UC San Diego
UC San Diego Previously Published Works

Title

Development of parallel forms of a brief smell identification test useful for longitudinal testing.

Permalink

https://escholarship.org/uc/item/0zd2d4vh

Authors

Moein, Shima
Sacan, Ahmet
Pourrezaei, Kambiz
et al.

Publication Date

2023-03-24

DOI

10.3758/s13428-023-02102-8

Peer reviewed

Development of parallel forms of a brief smell identification test useful for longitudinal testing

Shima T. Moein[1,2] · Ahmet Sacan[3] · Kambiz Pourrezaei[3] · Carol H. Yan[4] · Justin H. Turner[5] · Ryan Sharetts[2] · Richard L. Doty[1]

Abstract

Although there are numerous brief odor identification tests available for quantifying the ability to smell, none are available in multiple parallel forms that can be longitudinally administered without potential confounding from knowledge of prior test items. Moreover, empirical algorithms for establishing optimal test lengths have not been generally applied. In this study, we employed and compared eight machine learning algorithms to develop a set of four brief parallel smell tests employing items from the University of Pennsylvania Smell Identification Test that optimally differentiated 100 COVID-19 patients from 132 healthy controls. Among the algorithms, linear discriminant analysis (LDA) achieved the best overall performance. The minimum number of odorant test items needed to differentiate smell loss accurately was identified as eight. We validated the sensitivity of the four developed tests, whose means and variances did not differ from one another (Bradley–Blackwood test), by sequential testing an independent group of 32 subjects that included persons with smell dysfunction not due to COVID-19. These eight-item tests clearly differentiated the olfactory compromised subjects from normosmics, with areas under the ROC curve ranging from 0.79 to 0.83. Each test was correlated with the overall UPSIT scores from which they were derived. These brief smell tests can be used separately or sequentially over multiple days in a variety of contexts where longitudinal olfactory testing is needed.

Keywords UPSIT · Olfactory test · Longitudinal testing · Hyposmia · Anosmia

Introduction

The sense of smell is largely underappreciated until it becomes dysfunctional. Unlike vision, hearing, balance, and touch, this sense is rarely quantitatively assessed clinically,

✉ Shima T. Moein
shima.moein@pennmedicine.upenn.edu

1 Department of Otorhinolaryngology-Head and Neck Surgery, Perelman School of Medicine, Smell & Taste Center, University of Pennsylvania, Philadelphia, PA, USA

2 Research & Development Division, Sensonics International, 411 S Black Horse Pike, Haddon Heights, NJ 08035, USA

3 School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia, PA, USA

4 Department of Surgery, Division of Otolaryngology-Head and Neck Surgery, University of California San Diego, San Diego, CA, USA

5 Department of Otolaryngology-Head and Neck Surgery, Vanderbilt University Medical Center, Nashville, TN, USA

even though hundreds of thousands of persons seek medical help annually for problems with this primary sense. This has been markedly accentuated by the COVID-19 pandemic (Gerkin et al., 2020). Loss or distortion of the ability to smell impacts quality of life, including the flavor of foods and beverages, and impacts patient safety by compromising the ability to detect fire, leaking natural gas, spoiled food, and other environmental dangers (Devanand et al., 2014). It is now well established that olfactory dysfunction can be a harbinger for Alzheimer's and other neurodegenerative diseases and, in elderly populations, can triple the likelihood of mortality over the course of a half-decade (Liu et al., 2019; Devanand et al., 2014). Without quantitative testing, patients are often unaware of an olfactory deficit (Doty et al., 1987). Such testing is critical to accurately establish the veracity of a patient's complaint, the probability of malingering, the impact of treatments, and to what degree a patient's function is normal for his or her age and sex.

Although there are numerous brief odor identification tests available for quantifying the ability to smell, in many

cases their efficacy has not been evaluated and they lack parallel forms to make longitudinal testing possible without potential confounding from knowledge of prior test items. Despite the fact that tests with larger numbers of items do not appear to meaningfully suffer from these problems (Doty et al., 2021), out of practicality many clinicians and researchers prefer to use briefer tests. Unfortunately, in addition to lacking parallel forms, the empirical rationale for establishing the length of such tests is lacking.

The present study employed state-of-the-art machine learning techniques to generate a series of parallel brief odor identification tests of lengths which optimize practicality and both sensitivity and specificity in differentiating persons with abnormal smell function from those with normal smell function. Machine learning techniques take advantage of automated analytical algorithms and combinatorial search methods to maximize their predictive power. The training was focused on test items from tests administered to persons with an intact sense of smell and those with smell loss secondary to SARS-CoV-2 infections, a surrogate for virus-related smell loss.

Method

Experimental design

To achieve the goal of the study, we first determined the optimal number of odorant test items needed to accurately differentiate patients with a disorder known to have significant smell loss from healthy controls. The data to which the machine learning algorithms were applied came from smell tests that had been administered to 100 confirmed COVID-19 patients (age, 45.4 ±11.8, 67 males) and to 132 healthy subjects (age, 43.7 ±15.4, 48 males) tested prior to the outbreak of COVID-19 in Iran (Moein, Hashemian, Mansourafshar, et al., 2020a; Moein, Hashemian, Tabarsi, & Doty, 2020b).

The Persian version of the 40-item University of Pennsylvania Smell Identification Test (UPSIT®; Sensonics® International, Haddon Heights, NJ) was administered to all subjects. This version includes 40 odors out of a library of 51 odors that are used in different cultural versions of UPSIT®. The UPSIT® focuses on the comparative ability of subjects to identify odorants at the suprathreshold level (Doty et al., 1984). The subject releases each odorant by scraping an odorized strip with a pencil tip. He or she then indicates which of four written response alternatives corresponds to the perceived smell. A response must be made even if no smell is experienced or the smell seems different from those listed as the response alternatives (i.e., the test is forced-choice).

The dataset contained the UPSIT® test items that were correctly identified, along with subject demographic information, i.e., their age, gender, and educational level. The summary of the subjects' responses to the UPSIT® items is shown in Table 1. Note that while the names of the odors are indicated, there are response alternatives associated with each odorant. Thus, in this table, as well as elsewhere in the text, the name of an odorant actually reflects an UPSIT® item, which includes the odorant and associated response alternatives. All attributes, except for age, coconut, and pizza were significantly different between COVID-19-positive and healthy subjects (p value < 0.01) based on Student's t test for continuous attributes and Fisher's exact test for binary-valued attributes.

To examine the generalizability of the test scores beyond the initial COVID-19 cohort, an independent group of 32 validation subjects [mean (SD) age: 55.56 (16.52), 12 males] with or without smell loss, as determined from previous

Table 1 Odorants detection rate

Odorant	COVID-19	Healthy	Odorant	COVID-19	Healthy
Grass	26.0%	82.6%	Popcorn	27.0%	57.6%
Lemon	28.0%	84.1%	Daffodil	59.0%	88.6%
Apple	43.0%	93.9%	Rubber tire	55.0%	84.1%
Cheese	35.0%	77.3%	Peach	63.0%	91.7%
Winter-green	54.0%	95.5%	Peanut	62.0%	90.2%
Motor oil	27.0%	68.2%	Raspberry	59.0%	87.1%
Turpentine	48.0%	88.6%	Cinnamon	65.0%	91.7%
Pineapple	49.0%	89.4%	Clove	63.0%	89.4%
Garlic	56.0%	95.5%	Fish	38.0%	62.9%
Smoke	54.0%	93.2%	Ginger-bread	69.0%	92.4%
Grape	20.0%	58.3%	Chocolate	69.0%	92.4%
Mint	58.0%	95.5%	Bubble gum	66.0%	88.6%
Water-melon	58.0%	95.5%	Indian spice	64.0%	85.6%
Strawberry	50.0%	87.1%	Rose	54.0%	74.2%
Magnolia	54.0%	89.4%	Lilac	67.0%	87.1%
Natural gas	62.0%	97.0%	Onion	82.0%	100.0%
Menthol	48.0%	82.6%	Coconut	76.0%	90.2%
Paint thin-ner	53.0%	87.1%	Baby powder	86.0%	97.7%
Jasmine	50.0%	84.1%	Soap	86.0%	96.2%
Leather	55.0%	88.6%	Pizza	50.0%	58.3%

Correct response rates to odorant items are given separately for patients who tested positive for COVID-19, and for healthy subjects. For each odorant item, the percentage of the subjects that correctly identified it is shown. Odorant items are sorted by the difference in these percentages between positive and negative groups. All attributes, except for age, coconut, and pizza were significantly different between COVID-19-positive and healthy subjects (p value < 0.01)

UPSIT scores (Doty, 1995), were administered the brief tests that we developed (see Table S2). Sixty-seven subjects for whom such scores were available were contacted for the testing through phone calls or email messages; 45 agreed to participate and received the test kit via mail. Thirty-two completed test kits were returned to the researchers within the study time frame and their data were included in the analysis. Those with compromised smell function (i.e., UPSIT® scores < 35 for women and < 34 for men) had etiologies due to traumatic brain injury (two subjects), upper respiratory viral infection (one subject), nasal surgeries (two subjects), and unknown causes (21 subjects). Each subject was paid $20 for participating in the validation study.

Statistical analysis

The binary UPSIT® item response data were used to train and test popular machine learning algorithms available in MATLAB® version 2020a (MATLAB®, 2020), including logistic regression (Grimm & Yarnold, 1995), artificial neural networks (using ten hidden elements) (Haykin, 1998), decision trees (Breiman et al., 1984), k-nearest neighbor (kNN, $k = 3$ with city block distance metric), and the ensemble learning methods of random forests (Breiman, 2001), AdaBoost (short for adaptive boosting) (Freund & Schapire, 1999), and support vector machines (SVM) (Hearst et al., 1998). Parameter sweep analysis for number of hidden elements in neural networks and the number of nearest neighbors in the kNN method are provided in the supplementary Figs. S1 and S2. To compare the results with common smell test scoring approaches, a simple linear discriminant analysis (LDA) classifier, based on the total number of correctly identified odorant items, was also used.

Machine learning and feature selection

To reduce the data dimensionality, remove noise, and optimize the predictive performance of each method, feature selection was performed to select an optimal subset of odorant items. We used a sequential forward feature selection strategy where the selected odorant set was repeatedly extended so long as the inclusion of a new odorant into the selected set improved the cross-validation performance. As a trade-off between computing time and completeness of the search space, a modified sequential selection strategy was implemented such that at each selection iteration all combinations of 2–3 odorants were considered for inclusion. The performance of a combination of odorants was assessed until the optimized model was reached for each machine learning method.

A leave-one-out cross-validation procedure was applied to the data to assess the performance of each machine learning method. At each iteration, one subject was left out as a test sample; the machine learning model was trained on the remaining samples, which was then used to make a prediction for the left-out subject. This procedure was executed for each subject in the dataset at each feature selection iteration. The performance metrics we assessed included accuracy (ratio of correct predictions over all predictions), sensitivity (true-positive rate), specificity (true-negative rate), and the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, a graph displaying the true-positive rate against the false-positive rate. For some machine learning methods (e.g., SVM, random forest, and AdaBoost), several different combinations of sensitivity and specificities are possible depending on where the threshold for categorization is set. Therefore, in addition to the combination of optimized sensitivity and specificity, the threshold-independent AUC was also reported. The optimization of feature selection continued until no further improvement in performance could be achieved.

To provide tests useful for practical serial testing of smell function, an optimization search strategy for multiple sets was performed. The number of odorant items for these sets was decided by considering both practicality and having sufficient predictive performance based on the item selection results. The rationale was that, for example, four relatively brief tests could be administered sequentially, with a repeated administration of one of these tests for assessment of temporal reproducibility. Such tests could also be administered in different orders and at varied intervals, such as every day or every other day over the course of a week. In order to ensure heterogeneity among the tests, we made certain that any given odorant item was not present in more than three tests and no two tests shared more than half of odorant items. The sequential forward feature selection algorithm was employed for the items of the four tests simultaneously under these constraints. Here, optimization criterion was the arithmetic mean of the accuracy, sensitivity, specificity, and AUC, assessed from leave-one-out cross-validation of the sum-based LDA model.

Measurement of agreement

To investigate the level of agreement of the scores of the four final smell tests, the equality of mean and variance was simultaneously tested using the Bradley–Blackwood test (Bradley & Blackwood, 1989). This test provided an F statistic calculated from the regression of each pair of smell tests. For this test, considering the two scores provided by a pair of tests (i.e., T_1 and T_2), the difference and the sum of paired scores ($D = T_1 - T_2$ and $S = T_1 + T_2$) are used for testing the null slope and intercept in the regression of D on S. This simultaneous testing was achieved through the equation for the Bradley–Blackwood test

$$F = \frac{\left[\frac{\sum D_i^2 - SSE}{2}\right]}{\left[\frac{SSE}{n-2}\right]},$$

where n is the number of paired observations, $\sum D_i^2$ is the sum of the squares of the n observed differences and SSE is the residual sum of squares from the regression of D on S. The null hypothesis is the equality of the two measurements. The alpha level for the critical F statistic was set at 0.001. When F is greater than $F_{critical}$ (2, n -2), the null hypothesis of the equality of the two measurements is rejected (in the current context the acceptance of the null hypothesis is desired).

Test validation using the independent subject group

The 32 subjects self-administered each of the parallel tests on five consecutive days. The first smell test was performed for a second time on the fifth day (with a different order of the same odorants as Test 1). Each test provided a score between 0 and 8. A subject's overall score was also computed as the sum of the scores of the parallel tests (maximum possible: 40). The UPSIT® score and clinical category for smell function for each subject was compared with the parallel test scores. Spearman's rho was used to determine the correlation coefficient between the UPSIT scores and total scores. The AUC of the receiver operating characteristic curve was used to evaluate the strength of the parallel smell tests for clinical evaluation of the subjects' smell function category.

Results

Different machine learning algorithms resulted in different odorant sets

The classification performances achieved by each machine learning method throughout the feature selection process are summarized in Fig. 1. The number of odorants where each method achieved its highest performance varied between 7 and 29. As Fig. 1 indicates for all applied machine learning algorithms, after reaching the highest classification performance, the addition of more odorants either was inconsequential or even detrimental to the performance. The LDA, logistic regression, and Adaboost methods were more robust to the addition of more odorants and had only a slight decline in performance, whereas the other methods
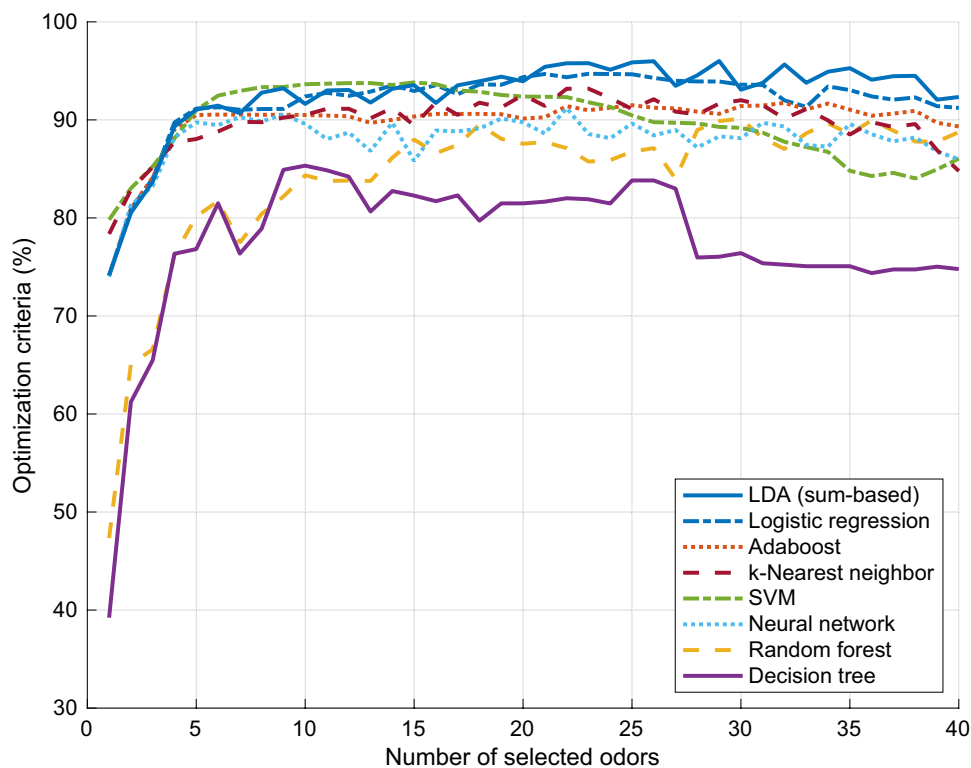

Fig. 1 Optimization criteria for different machine learning methods. *Note.* Optimization criteria (arithmetic mean of the accuracy, sensitivity, specificity, and AUC metrics) achieved during the feature selection strategy with increasing odorant counts for each machine learning method. Leave-one-out cross-validation was used to calculate the classification performance metrics

performed less well when more than the optimal number of odorants was used.

The optimal UPSIT® odorant sets for each of the machine learning methods and their corresponding cross-validation performance metrics are shown in Table 2. Our feature selection strategy was designed to optimize the arithmetic mean of the accuracy, sensitivity, specificity, and AUC metrics. The simple sum based linear LDA method using 29 odorants had the best overall performance, with an accuracy of 95.7%, sensitivity of 94.0% (true-positive rate), specificity of 97.0% (true-negative rate), and AUC of 0.97. Logistic regression using 23 odorants achieved equivalent specificity with slightly lower performance in other metrics. Compared to LDA, support vector machines (SVM) with 15 odorants achieved better sensitivity (97.0%) but had lower specificity (89.4%). k-Nearest neighbor (kNN, k = 3 with city block distance metric) with 23 features achieved 99.2% specificity at the cost of lower sensitivity (85.0%). AdaBoost and random forest methods performed similarly and had balanced sensitivity and specificity. Neural network (using ten hidden elements) and decision tree methods performed poorly with inferior sensitivity performance. The LDA decision threshold was 20 odorants, meaning a subject is predicted

to have smell loss (i.e., being COVID-19-positive) if they correctly identified 20 or fewer of the 29 selected odorants. Interestingly, LDA using all 40 UPSIT® odorants (decision threshold of 27) where no feature selection is performed had slightly lower accuracy and sensitivity but had the same specificity as the LDA method with 29 optimized features.

For each method, using the odorant sets from other methods did not produce a better performance than the optimized odorant set selected for that specific method. Thus, the differences in the odor sets selected by different methods were not due to the non-exhaustive nature of our feature selection strategy. While each method had a separate set of odors that optimized its performance, there were some odorants common to all methods. For instance, grass and lemon were selected by all eight machine learning models, while menthol was selected by all methods except SVM. Inclusion of the available demographic information, namely age and sex, did not result in any improvement beyond what was achievable with the odorant test data alone (see supplementary Table S1).

The order and incremental contributions of the selected odorants to the classification performance of the LDA model are shown in Fig. 2. The performance metrics were robust to

Table 2 Leave-one-out cross-validation performance of machine learning methods using features selected for optimized classification performance

Methods	Crit	Acc	Sens	Spec	AUC	Number of odors (selected odors)
LDA (sum-based)	96.0	95.7	94.0	97.0	0.97	29 (grass, apple, lemon, menthol, mint, baby powder, motor oil, natural gas, peach, grape, wintergreen, onion, popcorn, magnolia, soap, garlic, turpentine, gingerbread, peanut, bubble gum, indian spice, cinnamon, smoke, jasmine, rubber tire, chocolate, strawberry, daffodil, cheese)
Logistic regression	94.7	94.4	91.0	97.0	0.96	23 (grass, apple, lemon, menthol, natural gas, baby powder, onion, soap, motor oil, mint, leather, peanut, wintergreen, watermelon, pizza, jasmine, popcorn, magnolia, gingerbread, chocolate, indian spice, turpentine, rubber tire)
SVM	93.8	92.7	97.0	89.4	0.96	15 (grass, garlic, gingerbread, lemon, mint, menthol, fish, strawberry, pineapple, natural gas, indian spice, soap, rubber tire, chocolate, rose)
k-nearest neighbor	93.2	93.1	85.0	99.2	0.95	23 (grass, apple, motor oil, wintergreen, leather, gingerbread, clove, chocolate, pineapple, peanut, coconut, soap, baby powder, rubber tire, garlic, grape, mint, magnolia, cheese, lemon, daffodil, rose, onion)
Adaboost	92.4	91.4	90.0	92.4	0.95	19 (apple, grass, menthol, lemon, mint, smoke, onion, turpentine, popcorn, garlic, watermelon, magnolia, pizza, cheese, pineapple, bubble gum, wintergreen, peanut, motor oil)
LDA – UPSIT score	92.3	91.4	84.0	97.0	0.97	40 (grass, apple, lemon, menthol, mint, baby powder, motor oil, natural gas, peach, grape, wintergreen, onion, popcorn, magnolia, soap, garlic, turpentine, gingerbread, peanut, bubble gum, indian spice, cinnamon, smoke, jasmine, rubber tire, chocolate, strawberry, daffodil, cheese, leather, coconut, fish, pizza, pineapple, rose, lilac, watermelon, paint thinner, clove, raspberry)
Random forest	92.1	90.9	92.0	90.2	0.93	6 (grass, garlic, gingerbread, lemon, mint, menthol)
Neural network	91.1	89.2	83.0	93.9	0.91	7 (lemon, apple, popcorn, menthol, grass, natural gas, paint thinner
Decision tree	89.3	89.2	85.0	92.4	0.90	10 (apple, grass, menthol, lemon, grape, soap, cheese, rose, coconut, peach)

Optimization criteria (Crit), accuracy (Acc), sensitivity (Sens), and specificity (Spec) are shown as percentages and area under the receiver operating curve (AUC) is from 0 to 1. The number of features selected for each method is shown, followed by the list of features ordered by their contribution to the performance. Odorant names indicate which of the UPSIT® odorant/response sets were employed. LDA - UPSIT® score method using all 40 odorants where no feature selection is employed is also included for comparison. The best performance for each metric among tested machine learning methods is highlighted in bold
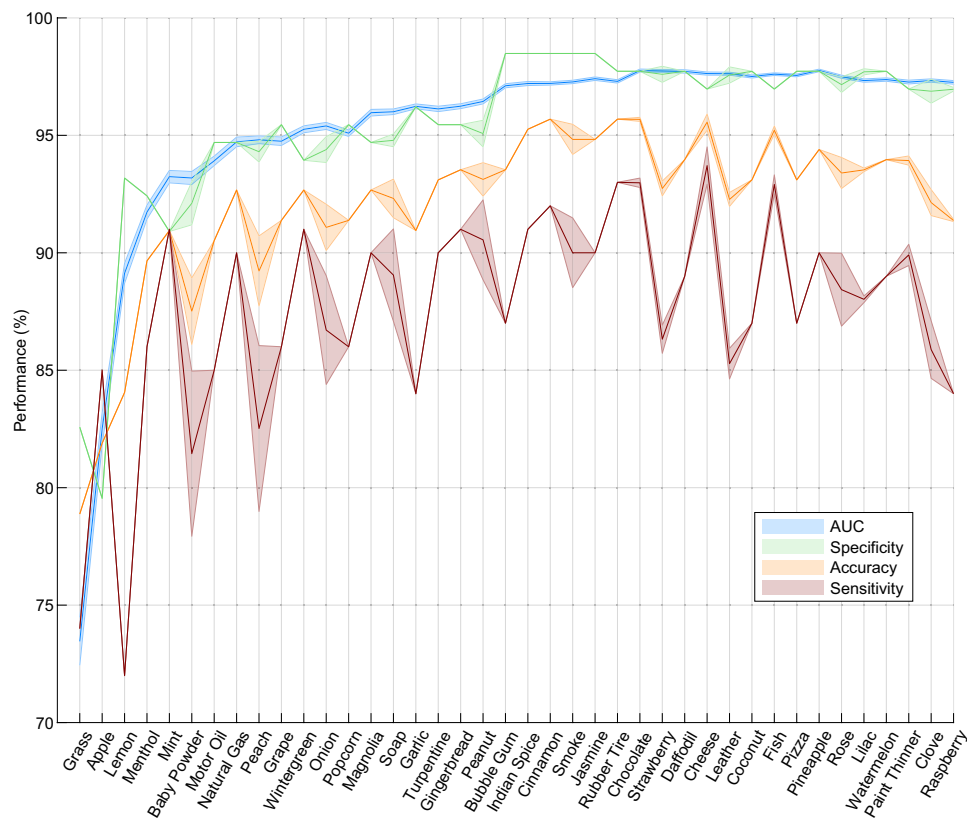
Fig. 2 Sequential feature selection performance metrics for the LDA model. *Note.* Odorants appear in the order of their inclusion to the selected feature set during the sequential feature selection algorithm. Ten-fold cross validation was used to quantify the variability.

Shaded areas show the range of performance values observed in ten-fold cross validations. Odorant names indicate which of the UPSIT® odorant/response sets were employed

cross-validation iterations, with little variability in all metrics, except for sensitivity, which had up to 5% variability within the ten-fold cross-validation iterations.

The largest contributions to the performance occurred for up to eight odorants and in certain machine learning methods (i.e., decision tree and random forest); brief tests with as few as 3–5 odorant items could not achieve

90% accuracy (Fig. 1). There were diminishing or negative returns for including additional odorants after the sweet spot of eight odorants. The classification performances of the machine learning methods were then restricted to the use of an optimized set with eight odorants (Table 3). The sum-based LDA method, which is the most practical of these machine learning methods, achieved 92.7% accuracy, 90.0%

Table 3 Leave-one-out cross-validation performance using only eight features

Method	Crit	Acc	Sens	Spec	AUC	Odors
SVM	93.3	93.1	94.0	92.4	0.93	Grass, garlic, gingerbread, lemon, mint, menthol, fish, strawberry
LDA (sum-based)	92.8	92.7	90.0	94.7	0.93	Apple, grass, menthol, lemon, natural gas, mint, baby powder, motor oil
Logistic regression	92.7	92.2	89.0	94.7	0.95	Apple, grass, menthol, lemon, natural gas, mint, baby powder, motor oil
Decision tree	91.3	91.4	89.0	93.2	0.91	Apple, lemon, menthol, grass, smoke, peanut, baby powder, chocolate
Random forest	91.0	89.7	86.0	92.4	0.93	Grass, apple, lemon, turpentine, peanut, fish, menthol, mint
k-nearest neighbor	91.0	90.5	87.0	93.2	0.93	Grass, garlic, strawberry, lemon, mint, menthol, apple, cheese
Adaboost	90.8	89.7	88.0	90.9	0.94	Grass, garlic, gingerbread, lemon, mint, menthol, fish, strawberry
Neural network	90.3	88.8	85.0	91.7	0.93	Grass, apple, lemon, menthol, mint, baby powder, motor oil, natural gas

Optimization criteria (Crit), accuracy (Acc), sensitivity (Sens), and specificity (Spec) are shown as percentages and area under the receiver operating curve (AUC) is from 0 to 1. Odorant names indicate which of the UPSIT® odorant sets were employed. The best performance for each metric is highlighted in bold

sensitivity, and 94.7% specificity. The LDA decision threshold was correct identification of six odorants, meaning a subject was predicted to be normosmic (i.e., not belonging to the COVID-19 group) if he or she correctly identified six or more of the eight odorants.

Brief smell tests optimized for sequential testing

In order to provide brief practical smell tests that could be administered for serial testing (screening) in a short period of time, we identified four eight-item sets of odorants. Sum-based LDA was the candidate for the extraction of such serial tests because of its high performance (Tables 2 and 3) and simple implementation without the use of overly complicated computer modeling. A constrained optimization procedure was used to select the odorants for each set to maximize the diversity among them. Specifically, an iterative feature selection strategy simultaneously added odorants to each of the four sets, limiting the overlap of odorants to less than 50% between any pair of the final sets. The choice of limiting each set to eight odorants was based on the sweet spot observed in Fig. 1 and Table 3 for the trade-off between practicality and classification performance.

The four optimized eight-odorant sets that were most useful for differentiating smell loss (COVID-19 patients) from normosmia (healthy controls) are shown in Table 4. These sets achieved an accuracy of 91–93% with 88–92% sensitivity, and 90–95% specificity. The LDA decision threshold for each set was six odorants, meaning a subject is predicted to be normosmic (i.e.., not belonging to the COVID-19 group) if he or she correctly identifies six or more of the eight odorants. These odorant sets could be successively employed in the same subjects on multiple days to minimize the likelihood of recalling previous sets of stimuli. This should make these specific odorant sets extremely useful in longitudinal tracking of the smell function of individuals when rapid screen is needed.

Finally, for an optimal sequential test, one expects to have low differences between parallel test scores for a given subject and large between-subject variability to assure that the full range of the scale is being used. For all subjects, the test scores for the sequential tests were computed. The level of agreement of each pair of tests was then measured using the Bradley–Blackwood test with $F_{Critical}$ (2,139) = 7.26. The computed F statistic for pair combinations of the optimized smell tests were 0.07, 2.19, 3.36, 2.02, 1.48, and 6.29, which were all less than the $F_{Critical}$, making them pass the simultaneous test of equality of t means and variances.

Test performance of the independent validation group

Each of the 32 subjects performed parallel tests on five consecutive days (four tests in Table 4 and the repetition of test 1 on the fifth day). The mean (SD) scores of the five consecutive tests were 5.06 (1.78), 5.03 (1.58), 5.53 (2.17), 5.25 (1.60), and 5.53 (2.13), and the mean total score of the of the five tests was 26.40 (7.33). A repeated measure analysis of variance indicated that the five test scores were not significantly different [F(4,124) = 1.11, p = 0.35]. The AUCs of ROCs for parallel tests were between 0.79 to 0.83 in detecting olfactory dysfunction and the AUC increased to 0.91 when the parallel test scores were totaled (Fig. 3).

Discussion

Alterations in the sense of smell are among the first signs of a number of serious medical diseases and disorders, including COVID-19 and such neurodegenerative diseases as Parkinson's and Alzheimer's. The sequential screening of smell function is rarely practiced in the clinical setting, despite the usefulness of such testing in identifying the onset of disorders such as COVID-19. One potential reason is the lack of parallel short tests to overcome the subject's recall bias.

In this study, a number of machine-learning models were compared using UPSIT® odorant items to establish optimal brief screening tests for differentiating patients with smell dysfunction from healthy controls. We used COVID-19 patients as a surrogate for smell dysfunction in general, as it provides a strong metric for such differentiation. Although most of the investigated models performed quite well, the

Table 4 Optimized odorant sets for multiple testing of smell function

Test #	Odors	Crit	Acc	Sens	Spec	AUC
1	Grass, apple, lemon, menthol, mint, baby powder, motor oil, natural gas	92.8	92.7	90.0	94.7	0.93
2	Watermelon, grass, turpentine, garlic, mint, onion, lemon, popcorn	92.0	92.2	89.0	94.7	0.92
3	Grass, apple, pineapple, cheese, motor oil, wintergreen, bubble gum, natural gas	91.6	90.9	92.0	90.2	0.93
4	Lemon, onion, grape, magnolia, gingerbread, natural gas, menthol, apple,	91.2	91.4	88.0	93.9	0.91

Leave-one-out cross-validation of each feature set was done using the LDA (sum-based) model. Optimization criteria (Crit), accuracy (Acc), sensitivity (Sens), and specificity (Spec) are shown as percentages and area under the receiver operating curve (AUC) is from 0 to 1. Feature sets were determined by simultaneous optimization of odorant combinations, under constraints to limit overlap (≤ 4 items) between sets
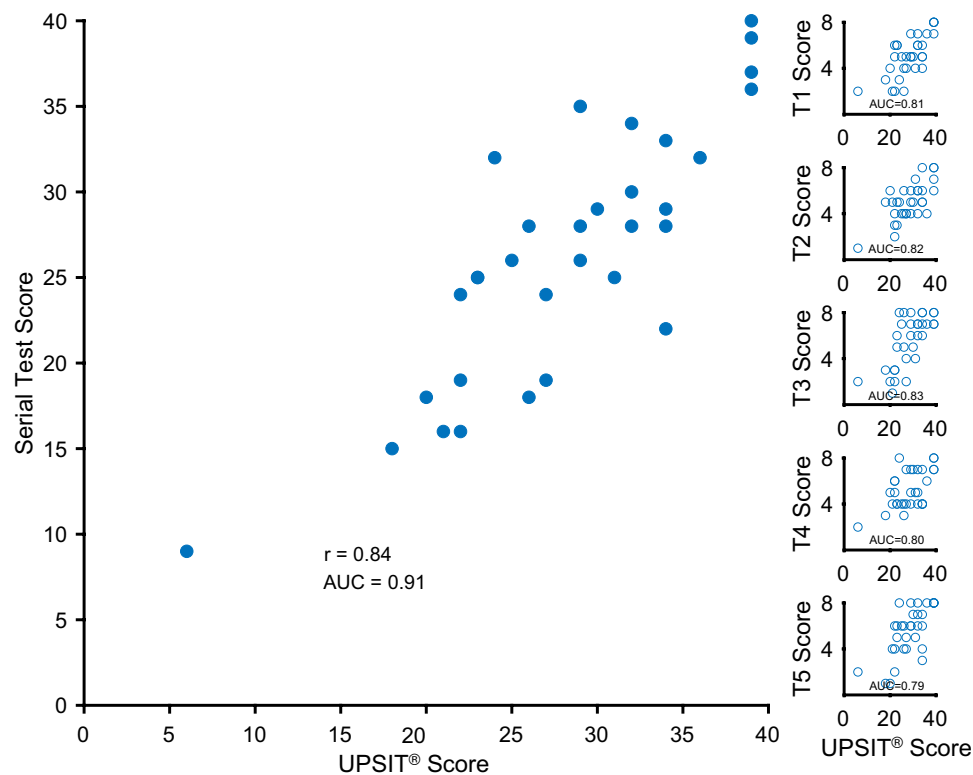
Fig. 3 Comparison of sequential test score with UPSIT® score. *Note.* Total score of the sequential tests are summed up as serial test score which is correlated with UPSIT score (*r* = 0.84, *p* < 0.0001). Area under the curve of the receiver operating characteristics curve analysis was 0.91 in differentiating normosmics from olfactory compromised subjects. The right subpanels compare each parallel smell test with UPSIT® score with the AUCs of the ROCs

sum based LDA model was superior, exhibiting 95.7% accuracy, 94.0% sensitivity, and 97.0% specificity. A minimum of eight odorant/response items was found to be needed to produce accuracy rates of 91–93%. We found that brief tests with as few as 3–5 odorant items could not achieve such high levels of accuracy.

In addition to determining the minimum number of odorant items needed to achieve high sensitivity, we identified four sets of short eight-item odorant tests useful for serial testing while minimizing the likelihood of remembering previously presented stimuli. Serial testing potentially allows for detection and continuous monitoring of smell dysfunction and may also be useful in tracking disease progression, recovery, and intervention effectiveness. Such testing may also be useful for testing persons who cannot accurately perform longer tests in single test sessions (e.g., due to fatigue and difficulties concentrating, as occurs in some elderly persons or patients with dementia).

To increase the generalizability to other populations and to focus on the predictive power of the smell tests, per se, we focused on using only the responses to the odorant items in the machine learning analysis and didn't emphasize non-olfactory demographic variables. Inclusion of the limited demographic information available, namely age and sex, did not result in any improvement beyond that achievable with odorant test data alone (see Supplementary File). If one wished to combine the smell test data developed in this study with other demographic data, such inclusion would likely necessitate a more sophisticated scoring.

To our knowledge, this is the first study to develop multiple parallel brief smell identification tests useful for detecting temporal changes in smell function with high sensitivity and specificity. The study has both strengths and weaknesses. Among its strengths are the use of data from a relatively large number of persons with and without smell dysfunction. Another strength was the systematic exploration of the efficacy of numerous machine learning algorithms to optimize sensitivity in detecting smell dysfunction. Several potential weaknesses should be acknowledged. First, one might argue that the current sample size was inadequate for the performed analyses. However, this is unlikely since the cross-validation procedure addresses this concern and demonstrated the opposite. Second, data augmentation (Shorten & Khoshgoftaar, 2019) was not performed in this study. Such augmentation is employed to reduce overfitting when training a machine learning model. However, since

the dataset contained relatively balanced numbers of participants in the two groups, this procedure was not necessary. A third concern might be that the tests we employed were reliant on smell identification and did not account for changes that may occur on other measures, such as an odor threshold. This is also unlikely, since it is well documented that identification tests such as the UPSIT® are strongly correlated with a number of types of olfactory tests, including detection threshold tests (Doty et al., 1994; Doty et al., 2019).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02102-8.

Data availability The datasets generated and analyzed during this study are not publicly available but are available from the corresponding author on reasonable request.

Code availability The codes used in this study will be shared publicly in GitHub from the time of publication.

Declarations

Ethics approval Ethical approval was waived by the local Ethics Committee of Institute for Research in Fundamental Sciences in view of the retrospective nature of the study and all the raw data were already collected through another approved project. All subjects gave written informed consent, and all elements of the study were approved by the appropriate IRBs.

Consent to participate Written informed consent was obtained from each subject.

Consent for publication Not applicable.

Competing interests RLD is a consultant to Destiny Pharma, Satsuma Pharmaceuticals, Merck Pharmaceuticals, and Johnson & Johnson. He receives royalties from Cambridge University Press, Johns Hopkins University Press, and John Wiley & Sons, Inc. and is president of, and a major shareholder in, Sensonics International, a manufacturer and distributor of smell and taste tests. STM is a physician scientist and RS is a research coordinator, both affiliated with Sensonics International. All other authors declare they have no relevant or financial competing interests.

References

Bradley, E. L., & Blackwood, L. G. (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician, 43*(4), 234–235. https://doi.org/10.1080/00031305.1989.10475665

Breiman, L. (2001). *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees* (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Devanand, D. P., Lee, S., Manly, J., Andrews, H., Schupf, N., Doty, R. L., Stern, Y., Zahodne, L. B., Louis, E. D., & Mayeux, R. (2014). Olfactory deficits predict cognitive decline and Alzheimer dementia in an Urban Community. *Neurology, 84*(2), 182–189. https://doi.org/10.1212/wnl.0000000000001132

Doty, R. L. (1995). The smell identification test™ administration manual. Haddon heights,NJ; Sensonics International.

Doty, R. L., Shaman, P., & Dann, M. (1984). Development of the University of Pennsylvania Smell Identification Test: A standardized microencapsulated test of olfactory function. *Physiology & Behavior, 32*(3), 489–502. https://doi.org/10.1016/0031-9384(84)90269-5

Doty, R. L., Reyes, P. F., & Gregor, T. (1987). Presence of both odor identification and detection deficits in Alzheimer's disease. *Brain Research Bulletin, 18*(5), 597–600. https://doi.org/10.1016/0361-9230(87)90129-8

Doty, R. L., Smith, R., Mckeown, D. A., & Raj, J. (1994). Tests of human olfactory function: Principal components analysis suggests that most measure a common source of variance. *Perception & Psychophysics, 56*(6), 701–707. https://doi.org/10.3758/bf03208363

Doty, R. L., Wylie, C., Potter, M., Beston, R., Cope, B., & Majam, K. (2019). Clinical validation of the olfactory detection threshold module of the Snap & Sniff® olfactory test system. *International Forum of Allergy & Rhinology, 9*(9), 986–992. https://doi.org/10.1002/alr.22377

Doty, R. L., Popova, V., Wylie, C., Fedgchin, M., Daly, E., Janik, A., Ochs-Ross, R., Lane, R., Lim, P., Cooper, K., Melkote, R., Jamieson, C., Singh, J., & Drevets, W. C. (2021). Effect of Esketamine nasal spray on olfactory function and nasal tolerability in patients with treatment-resistant depression: Results from four multicenter, randomized, double-blind, placebo-controlled, phase III studies. *CNS Drugs, 35*(7), 781–794. https://doi.org/10.1007/s40263-021-00826-9

Freund, Y., & Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society of Artificial Intelligence, 14*(5), 771–780.

Gerkin, R. C., Ohla, K., Veldhuizen, M. G., Joseph, P. V., Kelly, C. E., Bakke, A. J., Steele, K. E., Farruggia, M. C., Pellegrino, R., Pepino, M. Y., Bouysset, C., Soler, G. M., Pereda-Loth, V., Dibattista, M., Cooper, K. W., Croijmans, I., Di Pizio, A., Ozdener, M. H., Fjaeldstad, A. W., … Baguma, M. (2020). Recent smell loss is the best predictor of COVID-19 among individuals with recent respiratory symptoms. *Chemical Senses*, 46, bjaa081. https://doi.org/10.1093/chemse/bjaa081

Grimm, L. G., & Yarnold, P. R. (1995). *Reading and understanding multivariate statistics*. American Psychiatric Association.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. Prentice-Hall.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications, 13*(4), 18–28. https://doi.org/10.1109/5254.708428

Liu, B., Luo, Z., Pinto, J. M., Shiroma, E. J., Tranah, G. J., Wirdefeldt, K., Fang, F., Harris, T. B., & Chen, H. (2019). Relationship between poor olfaction and mortality among community-dwelling older adults. *Annals of Internal Medicine, 170*(10), 673. https://doi.org/10.7326/m18-0775

MATLAB®. (2020). Version 9.8 (R2020a). The MathWorks Inc, Natick, Massachusetts, United States.

Moein, S. T., Hashemian, S. M. R., Mansourafshar, B., Khorram-Tousi, A., Tabarsi, P., & Doty, R. L. (2020a). Smell dysfunction:

A biomarker for COVID-19. *International Forum of Allergy & Rhinology, 10*(8), 944–950. https://doi.org/10.1002/alr.22587

Moein, S. T., Hashemian, S. M. R., Tabarsi, P., & Doty, R. L. (2020b). Prevalence and reversibility of smell dysfunction measured psychophysically in a cohort of COVID-19 patients. *International Forum of Allergy & Rhinology, 10*(10), 1127–1135. https://doi.org/10.1002/alr.22680

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for Deep Learning. *Journal of Big Data, 6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0