends* **18**, 66–72 (2024).
2. Bzdok, D., Varoquaux, G. & Steyerberg, E. W. Prediction, not association, paves the road to precision medicine. *JAMA Psychiat.* **78**, 127–128 (2021).
3. CottaRamusino, M. et al. Diagnostic performance of molecular imaging methods in predicting the progression from mild cognitive impairment to dementia: an updated systematic review. *Eur. J Nucl. Med. Mol. Imaging* **51**, 1876–1890 (2024).
4. Roheger, M., Liebermann-Jordanidis, H., Krohm, F., Adams, A. & Kalbe, E. Prognostic factors and models for changes in cognitive performance after multi-domain cognitive training in healthy older adults: A systematic review. *Front. Hum. Neurosci.* **15**, 636355. https://doi.org/10.3389/fnhum.2021.636355 (2021).
5. Dwyer, D. B., Falkai, P. & Koutsouleris, N. Machine learning approaches for clinical psychology and psychiatry. *Annu. Rev. Clin. Psychol.* **14**, 91–118 (2018).
6. Arbabshirani, M. R., Plis, S., Sui, J. & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage* **145**, 137–165 (2017).
7. Rankin, D. et al. Identifying key predictors of cognitive dysfunction in older people using supervised machine learning techniques: observational study. *JMIR Med. Inform.* **8**, 20995. https://doi.org/10.2196/20995 (2020).
8. Ansart, M. et al. Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. *Med. Image Anal.* **67**, 101848. https://doi.org/10.1016/j.media.2020.101848 (2021).
9. Ahmad, S., El-Affendi, M. A., Anwar, M. S. & Iqbal, R. Potential future directions in optimization of students' performance prediction system. *Comput. Intell. Neurosci.* **1**, 6864955. https://doi.org/10.1155/2022/6864955 (2022).
10. Domingos, P. A few useful things to know about machine learning. *Comm. ACM.* **55**, 78–87 (2012).
11. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804. https://doi.org/10.1016/j.patter.2023.100804 (2023).
12. Sasse, L., & Nicolaisen-Sobesky, E. On Leakage in Machine Learning Pipelines. *arXiv* preprint arXiv: 2311.04179, (2024).
13. Hamdan, S. et al. Confound-leakage: confound removal in machine learning leads to leakage. *GigaScience* **12**, giad071. https://doi.org/10.1093/gigascience/giad071 (2023).
14. Spisak, T. Statistical quantification of confounding bias in machine learning models. *GigaScience*, **11**, giac082. https://doi.org/10.1093/gigascience/giac082 (2022).
15. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 1–758 (Springer, 2009).
16. Ardila, A. The executive functions in language and communication. In *Cognition and acquired language disorders* (ed. Peach, R. K. & Shapiro, L. P.) 147–166 (Mosby, 2012).
17. Baddeley, A. Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* **4**, 829–839 (2003).
18. Levelt, W. J. Accessing words in speech production: Stages, processes, and representations. *Cogn.* **42**, 1–22 (1992).
19. Goldstein, S., Naglieri, J. A., Princiotta, D., & Otero, T. M. Introduction: A history of executive functioning as a theoretical and clinical construct. In *Handbook of Executive Functioning*. (ed. Goldstein, S. & Naglieri, J. A.) 3–12 (Springer Science, 2014).
20. Ward, J. *The Student's Guide to Cognitive Neuroscience*. (Psychology Press, 2015).
21. Friedman, N. et al. Individual differences in executive functions are almost entirely genetic in origin. *J. Exper. Psychol.* **137**, 201–225 (2008).
22. Diamond, A. Executive functions. *An. Rev. Psy.* **64**, 135–168 (2013).
23. Miyake, A. et al. The unity and diversity of executive functions and their contributions to complex 'Frontal Lobe' tasks: a latent variable analysis. *Cogn. Psychol.* **41**, 49–100 (2000).
24. Löffler, C., Frischkorn, G. T., Hagemann, D., Sadus, K. & Schubert, A. L. The common factor of executive functions measures nothing but speed of information uptake. *Psychol. Res.* **88**, 1092–1114 (2024).
25. Barch, D. M. The cognitive neuroscience of schizophrenia. *Annu. Rev. Clin. Psychol.* **1**, 321–353 (2005).
26. Guarino, A. et al. Executive functions in Alzheimer disease: a systematic review. *Front. Neurosci.* **10**, 437 (2019).
27. Kudlicka, A., Clare, L. & Hindle, J. V. Executive functions in Parkinson's disease: Systematic review and meta-analysis. *Mov. Disord.* **26**, 2305–2315 (2011).
28. Nigg, J. T., Blaskey, L. G., Huang-pollock, C. L., & Rappley, M. D. Neuropsychological Executive Functions and DSM-IV ADHD. Subtypes. *J. Am. Acad. Child Adolesc. Psych.* **41,** 59–66 (2002).
29. Tavares, J. V. T. et al. Distinct profiles of neurocognitive function in unmedicated unipolar depression and bipolar II depression. *Biol. Psychol.* **62**, 917–924 (2007).
30. Salthouse, T., Atkinson, T. & Berish, D. Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *J. Exper. Psychol.* **132**, 566–594 (2003).

31. Novick, J. M., Trueswell, J. C. & Thompson, S. L. Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cogn Affec. Behav. Neurosci.* **5**, 263–281 (2005).
32. Laver, J. *Principles of Phonetics*. (Cambridge University Press, 1994).
33. Eyben, F. et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *Transac. Affect. Com.* **7**, 190–202 (2015).
34. Hecker, P., Steckhan, N., Eyben, F., Schuller, B. W. & Arnrich, B. Voice analysis for neurological disorder recognition – A systematic review and perspective on emerging trends. *Front. Digit. Health* **4**, 842301. https://doi.org/10.3389/fdgth.2022.842301 (2022).
35. Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F. & Green, J. R. Speech as a biomarker: Opportunities, interpretability, and challenges. *Perspect. ASHA Spec. Interest Groups* **7**, 276–283 (2022).
36. Robin, J. et al. Evaluation of speech-based digital biomarkers: review and recommendations. *Digl Biomark.* **4**, 99–108 (2020).
37. Martínez-Sánchez, F., Meilán, J. J. G., Carro, J. & Ivanova, O. A prototype for the voice analysis diagnosis of Alzheimer's disease. *J. Alzheimers. Dis.* **64**, 473–481 (2018).
38. Parola, A., Simonsen, A., Bliksted, V. & Fusaroli, R. Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. *Schizo. Res.* **216**, 24–40 (2020).
39. Speer, S. R. & Ito, K. Prosody in first language acquisition–Acquiring intonation as a tool to organize information in conversation. *Lang Ling Com.* **3**, 90–110 (2009).
40. Alexander, M. P., Benson, D. F. & Stuss, D. T. Frontal lobes and language. *Brain & Lang.* **37**, 656–691 (1989).
41. Ross, E. D. The aprosodias: Functional-anatomical organization of the affective components of language in the right hemisphere. *Arch. Neurol.* **140**, 695–710 (1981).
42. Keulen, S. et al. Psychogenic foreign accent syndrome: a new case. *Front. Neurosci.* **10**, 143 (2016).
43. Roy, A., Allain, P., Roulin, J. L., Fournet, N. & Le Gall, D. Ecological approach of executive functions using the behavioural assessment of the dysexecutive syndrome for children (BADS-C): Developmental and validity study. *J. Neuropsych.* **37**, 956–971 (2015).
44. Breitenstein, C., Van Lancker, D., Daum, I. & Waters, C. H. Impaired perception of vocal emotions in Parkinson's disease: influence of speech time processing and executive functioning. *Brain & Cogn.* **45**, 277–314 (2001).
45. Nevler, N. et al. Automatic measurement of prosody in behavioral variant FTD. *Neurol.* **89**, 650–656 (2017).
46. Filipe, M. G., Frota, S. & Vicente, S. G. Executive functions and prosodic abilities in children with high-functioning autism. *Front. Psych.* **9**, 359 (2018).
47. Alghowinem, S., Gedeon, T., Goecke, R., Cohn, J. F. & Parker, G. Interpretation of depression detection models via feature selection methods. *IEEE Trans. Affect. Comput.* **14**, 133–152 (2020).
48. Cummins, N., Epps, J., Sethu, V., Breakspear, M. & Goecke, R. Modeling spectral variability for the classification of depressed speech. In *Proc. Interspeech*. 857–861 (2013).
49. Moore, I. I. E., Clements, M. A., Peifer, J. W. & Weisser, L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *Transact. Biomedic.* **55**, 96–107 (2007).
50. Williamson, J. R. et al. Vocal biomarkers of depression based on motor incoordination. *Proc. Aud.* **3**, 41–48 (2013).
51. Engelhardt, P. E., Nigg, J. T. & Ferreira, F. Is the fluency of language outputs related to individual differences in intelligence and executive function?. *Acta Psychol.* **144**, 424–432 (2013).
52. Camilleri, J. A. et al. SpEx: a German-language dataset of speech and executive function performance. *Sci. Rep.* **14**, 9431. https://doi.org/10.1038/s41598-024-58617-3 (2024).
53. *Wiener Testsystem*. (SCHUHFRIED GmbH, 2016).
54. Stoet, G. PsyToolkit: A software package for programming psychological experiments using Linux. *Behav. Res. Methods* **42**, 1096–1104 (2010).
55. Reitan, R. M. Validity of the trail making test as an indicator of organic brain damage. *Percept. Mot. Skills* **8**, 271–276 (1958).
56. Raven, J. C., Raven, J. & Court, J. H. *SPM Manual (Deutsche Bearbeitung und Normierung von St. Bulheller und H. Häcker)*. (Swets & Zeitlinger B.V, 1998).
57. Grant, D. A. & Berg, E. A. A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *J. Exp. Psychol.* **38**, 404–411 (1948).
58. Kaller, C. P., Unterrainer, J. M. & Stahl, C. Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. *Psychol. Assess.* **24**, 46–53 (2012).
59. Meiran, N. Reconfiguration of processing mode to task performance. *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 1423–1442 (1996).
60. Schellig, D., Schuri, U. & Arendasy, M. *NBN-NBACK-nonverbal*. (SCHUHFRIED GmbH, 2009).
61. Sturm, W. & Willmes, K. *NVLT Non-Verbal Learning Test*. (SCHUHFRIED GmbH, 2016).
62. Schellig, D. & Hättig, H. A. Die Bestimmung der visuellen Merkspanne mit dem Block-Board. *Z. Neuropsychol.* **4**, 104–112 (1993).
63. Kaiser, S., Aschenbrenner, S., Pfüller, U., Roesch-Ely, D., & Weisbrod, M. *Response Inhibition*. (SCHUHFRIED GmbH, 2016).
64. Simon, J. R. & Wolf, J. D. Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics* **6**, 99–105 (1963).
65. Schuhfried, G. *Interferenz nach Stroop*. (SCHUHFRIED GmbH, 2016).
66. Sturm, W. *Wahrnehmungs- und Aufmerksamkeitsfunktionen: Geteilte Aufmerksamkeiten*. (SCHUHFRIED GmbH, 2016).
67. Mackworth, N. H. The breakdown of vigilance during prolonged visual search. *J. Exper. Psych.* **1**, 6–21 (1948).
68. Goodglass, H., & Kaplan, E. *The Assessment of Aphasia and Related Disorders*. (Lea & Febiger, 1972).
69. Amunts, J., Camilleri, J. A., Eickhoff, S. B., Heim, S. & Weis, S. Executive functions predict verbal fluency scores in healthy participants. *Sci. Rep.* **10**, 1–11 (2020).
70. Amunts, J. et al. Comprehensive verbal fluency features predict executive function performance. *Sci. Rep.* **11**, 1–14 (2021).
71. Eyben, F., Wöllmer, M. & Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. *Proc. Multimed.* **18**, 1459–1462 (2010).
72. Van Rossum, G., & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, 2009).
73. Hamdan, S. et al. Julearn: An easy-to-use library for leakage-free evaluation and inspection of ML models. *Gigabyte*, gigabyte 113 https://doi.org/10.46471/gigabyte.113 (2024).
74. Molinaro, A. M., Simon, R. & Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinform.* **21**, 3301–3307 (2005).
75. Dromey, C., Silveira, J. & Sandor, P. Recognition of affective prosody by speakers of English as a first or foreign language. *Speech Comm.* **47**, 351–359 (2005).
76. Volin, J., Tykalová, T., & Boril, T. Stability of prosodic characteristics across age and gender groups. *Inter Speech* 3902–3906 (2017).
77. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **6**, 1–21 (2012).
78. Wolpert, D. H. The lack of a priori distinctions between learning algorithms. *Neural. Com.* **8**, 1341–1390 (1996).
79. Byeon, H. Is the Random Forest algorithm suitable for predicting Parkinson's disease with mild cognitive impairment out of Parkinson's disease with normal cognition?. *Int. J. Enviro.* **17**, 2594 (2020).

80. Cordova, M. et al. Heterogeneity of executive function revealed by a functional random forest approach across ADHD and ASD. *Neuroimage Clin.* **26**, 102245. https://doi.org/10.1016/j.nicl.2020.102245 (2020).
81. Adnan, M. N., Ip, R. H., Bewong, M. & Islam, M. Z. BDF: A new decision forest algorithm. *Inform. Sci.* **569**, 687–705 (2021).
82. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data?. *Adv. Neural Inf. Process. Syst.* **35**, 507–520 (2022).
83. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
84. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiat.* **77**, 534–540 (2020).
85. Wright, S. Correlation and causation. *J. Agric.* **20**, 557–585 (1921).
86. Nembrini, S., König, I. R. & Wright, M. N. The revival of the Gini importance?. *Bioinformatics* **34**, 3711–3718 (2018).
87. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **2**, 2825–2830 (2011).
88. North, B. V., Curtis, D. & Sham, P. C. A note on the calculation of empirical P values from Monte Carlo procedures. *Am. J. Hum. Genet.* **71**, 439–441 (2002).
89. Baddeley, A. D. & Hitch, G. Working memory. *Psych. of learn & motiv.* **8**, 47–89 (1974).
90. LaCroix, A. N., Blumenstein, N., Tully, M., Baxter, L. C. & Rogalsky, C. Effects of prosody on the cognitive and neural resources supporting sentence comprehension: A behavioral and lesion-symptom mapping study. *Brain Lang.* **203**, 104756 (2020).
91. Salthouse, T. A. What cognitive abilities are involved in Trail Making performance?. *Intell.* **39**, 222–232 (2011).
92. MacPherson, S. E., Allerhand, M., Cox, S. R. & Deary, I. J. Individual differences in cognitive processes underlying Trail Making Test-B performance in old age: The Lothian Birth Cohort 1936. *Intell.* **75**, 23–32 (2019).
93. Karimpoor, M. et al. Tablet-based functional MRI of the trail making test: effect of tablet interaction mode. *Front. Hum. Neurosci.* **11**, 496 (2017).
94. Kreitewolf, J., Friederici, A. D. & von Kriegstein, K. Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. *Neuroimage* **102**, 332–344 (2014).
95. Sánchez-Cubillo, I. et al. Construct validity of the Trail Making Test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *J. Int. Neuropsychol. Soc.* **15**, 438–450 (2009).
96. Yap, P. et al. Development trends of white matter connectivity in the first years of life. *PLoS ONE* **6**, e24678. https://doi.org/10.1371/journal.pone.0024678 (2011).
97. Tamarit, L., Goudbeek, M., & Scherer, K. R. Spectral slope measurements in emotionally expressive speech. In *Proc. of Speech*. Vol. 7, 169–183 (2008).
98. Le, P., Ambikairajah, E., Epps, J., Sethu, V. & Choi, E. H. C. Investigation of spectral centroid features for cognitive load classification. *Speech Comm.* **54**, 540–551 (2011).
99. Hasan, M. R., Jamil, M., & Rahman, M. G. R. M. S. Speaker identification using mel frequency cepstral coefficients. *Variat.* **1**, 565–568 (2004).
100. Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S. & Scheinost, D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat. Comm.* **15**, 1829 (2024).
101. Diamantidis, N. A., Karlis, D. & Giakoumakis, E. A. Unsupervised stratification of cross-validation for accuracy estimation. *Artif. Intell.* 1–16 (2000).

Acknowledgements

This study was supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/2–1, EI 816/16 – 1 and EI 816/21 – 1), the National Institute of Mental Health (R01-MH074457), the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain", the Virtual Brain Cloud (EU H2020, no. 826421) & the National Institute on Aging (R01AG067103).

Author contributions

G.K., J.A.C., S.W. conceived the project and designed the study. S.H., S.H., S.B.E., K.R.P. contributed essential resources. G.K. with contributions from S.W. and all other authors wrote the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-24325-9.

Correspondence and requests for materials should be addressed to G.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.