САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

на правах рукописи

Кононов Александр Станиславович

РАЗРАБОТКА МЕТОДА ДИАГНОСТИКИ РАКА ЛЕГКИХ НА ОСНОВЕ ОНЛАЙН АНАЛИЗА ВЫДЫХАЕМОГО ВОЗДУХА С ИСПОЛЬЗОВАНИЕМ МЕТАЛЛООКСИДНЫХ ГАЗОЧУВСТВИТЕЛЬНЫХ СЕНСОРОВ

Диссертация на соискание ученой степени
кандидата химических наук

Научная специальность 1.4.2. Аналитическая химия

Научный руководитель
доктор физ.-мат. наук, профессор
Ганеев Александр Ахатович

Санкт-Петербург

2021

Оглавление

Введение **4**

Глава 1. Обзор литературных данных **9**

1.1. Потенциальные биомаркеры рака легкого в выдыхаемом воздухе 9

1.2. Методы пробоотбора и пробоподготовки при анализе выдыхаемого воздуха 11

1.3. Методы анализа выдыхаемого воздуха, пригодные для выявления рака легкого 14

1.4. Методы обработки многомерных данных 26

Глава 2. Используемые методы исследования и приборы **41**

2.1. Описание характеристик сенсоров 41

2.2. Методика приготовления модельных газовых смесей 46

2.3. Анализ модельных газовых смесей и проб выдыхаемого воздуха в медицинском исследовании с использованием МС 1 48

2.4. Анализ модельных газовых смесей для переноса градуировочных зависимостей с использованием МС 2.1 и МС 2.2 58

Глава 3. Разработка метода онлайн-анализа выдыхаемого воздуха для диагностики рака легких с использованием мультисенсорной системы **62**

3.1. Описание медицинского исследования 62

3.2. Описание процедуры проведения анализа ВВ пациентов 63

3.3. Выбор наиболее эффективного алгоритма обработки данных и классификационной модели 65

3.4. Анализ полученных результатов 72

Выводы 76

Глава 4. Разработка метода переноса градуировочной зависимости и стандартизации откликов между двумя мультисенсорными системами **77**

4.1. Описание дизайна исследования 77

4.2. Оценка результатов стандартизации при классификации индивидуальных образцов ЛОС 80

4.3. Оценка результатов стандартизации при классификации смесей ЛОС 83

Выводы 87

Заключение **88**

Список сокращений и условных обозначений **89**

Список литературы **91**

Личный вклад автора состоял в сборе и анализе литературных данных, активном участии в постановке задач, исследовании, планировании, подготовке и проведении экспериментов, исследовании физико-химических свойств сенсоров и обработке полученных данных, а также в анализе, интерпретации и обобщении полученных результатов, подготовке докладов и публикаций.

Благодарности. Искренне благодарю всех, кто способствовал выполнению данной работы. Особую благодарность выражаю Ганееву Александру Ахатовичу за наставничество на всех этапах научно-исследовательской работы, экспериментальный опыт, обучение критическому мышлению и умению выявлять суть проблемы. Благодарю Джагацпаняна Игоря Эдуардовича за многократные продуктивные обсуждения тонкостей полупроводниковых сенсоров и газоанализаторов, а также за помощь в подготовке статей для публикации.

Выражаю благодарность моим соавторам и коллегам: Коротецкому Борису Александровичу, Губаль Анне Романовне, Чучиной Виктории Александровне, Нефедову Андрею Олеговичу, Васильеву Алексею Андреевичу, Арсеньеву Андрею Ивановичу.

В заключение благодарю супругу, родителей, брата, друзей и близких людей за поддержку.

Введение

Раннее выявление рака легких (РЛ), как правило, связано со значительным улучшением эффективности его лечения. Однако используемые в настоящее время методы ранней диагностики РЛ обладают недостаточной эффективностью, что приводит к выявлению болезни на поздней стадии и, как следствие, к высокой смертности. В связи с этим разработка высокопроизводительного и надежного метода диагностики является важной задачей, которая требует наискорейшего решения. Анализ выдыхаемого воздуха (ВВ) для определения ряда органических соединений, являющихся признанными биомаркерами РЛ, становится многообещающим методом раннего выявления РЛ. Это направление исследований привлекает все больший интерес, что подтверждается ежегодно увеличивающимся количеством научных публикаций по данной тематике. В этой области в принципе возможно создание не только скринингового метода ранней диагностики РЛ, но и метода, позволяющего контролировать состояние больного РЛ как до лечения, так и после. Однако при создании подобного метода необходимо выполнить ряд условий, зачастую противоречивых. Метод должен иметь малое время пробоотбора и анализа, быть относительно дешевым и неинвазивным, и, по возможности, работать в онлайн-режиме. Важнейшим требованием, предъявляемым не только к рассмотренным, но и к любым другим методам диагностики РЛ является высокие уровни специфичности и прогностичности положительного результата [1]. Соответствующие величины должны быть не менее 98-99%, в противном случае резко возрастает количество неоправданных биопсий, сопряжённых с риском осложнений, увеличивается использование дополнительных методов обследования и увеличивается его стоимость. Требования к чувствительности и прогностичности отрицательного результата менее жесткие – для чувствительности не менее 90%, для прогностичности отрицательного результата – 85% [2].

На данном этапе развития анализа ВВ все используемые для диагностики РЛ методы анализа газовая хромато-масс-спектрометрия (GC-MS), масс-спектрометрия с реакцией переноса протона (PTR-MS), поликапиллярная спектрометрия ионной подвижности (MCC-IMS) не полностью удовлетворяют предъявляемым требованиям: GC-MS имеет низкую производительность, высокую трудоемкость и возможность

использования этого метода только в оффлайн-режиме, а методы, которые можно использовать в онлайн-режиме: PTR-MS, MCC-IMS – являются недостаточно чувствительными. Впрочем, все эти методы достаточно сложно использовать для непосредственного контроля состояния больного РЛ. Значительно проще для этих целей использовать мультисенсорную систему (МС) для распознавания образов ВВ типа «электронный нос» (ЭН), для которой могут быть достигнуты приемлемые уровни специфичности и прогностичности положительного результата, хотя и они требуют улучшения. Отметим, что ныне существующие ЭН системы не позволяют полностью решить проблемы диагностики РЛ, что в значительной степени связано со свойствами используемых для этих целей сенсоров. Среди недостатков ныне используемых ЭН недостаточная перекрестная чувствительность по основным биомаркерам РЛ и недостаточная долговременная стабильность их аналитических характеристик. Эти недостатки присущи многим типам сенсоров, но в наименьшей степени они касаются металлооксидных сенсоров, которые, правда обладают другим недостатком - производственной вариабельностью. Существующие технологии изготовления не позволяют получить сенсоры, имеющие идентичные характеристики и, следовательно, идентичный характер отклика к аналиту. Это препятствует масштабному производству мультисенсорных систем, при котором можно было бы собирать данные в общую базу и использовать единую классификационную модель для всех приборов. Для решения этой проблемы существует ряд методов по устранению инструментальной вариации, которые часто называют переносом градуировочных зависимостей. Этот подход состоит в том, чтобы преобразовать данные с дополнительных устройств (на которых измерены тестовые образцы) в соответствие с ведущим или основным устройством (на данных которого обучена модель прогнозирования). Набор образцов для стандартизации измеряется как на основном, так и на том устройстве, которое необходимо стандартизировать. Затем применяются алгоритмы регрессии для установления зависимости между переменными.

В связи с этим разработка нового прямого метода диагностики, включающего в себя как разработку новых сенсоров, так и мультисенсорной системы на их основе с возможностью оперировать единой базой данных, для создания системы диагностики РЛ по выдыхаемому воздуху является очень важной и актуальной задачей.

Целью данной работы разработка методологии онлайн-анализа ВВ с помощью системы газочувствительных металлооксидных сенсоров для диагностики РЛ. В связи с поставленной целью решались следующие **задачи**:

1. Разработка схемы онлайн-анализа ВВ с помощью системы газочувствительных металлооксидных сенсоров, не требующей дополнительной пробоподготовки;
2. Определение относительных чувствительностей ЛОС для предварительного отбора сенсоров;
3. Проведение сравнительного медицинского исследования и анализа ВВ пациентов группы больных РЛ и здоровых людей;
4. Выбор эффективного алгоритма обработки данных, позволяющих эффективно разделять группы больных РЛ и здоровых людей с высокой чувствительностью и специфичностью безотносительно внешних факторов состояния пациента (возраст, пол, курение и др.) и основываясь исключительно на откликах мультисенсорной системы;
5. Проведение исследования анализа ЛОС на двух сенсорных системах с идентичными группами сенсоров и разработка подхода для стандартизации мультисенсорных систем.

Научная новизна:

1. Предложена, создана и апробирована схема онлайн-анализа ВВ с помощью системы газочувствительных металлооксидных сенсоров, не требующая дополнительной пробоподготовки. Эта система сочетает в себе онлайн-измерение и временное интегрирование сигнала, высокую скорость продувки, и, как следствие, высокое быстродействие с минимизацией эффектов памяти;
2. Разработан и апробирован алгоритм обработки экспериментальных данных, позволяющий эффективно разделять больных РЛ и здоровых людей с высокой чувствительностью (90.5 ± 2.6)%, специфичностью (98.1 ± 1.5)%, точностью (94.0 ± 1.6)%, ROC AUC 0.961 ± 0.018, прогностичностью положительного результата (98.3 ± 1.3)% и прогоностичностью отрицательного результата (89.9 ± 2.7)%;
3. Разработан и апробирован алгоритм обработки данных для оценки результативности переноса градуировочных зависимостей между двумя мультисенсорными системами с помощью стандартизации откликов на модельных

задачах классификации.

Практическая значимость работы:

1. Разработана система онлайн-анализа ВВ с использованием ячейки из 6 газочувствительных МО сенсоров, позволяющая за 25-30 минут проанализировать ВВ одного пациента при 3 температурных режимах;

2. Разработана схема онлайн-анализа и алгоритм обработки данных, позволяющая эффективно разделять группы больных РЛ и здоровых людей с высокой чувствительностью (90.5 ± 2.6)%, специфичностью (98.1 ± 1.5)%, точностью (94.0 ± 1.6)%, ROC AUC 0.961 ± 0.018, прогностичностью положительного результата (98.3 ± 1.3)% и прогоностичностью отрицательного результата (89.9 ± 2.7)%;

3. Разработаны методические подходы к стандартизации сенсорных систем с идентичными сенсорами с помощью метода переноса градуировочной зависимости, что позволяет использовать и обрабатывать результаты анализа ВВ с нескольких мультисенсорных систем в единой базе.

Основные положения, выносимые на защиту:

1. Система онлайн-анализа ВВ с использованием массива газочувствительных МО сенсоров для диагностики РЛ;

2. Алгоритм обработки экспериментальных данных, позволяющий эффективно разделять группы больных РЛ и здоровых людей с высокой чувствительностью (90.5 ± 2.6)%, специфичностью (98.1 ± 1.5)%, точностью (94.0 ± 1.6)%, ROC AUC 0.961 ± 0.018, прогностичностью положительного результата (98.3 ± 1.3)% и прогоностичностью отрицательного результата (89.9 ± 2.7)%.

Публикации и апробация работы:

Результаты диссертационной работы докладывались и обсуждались на следующих конференциях и конкурсах: Международная студенческая конференция “Science and Progress - 2018” (Санкт-Петербург, 2018), конкурс междисциплинарных студенческих и аспирантских проектов «Start-up СПбГУ — 2018» (Санкт-Петербург, 2018), VI Петербургский международный онкологический форум «Белые ночи 2020», Национальная (Всероссийская) конференция по естественным и гуманитарным наукам с международным участием «Наука СПбГУ – 2020» (Санкт-Петербург, 2020), Международная конференция по естественным и гуманитарным наукам «Science SPbU –

2020», Петербургский международный онкологический форум «Белые ночи 2021» (Санкт-Петербург, 2020).

По теме работы опубликованы 3 статьи в журналах, индексируемых в базах WoS и Scopus:

1. A.A. Ganeev, A.R. Gubal, G.N. Lukyanov, A.I. Arseniev, A.A. Barchuk, I.E. Jahatspanian, I.S. Gorbunov, A.A. Rassadina, V.M. Nemets, A.O. Nefedov, B.A. Korotetsky, N.D. Solovyev, E. Iakovleva, N.B. Ivanenko, A.S. Kononov, M. Sillanpaa and T. Seeger. Analysis of exhaled air for early-stage diagnosis of lung cancer: opportunities and challenges // Russian Chemical Reviews (2018) 87 (9), pp. 904-921, DOI: 10.1070/RCR4831;
2. A. Kononov, B. Korotetsky, I. Jahatspanian, A. Gubal, A. Vasiliev, A. Arsenjev, A. Nefedov, A. Barchuk, I. Gorbunov, K. Kozyrev, A. Rassadina, E. Iakovleva, M. Sillanpaa, Z. Safaei, N. Ivanenko, N. Stolyarova, V. Chuchina, A.Ganeev. Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer // Journal of breath research (2019) 14 (1), 016004, DOI: 10.1088/1752-7163/ab433d;
3. A. Arseniev, A. Nefedova, A. Ganeev, A. Nefedov, S. Novikov, A. Barchuk, S. Kanaev, I. Jahatspanian, A. Gubal, A. Kononov, S. Tarkov, N. Aristidov. Combined diagnostics of lung cancer using exhaled breath analysis and sputum cytology // Problems in oncology (2020) 66 (4), pp. 381-384, DOI: 10.37469/0507-3758-2020-66-4-381-384.

Работа выполнена в Институте Химии Федерального Государственного Бюджетного Образовательного Учреждения Высшего Образования «Санкт-Петербургский Государственный Университет» (2017-2021 гг.).

Глава 1. Обзор литературных данных

1.1. Потенциальные биомаркеры рака легкого в выдыхаемом воздухе

Анализ ВВ, в частности, для диагностических целей на данный момент является активно развивающейся областью исследований. [3]. Возможность использования анализа ВВ для выявления рака легких (РЛ) изучается в течение многих лет, и сейчас привлекает все большее внимание исследователей благодаря быстрому развитию метаболомики [4]. Метаболомический анализ ВВ обычно направлен на количественное определение метаболитов с низкой молекулярной массой (менее 1000 а.е.м.) [5]. Изменение концентраций таких соединений может быть вызвано различными патофизиологическими процессами, генетическими модификациями или факторами окружающей среды, влияющими на живые системы [5]. Такие изменения в ВВ могут являться предупреждающими признаками таких заболеваний как РЛ [6].

Летучие органические соединения (ЛОС), содержащиеся в ВВ, образуются в ходе реакций обмена, происходящих как в организме человека, так и в микробиоте. При патологических состояниях в симбиозе микробиоты неизбежно происходят сдвиги метаболизма, и, как следствие, происходит изменение продуцируемых веществ, в том числе низкомолекулярных. Такие соединения могут быть обнаружены в ВВ человека. В случае патологии перемены в спектре низкомолекулярных метаболитов микрофлоры в принципе могут быть детектированы с последующим диагностированием РЛ на ранних стадиях.

В выдохе человека присутствуют несколько сотен соединений, но только некоторые из них могут быть полезны для обнаружения РЛ на ранней стадии заболевания [2]. Для постановки надежного диагноза требуется идентификация определенных соединений, наличие или концентрация которых однозначно коррелирует с заболеванием. Согласно Всемирной организации здравоохранения: биомаркер – это любое вещество, структура или процесс, которые могут быть измерены в организме или его продуктах, а также влияют или предсказывают частоту исхода или заболевания [7]. Отметим, что биомаркеры для здоровых и больных людей отличаются, как правило, не их наличием/отсутствием, а диапазонами концентраций. Механизмы образования

потенциальных биомаркеров РЛ в выдохе человека подробно рассмотрены в данной работе [8].

Можно выделить некоторое количество соединений, информативность которых была показана в ряде работ. В таблице 1 представлены биомаркеры РЛ, для которых было показано значимое разделение между группой РЛ и группы здоровых (контрольной группы) и которые встречается не менее, чем в двух работах [3]. Биомаркеры сгруппированы по классам с указанием их возможной природы происхождения [9].

Таблица 1. Информативные биомаркеры РЛ в выдохе человека (в скобках указано количество работ, в которых биомаркер отмечен как информативный)

Класс соединений	**Потенциальный эндогенный источник**	**Основные соединения и/или производные**	**Экзогенный источник**
Алканы/ Алкены/ Алкадиены	Оксидативный стресс (пероксидация полиненасыщенных жирных кислот)	Изопрен (4), декан (3), бутан (3), пентан (3), ундекан (2), метилциклопентан (2), 4-метилоктан (2), пропан (2), 2-метилпентан (2), гептан (2)	Окружающая среда, пластик или топливо
Спирты	Метаболизм углеводородов, абсорбированных через желудочно-кишечный тракт	Пропан-1-ол (5), пропан-2-ол (3)	Окружающая среда, пища, дезинфицирующие средства
Альдегиды	Метаболизм спиртов; Пероксидация липидов	Гексаналь (4), гептаналь (3), пропаналь (3), бутаналь (2), пентаналь (2), октаналь (2), нонаналь (2)	Окружающая среда, пища, пищевые отходы, сигаретный дым
Кетоны	Окисление жирных кислот; Метаболизм белков	Бутан-2-он (5), ацетон (3), пентан-2-он (2)	Окружающая среда, пища, пищевые отходы, лекарства, ароматизаторы, краски

Карбоновые кислоты	Метаболизм аминокислот	Уксусная кислота (2), пропионовая кислота (2)	Пищевые консерванты, растворители, полимеры
Ароматические соединения	-	Этилбензол (4), стирол (4), бензальдегид (2), бензол (3), пропилбензол (2), 1,2,4-триметилбензол (2), о-ксилол (2)	Бензин, сигаретный дым, топливо, смолы, масла

На сегодняшний день опубликовано значительное количество работ с частично противоречивыми результатами: средняя концентрация биомаркера в ВВ испытуемых с РЛ может в одном исследовании быть значимо выше, а в другом – значимо ниже средней концентрации биомаркера в группе здоровых людей [6]. Также отметим, что разные группы исследователей использовали различные методы отбора и пробоподготовки проб и выявления биомаркеров. Отсутствие стандартной процедуры анализа ВВ является основной причиной расхождений в получаемых результатах.

В одной из обзорных работ, посвященных обзору потенциальных биомаркеров РЛ, было показано, что использование одного вещества недостаточно для успешного разделения группы РЛ и группы здоровых людей [3]. Наоборот, исследователи отмечают, что для диагностического теста необходим именно набор веществ, формирующий профиль ВВ пациента [3,6].

1.2. Методы пробоотбора и пробоподготовки при анализе выдыхаемого воздуха

Отбор проб является одним из важных этапов анализа ВВ. Существует ряд параметров, на которые необходимо обращать внимание, чтобы избежать ошибочных предположений о происхождении тех или иных идентифицированных соединений. К этим параметрам относятся тип ВВ (объем используемого дыхания), техника дыхания, кратность отбора, способ отбора, влияние ЛОС, присутствующих в окружающей среде, условия хранении и транспортировки проб. Все эти параметры подробно рассмотрены и обсуждены в работах [10–12]. В тех случаях, когда состав ВВ анализируется в онлайн-

режиме или в режиме реального времени, стадии отбора проб и предварительного концентрирования могут быть пропущены.

1.2.1. Особенности пробоотбора выдыхаемого воздуха

Для анализа состава ВВ можно отбирать смешанный экспираторный воздух или только альвеолярный воздух. При использовании первого варианта высок риск загрязнения пробы экзогенными соединениями из полости рта и мертвого пространства (носоглотка, трахея, бронхи и бронхиолы вплоть до их перехода в альвеолы), что может скомпрометировать результат анализа [10]. Альвеолярный воздух богат летучими соединениями крови, поэтому применение метода альвеолярного отбора считается более точным, обеспечивая представительность и постоянство качества пробы [13,14].

Использование различных техник дыхания, таких как задержка дыхания, гипервентиляция, дыхание против сопротивления и др., направлено как правило, либо на накопление выделяемых газов, либо на разделение фракций ВВ за один выдох [13,14].

Отбор пробы может быть достигнут за один или несколько полных выдохов. Анализ состава многократного выдыхания является более воспроизводимым с точки зрения состава пробы [10], однако однократное выдыхание, как правило, занимает меньше времени и более приемлемо для пациентов.

Следует отдельно упомянуть проблему конденсации водяного пара, присутствующего в ВВ, и перераспределения компонентов ВВ между конденсатом и газообразной фазой. Водяные пары, которыми насыщен ВВ, участвуют в переносе многих летучих и нелетучих соединений посредством растворения молекул (согласно коэффициентам распределения) внутри аэрозольной частицы [15,16]. В водяных парах аккумулируются все нелетучие соединения, такие как пероксид водорода, аденозин, лейкотриены, изопростаны, пептиды и цитокины [17]. Кроме того, полярные органические и неорганические соединения, такие как спирты, кетоны, карбоновые кислоты, аммиак и оксиды азота, могут частично концентрироваться в конденсате ВВ [18]. Для получения наиболее полной информации о составе ВВ иногда анализируют не только выдох, но и отдельно конденсат ВВ. Для борьбы с неконтролируемой конденсацией паров воды в пробоотборных устройствах и коммуникациях все элементы системы термостатируются при 37-40ºС.

1.2.2. Способы хранения проб выдыхаемого воздуха

Хранение проб ВВ может быть реализовано различными способами [10,19]. Наиболее распространенный и рекомендуемый способ отбора ВВ – использование пробоотборных тедларовых пакетов [20,21]. Пакеты изготавливают из таких химически инертных полимерных материалов, как поливинилфторид, перфторалкоксидные полимеры, политетрафторэтилен и поливинилиденхлорид [22]. Такие пакеты обладают рядом преимуществ: они непроницаемы для диффузии газов (если они дополнительно покрыты алюминиевой фольгой) [23], удобны в использовании (можно применять многократно, если после предыдущей пробы тщательно продувать очищенным воздухом, азотом или аргоном). Несмотря на все преимущества, мешки имеют недостатки: пластификаторы и растворители, используемые при производстве полимера, такие как фенол и *N,N*-диметилацетамид, могут высвобождаться в относительно высоких концентрациях, загрязняя пробу [24]. Пакеты уязвимы для проколов. Некоторые компоненты, например, гексан-1-аль и 2-метилбута-1,3-диен, не могут храниться в мешках более нескольких часов [25,26].

Другой способ – использование газонепроницаемых шприцев. Шприц объемом 50 мл соединяется с мундштуком, в который выдыхает пациент. Во время выдоха с помощью шприца отбирается примерно 20-30 мл ВВ, который затем переносят в вакуумированные стеклянные пробирки, где проба хранится до проведения анализа [13]. Еще одной формой хранения проб ВВ является конденсат ВВ [27].

Относительно недавно разработан дыхательный пробоотборник Bio-VOC [28]. Это устройство позволяет собирать альвеолярный воздух, а после завершения сбора проб ЛОС концентрируют с использованием системы твердофазной микроэкстракции (ТФМЭ) [29,30]. Основным недостатком является малый объем собираемого воздуха (100-150 мл) [29–31].

1.2.3. Предварительное концентрирование

Содержание ЛОС в ВВ может варьироваться от нескольких мкмоль×л$^{-1}$ до нескольких фмоль×л$^{-1}$ [13,32]. Поэтому в зависимости от используемого метода анализа состава ВВ необходимо прибегать к промежуточному этапу между отбором пробы и анализом для повышения содержания целевых компонентов.

Наиболее часто в качестве метода предварительного концентрирования при анализе ВВ используется концентрирование на твердых сорбентах с последующей термодесорбцией [33,34]. Это позволяет достичь пределов обнаружения на уровне ppt при объеме пробы до 1 л [35,36]. Несмотря на большой ассортимент твердых сорбентов с различной силой удерживания, рабочей температурой и гидрофобностью, один сорбент не способен адсорбировать все соединения, присутствующие в пробе ВВ, что связано с широким диапазоном летучести детектируемых ЛОС. Поэтому применяют многокомпонентные сорбционные трубки, в которых последовательно упакованы различные сорбенты с увеличением силы удерживания [37]. Концентрирование проводят при комнатной или более низкой температуре, а практически полную термодесорбцию аналитов с поверхности сорбента при 250-300 С.

Ключевыми источниками ошибки (потеря аналита или появление артефактов) при этом методе предварительного концентрирования являются деградация адсорбированных аналитов при хранении [38], термическое разложение или изомеризация некоторых соединений в процессе термодесорбции [39,40], и деградация материала сорбента [41,42].

1.3. Методы анализа выдыхаемого воздуха, пригодные для выявления рака легкого

1.3.1. Методы анализа с количественным определением летучих органических соединений в выдыхаемом воздухе

Метод газовой хромато-масс-спектрометрии (GC-MS), возможно, самый универсальный и чувствительный для определения ЛОС в выдохе, позволяющий анализировать большое количество соединений в диапазоне от ppb до ppt. Поэтому можно сказать, что GC-MS является золотым стандартом при определении низких содержаний ЛОС в выдохе человека [43].

Несмотря на свою универсальность и низкие пределы обнаружения, метод GC-MS имеет ряд недостатков, связанных в первую очередь с пробоотбором и пробоподготовкой. Сама же процедура проведения анализа и обработки его результатов при современном уровне автоматизации, оснащенности селективными детекторами и доступности разнообразных хроматографических колонок обычно не вызывает больших затруднений. Но внедрение метода GC-MS в клинических условиях имеет ряд ограничений из-за

высоких затрат, трудности использования, а также необходимости в высококвалифицированных химиках-аналитиках для управления оборудованием и интерпретацией результатов.

Кроме того, анализ методом GC-MS является затратным по времени и не является методом онлайн-анализа. Отметим, что потеря и деградация аналитов, в частности реакционноспособных или термически лабильных метаболитов, и возможные загрязнения являются важными до сих пор полностью нерешенными проблемами, которые необходимо преодолеть для улучшения качества данных, получаемых в этом виде анализа [10,44,45].

В методе **масс-спектрометрии с реакцией переноса протона (PTR-MS)** используется предварительное формирование реактант-иона H_3O^+ в разряде низкого давления в парах воды в полом катоде и короткой дрейфовой трубке. Затем эти ионы поступают в дрейфовую трубку с постоянным аксиальным полем, на входе в которую вводится анализируемая проба. В конце трубки находится столкновительная ячейка, в которой происходит реакция протонирования аналита (M):

$$H_3 0^+ + M \Leftrightarrow MH^+ + H_2O \quad (1)$$

Далее ионы поступают в масс-спектрометр, как правило, квадрупольный. Для количественного определения аналита используют отношение интенсивности сигнала аналита к интенсивности сигнала прекурсора H_3O^+.

Метод PTR-MS имеет высокую чувствительность: пределы обнаружения в ряде случаев находятся на уровне ppt [46]. Достоинства метода PTR-MS, как и других онлайн-систем, особенно проявляются при определении неустойчивых соединений, в частности альдегидов [46].

Основные проблемы данного метода – частичная фрагментация аналитов, многочисленные интерференции и, как следствие, сложность интерпретации масс-спектров и количественного определения ряда соединений. Кроме того, влажность анализируемого воздуха существенно влияет на чувствительность метода и на относительные интенсивности сигналов фрагментов протонированного аналита [47]. Отметим, что для ряда соединений, например, пропан-1-ола, невозможно использовать его протонированную форму MH^+, поскольку она нестабильна, хотя для многих других соединений возможно детектирование протонированных компонентов MH^+. Одним из главных недостатков метода PTR-MS является ограничение круга определяемых

соединений только теми из них, для которых сродство к протону для MH^+ больше, чем в ионе H_3O^+.

Одним из методов, позволяющих определять содержание ЛОС в ВВ, является **масс-спектрометрия выбранных ионов в потоке (SIFT-MS)**. Этот метод основан на предварительном выделении из смеси компонентов, возбуждаемых во влажном воздухе в радиочастотном разряде, одного из ионов-реактантов - H_3O^+, O_2^+ или NO^+ (с помощью квадрупольного масс-фильтра) – с последующей химической ионизацией широкого круга соединений в дрейфовой трубке и детектированием ионов с помощью масс-спектрометра. По принципу действия метод SIFT-MS близок методу PTR-MS, различие состоит в том, что в методе SIFT-MS используется предварительное выделение одного из ионов-реактантов с помощью масс-фильтра, а в ионном источнике PTR-MS формируется только один реактант-ион – H_3O^+, но подобраны такие условия, что интенсивности других молекулярных ионов значительно ниже. Подобный подход не только упрощает систему, но и позволяет достичь в методе PTR-MS более высокие чувствительности и более низких пределов обнаружения, чем в методе SIFT-MS [47].

В то же время SIFT-MS относится к немногим методам, которые используются для количественного определения ряда потенциальных биомаркеров РЛ, в частности, ацетальдегида, пропан-1-ола, пропан-2-ола, уксусной кислоты, метилформиата, этилбензола, изопрена и др. [8,48]. Особенное внимание было уделено определению ацетальдегида в выдохе и в газовой среде, в которой находятся растущие раковые клетки [8,49,50].

Пределы обнаружения ряда малоатомных ЛОС находятся на уровне единиц ppb, достаточном для определения потенциальных маркеров РЛ. В то же время с помощью метода SIFT-MS до сих пор не получены результаты с приемлемыми уровнями специфичности и чувствительности метода.

Спектрометрия ионной подвижности применяется для анализа ВВ преимущественно в виде варианта с поликапиллярной колонкой **(MCC-IMS)** [51–53]. В результате получаются двумерные «образы» выдоха. С помощью подобного подхода в работе [51] был исследован ВВ 19 пациентов с подтвержденной немелкоклеточной карциномой легкого с различной гистологией, с использованием гибкой бронхоскопии с видеочипами. Всего было зарегистрировано 72 пиков, 5 из которых существенно отличались для легкого с РЛ и здорового легкого. Для аденокарциномы выделялся пик,

соответствующий, по-видимому, димеру н-декана, а для плоскоклеточного рака – бутан-2-олу, или 2-метилфурану, или нонаналю. Чувствительность, специфичность, прогностичность положительного и отрицательного результатов составили для аденокарциномы 100%, 75%, 80% и 100%, для плоскоклеточного рака - 78%, 78%, 80%, 75% (бутан-2-ол) и 78%, 78%, 80%, 88% (нонаналь). Отметим, что предложенную в работе [51] методологию сложно назвать неинвазивной, поскольку в больное и здоровое легкое необходимо вводить зонды, с помощью которых определяется разность интенсивностей различных компонентов, присутствующих в выдохе.

При использовании метода MCC-IMS для диагностики злокачественной плевральной мезотелиомы по ВВ были получены [52] близкие к результатам предыдущей работы величины: чувствительность, специфичность, прогностичность положительного и отрицательного результатов составили 96%, 67%, 76%, 93% соответственно. Отметим, что достигнутый уровень прогностичности положительного результата недостаточен для использования MCC-IMS как единственного метода для проведения скринингового обследования, поскольку потребуется дополнительное обследование весьма значительной части пациентов, у которых отсутствуют онкологические заболевания.

Использование IMS без МСС для диагностики РЛ, так и других заболеваний по выдоху малоэффективно [54], что связано с низкой селективностью метода. Отметим, что важной особенностью системы MCC-IMS является возможность прямого анализа без использования пробоотборных пакетов и ТФМЭ [52].

1.3.2. Методы анализа выдыхаемого воздуха на основе мультисенсорных систем, работающих по принципу распознавания образов

Наряду с методами прямого определения ЛОС одним из перспективных подходов для реализации диагностики РЛ по ВВ на ранних стадиях является использование мультисенсорной системы (МС) типа «электронный нос» (ЭН). Под этим термином понимают компактный и относительно недорогой газоанализатор, состоящий из массива неселективных химических сенсоров и системы распознаванию образов [55]. Принцип работы ЭН заключается в формировании многомерного отклика от массива сенсоров, обладающих различной перекрестной чувствительностью, и последующей обработке отклика с помощью хемометрических методов для получения так называемого образа конкретной газовой смеси, в нашем случае выдыхаемого воздуха. Подобный образ можно

назвать «отпечатком дыхания» по аналогии с отпечатком пальцев. На основе обучающего набора данных, включающих в себе образы выдоха группы пациентов с каким-либо заболеванием и группы пациентов с подтвержденным отсутствием заболевания (контрольной группы), обучается математическая модель-классификатор, позволяющая делать прогноз о принадлежности испытуемого по его «отпечатку дыхания».

Ключевую роль при разработке инструмента диагностики на базе ЭН играет тип сенсоров. Так для исследования возможностей диагностики выявления РЛ используют: сенсоры на проводящих полимерах, пьезоэлектрические кварцевые резонаторы, сенсоры на поверхностно акустических волнах (ПАВ), оптические сенсоры, полупроводниковые металлооксидные (МО) и др. Преимущества и ограничения сенсоров данных типов тесно связаны с различным характером формирования аналитического сигнала. В таблице 2 представлены основные достоинства и недостатки вышеуказанных типов сенсоров [2,56].

Таблица 2. Достоинства и недостатки сенсоров для системы ЭН

Тип сенсоров	**Принцип**	**Достоинства**	**Недостатки**
Пьезоэлектрические кварцевые резонаторы / сенсоры на ПАВ	Изменение резонансной частоты	Высокая чувствительность, быстрый отклик	Сложный процесс изготовления, чувствительность к влажности и температуре, низкая стабильность при высоких температурах
Оптические сенсоры	Изменение оптической плотности, интенсивности флуоресценции, люминесценции	Высокая чувствительность, срок службы	Сложность в миниатюризации, высокая стоимость
Полупроводниковые металлооксидные сенсоры	Изменение сопротивления или проводимости сенсора	Низкая стоимость, время отклика, долговечность, самоочищение	Низкая селективность, относительно высокая потребляемая мощность
Проводящие	Изменение	Низкая стоимость	Время отклика и

полимеры	сопротивления, массы, оптических свойств	изготовления, низкое энергопотребление	релаксации, низкая стабильность, низкая чувствительность, дрейф сигнала
Сенсоры на основе полевых транзисторов	Изменение электрического тока	Высокая адсорбционная способность	Время отклика, низкая чувствительность к ЛОС

1.3.2.1. Проводящие полимеры

Принцип действия газовых сенсоров на проводящих полимерах заключается в изменении сопротивления сенсоров из-за адсорбции газов поверхностью сенсоров [57]. Эти сенсоры работают при температуре окружающей среды и могут быть покрыты различными материалами для повышения чувствительности датчиков к определенным ЛОС [57]. В работе McWilliams и соавт. [57] была исследована возможность ранней диагностики РЛ с помощью системы ЭН Cyranose 320, в котором был использован массив из 32 сенсоров на основе проводящих полимеров. Был проанализирован ВВ 25 пациентов с РЛ (I и II стадии) и группы повышенного риска из 166 активных и бывших курильщиков без РЛ. Результаты показали значимое влияние параметра курения и пола испытуемых: эффективность разделения была выше для бывших курильщиков, чем для активных, по крайней мере в случае аденокарциномы (ROC AUC – площадь под кривой взаимной зависимости вероятностей ложноположительных и истинно положительных результатов. ROC AUC для бывших курильщиков мужчин – 0.846, для бывших курильщиков женщин – 0.816, для активных курильщиков мужчин – 0.745, для активных курильщиков женщин – 0.725). Авторы предполагают, что изменение профиля ЛОС, вызванные активным курением, до некоторой степени маскируют ЛОС, связанные с возникновением опухоли. Причем у мужчин подобные изменения ЛОС вследствие курения выражены более сильно, чем у женщин. Чувствительность и специфичность разработанного метода составили 88.0% и 81.3% соответственно.

1.3.2.2. Сенсоры на поверхностных акустических волнах

В работе Chen и соавт. [58] была использована пара сенсоров на ПАВ. Первый сенсор был покрыт полиизобутиленовой пленкой, а второй был использован в качестве сравнения. Была использована стадия предварительного концентрирования образцов ВВ с помощью ТФМЭ с последующим инжектированием в газохроматографическую капиллярную колонку. Элюированные ЛОС подавали на сенсоры с ПАВ и фиксировали изменение частоты. Полученные данные были проанализированы с использованием искусственной нейронной сети. В результате для общей выборки из 10 пациентов удалось достичь 80% чувствительности и 80% специфичности.

1.3.2.3. Пьезоэлектрические кварцевые резонаторы

Кварцевые резонаторы состоят из кристаллов кварца, покрытых специфичными металлопорфиринами. ЛОС сорбируется на поверхности металлопорфиринов, изменяя массу кристалла и частоту его колебаний. Такие изменения фиксируются и используются для обучения классификаторов. Gasparri и соавт. [59] использовали массив сенсоров на основе кварцевых микровесов, покрытых различными металлопорфиринами, для дискриминации 70 испытуемых с РЛ и 76 пациентов из контрольной группы. Достигнутые чувствительность и специфичность составили 81% и 91% соответственно. При этом была достигнута большая чувствительность к РЛ на стадии I по сравнению со стадиями II-IV (92% и 58% соответственно).

1.3.2.4. Оптические сенсоры

Принцип работы оптических сенсоров основан на изменении оптических характеристик при взаимодействии с ЛОС. В простом варианте через массив таких сенсоров продувается анализируемая проба, в результате чего цвет сенсоров изменяется и после фиксированного времени анализируется полученное изображение сенсоров. В работе Mazzone и соавт. [60] использовали систему, состоящую из 24 одноразовых оптических сенсоров для идентификации пациентов с РЛ и определения гистологического типа. В данном исследовании приняло участие 229 человек: 92 пациентов с РЛ (41 – I-II стадия немелкоклеточного рака, 42 – III-IV стадия) и 137 – контрольная группа с повышенным риском заболевания. Все испытуемые с РЛ были

разделены на группы в соответствии с гистологическим типом рака (аденокарцинома, плоскоклеточный, мелкоклеточный рак), и образцы ВВ каждой из групп по отдельности сравнивали с образцами ВВ контрольной группы. Также была оценена возможность разделения групп пациентов с РЛ ранних (I-II) и поздних (III-IV) стадий и возможности прогнозирования выживаемости. Было показано, что модели, построенные для каждого типа рака отдельно, более точны, чем обобщенная модель. Достигнутые показатели чувствительности и специфичности варьировались в диапазоне 70-91% и 73-95% соответственно в зависимости от гистологического типа. Различия между ранними и поздними стадиями определялись с чувствительностью 81% и специфичностью 93%, а выживаемость (менее 12 месяцев или более 12 месяцев) оценивалась с чувствительностью 70% и специфичностью 86%.

1.3.2.5. Сенсоры на основе полевых транзисторов

Набор полевых транзисторов на основе кремниевых нанотрубок применены в работе Shehada и соавт. [61] для обнаружения и классификации РЛ, рака желудка, бронхиальной астмы и хронической обструктивной болезни легких (ХОБЛ). Общее число испытуемых составило 374 человека. Размер выборки испытуемых с РЛ составил 149 человек, с раком желудка – 40 человек, с астмой или ХОБЛ – 56 человек, а контрольная группа состояла из 129 человек. При этом испытуемые с РЛ и раком желудка были дополнительно разделены на две группы в соответствии со стадией заболевания: ранняя (I и II) и поздняя (III и IV). В результате чувствительность и специфичность для построенных бинарных классификаторов составили: 87% и 82% (РЛ против контрольной группы), 92% и 80% (РЛ против астмы), 97% и 90% (РЛ против рака желудка) соответственно. При этом авторами было отмечено, что способность к разделению группы пациентов с астмой от контрольной группы была довольно низкой (чувствительность – 48%, специфичность – 91%). Авторы связывают данный факт с тем, что для астмы характерен лишь один маркер – пентан, а не набор маркеров, как для онкозаболеваний. При определении стадии заболевания для пациентов с РЛ достигнута 34% чувствительность и 95% специфичность.

1.3.2.6. Полупроводниковые металлооксидные сенсоры

Кондуктометрические газочувствительные металлооксидные (МО) сенсоры наиболее часто используется в ЭН системах из-за их низкой стоимости, стабильности и чувствительности к широкому спектру соединений [62]. В качестве сенсорных материалов наиболее часто используют нанокристаллические оксиды SnO_2, ZnO, WO_3 и др., легированные Pd, Pt или другими катализаторами. Эти оксиды являются широкозонными полупроводниками с проводимостью n-типа. Поверхность МО сенсоров обладает высокими адсорбционными свойствами и реакционной способностью, что обусловлено наличием свободных электронов в зоне проводимости полупроводника, поверхностных и объемных кислородных вакансий, а также активного хемосорбированного кислорода. Сенсоры стабильны на воздухе при нагревании до 500-600 °C и могут быть получены в высокодисперсном состоянии с размером кристаллитов 3-50 нм и величиной удельной поверхности до 100-150 $м^2/г$ [63].

При контакте сенсора с газовой средой, на его поверхности происходит адсорбция атомов и молекул летучих веществ. При этом возможна как физическая адсорбция за счет слабых сил притяжения с энергией связи 0.01-0.1 эВ, так и химическая адсорбция с возникновением химического соединения за счет сил обменного типа с энергией связи порядка 1 эВ [64].

Практически всегда химическая адсорбция является активированной, т.е. частица газа должна затратить энергию на преодоление потенциального барьера, которая затем возвращается в результате акта адсорбции. Активированная адсорбция протекает с меньшей скоростью, которая увеличивается при повышении температуры [65]. В подавляющем большинстве случаев газовые сенсоры работают в воздушной среде, в которой основное влияние на их электрофизические и газочувствительные свойства оказывает адсорбция молекул и атомов кислорода и молекул воды.

Газы-восстановители вступают во взаимодействие с хемосорбированным кислородом, что приводит к снижению плотности отрицательного заряда на поверхности и повышению электропроводности. Заметное изменение величины проводимости сенсора может быть зарегистрировано при наличии аналитов в концентрациях 0.1-10 ppm [63].

Структурные изменения такие как изменения размеров и геометрии МО зерен приводят к изменениям их проводимости и каталитических свойств. Разрушение МО пленки после значительного времени эксплуатации и разделение фаз между оксидом

металла и добавками являются дополнительными факторами, влияющими на стабильность сенсора. Воздействие соединений, способных необратимо связываться с оксидом металла приводит к ингибированию каталитической активности и загрязнению [62,66]. В качестве таких ингибиторов могут выступать азот-, фосфор- и серосодержащие соединения [67].

Так как на практике работа с сенсорами происходит не вакууме, а в воздушной среде, необходимо учитывать, что поверхность полупроводникового сенсора содержит значительное количество хемосорбированного кислорода. В различных температурных режимах можно наблюдать различные формы хемосорбированного кислорода: 80-150ºС – кислород восстанавливается до молекулярного аниона O_2^-, 150-260ºС – дальнейшее восстановление до атомарного аниона O^-, 260-460ºС – анионы O^{2-}. Поэтому для молекул-восстановителей наиболее вероятно взаимодействие с хемосорбированным кислородом, чем независимая адсорбция на поверхности чувствительного слоя [68]. Рабочий температурный диапазон таких сенсоров обычно составляет от 200°С до 600°С.

В стандартном варианте процедуру анализа ВВ можно в принципе разделить на 3 этапа. Сначала через ячейку из МО сенсоров пропускают референтный газ (например, воздух помещения, в котором находится испытуемый), который формирует базовую линию. Далее с помощью крана подается проба ВВ в течение определенного времени. Далее кран снова переключается на референтный газ. На всех этапах при этом регистрируется зависимость проводимости каждого сенсора от времени. Пример такой зависимости проиллюстрирован на рисунке 1.

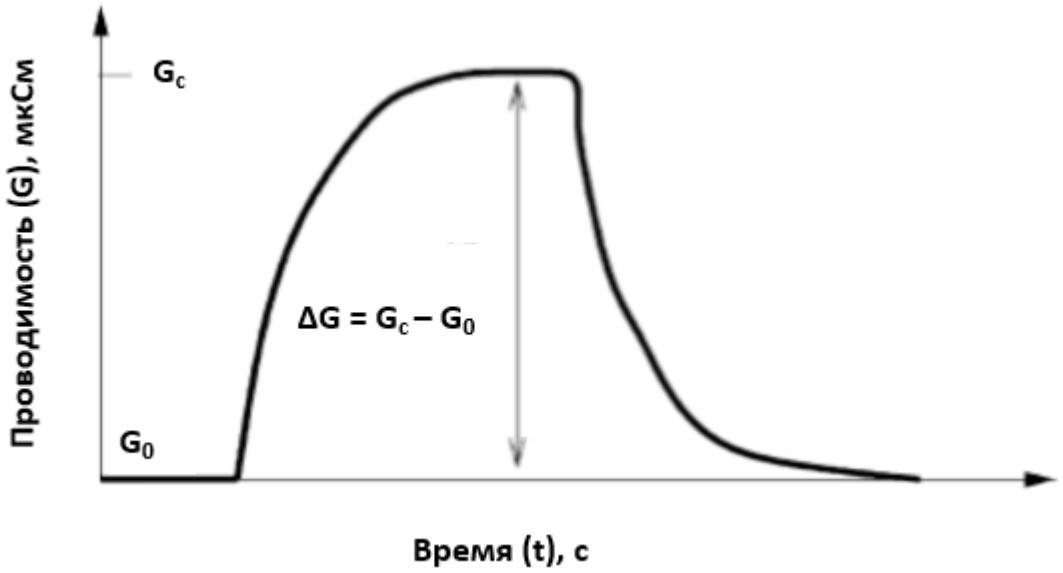

Рисунок 1. Пример зависимости проводимости сенсора (G) от времени (t) при подаче анализируемого газа

Из полученных зависимостей проанализированных образцов могут быть извлечены различные признаки. Наиболее распространено использование $\Delta G/G_0$. Также в качестве информативных признаков используют G_c/G_0, G_{max}, интегралы различных зон, производные 1-го и 2-го порядка, значение проводимости при определенном времени относительно подачи пробы, время достижения определенной доли изменения проводимости.

В таблице 3 содержится информация о работах, в которых была исследована возможность разделения пациентов на группы РЛ и контрольную группу с использованием ЭН на основе МО сенсоров [56,69].

Таблица 3. Сравнение критериев информативности разработанных тестов по диагностике РЛ в пилотных исследованиях с использованием ЭН систем на основе МО сенсоров. Основные критерии разделения группы здоровых и группы РЛ – чувствительность (Se), специфичность (Sp) и точность (Acc).

Характеристика выборки	**Se**	**Sp**	**Acc**	**Работа**
N=101 (43 РЛ, 58 контроль)	95.3%	90.5%	92.6%	[70]
N=89 (16 РЛ, 73 контроль)	-	-	*	[71]

N=18 (9 РЛ, 9 контроль)	100%	88.9%	94.4%	[72]
N=89 (47 РЛ, 42 контроль)	93.6%	83.3%	-	[73]
N=76 (31 РЛ, 45 контроль)	-	-	88%	[74]
N=37 (12 РЛ, 25 контроль)	83%	88%	-	[75]
N=84 (32 РЛ, 52 контроль)	85%	84%	-	[76]
N=290 (144 РЛ, 146 контроль)	94.4%	32.9%	-	[77]
N=145 (52 РЛ, 93 контроль)	83%	84%	-	[78]
N=16 (6 РЛ, 10 контроль)	85.7%	100%	93.8%	[79]

*-чувствительность, специфичность и точность не указаны в работе. При разделении достигнуты следующие уровни значимости: 0.045, 0.025, 0.001 для каждого канала ЭН системы.

Отдельно стоит выделить работы, где были исследованы коммерчески доступные системы анализа ЛОС для диагностики РЛ [74,78]. Например, van de Goor и соавт. [78] использовали пять ЭН систем Aeonose с применением искусственной нейронной сети для классификации пациентов на группу с РЛ и группу здоровых людей (60 и 107 человек соответственно). Результаты показали диагностическую точность 83% при чувствительности 83%, специфичности 84% и ROC AUC 0.84. Были показаны сопоставимые результаты с чувствительностью 88%, специфичностью 86% и диагностической точностью 86%. В другом исследовании группа de Vries и сооавт. [74] использовала SpiroNose в сочетании с оборудованием для исследования функции легких для классификации пациентов на группы с РЛ, ХОБЛ, астмы и группы здоровых пациентов (45 - РЛ, 31 - контроль). Результаты показали, что пациенты с РЛ и здоровые люди из контрольной группы были достаточно хорошо различимы ($p < 0.001$), а точность при перекрестной проверки составила 88% с ROC-AUC 0.95 ± 0.11.

Для анализа ВВ процедура отбора проб имеют первостепенное значение. Согласно обзору, представленному Krilaviciute и др. [20], из 73 исследований, связанных с диагностикой рака легких по ВВ, только в шести из них был реализован режим прямого онлайн-измерения, в то время как в остальных работах использовался предварительный отбор проб. Другими словами, ВВ в большинстве случаев собирается в специальные контейнеры для хранения и транспортировки в аналитические помещения. Кроме того, в большинстве исследований используются дополнительные процедуры сорбции для концентрирования ЛОС [20]. Очевидно, что такие процедуры могут вызвать потерю

соответствующих соединений и загрязнение образца, связанное с материалом сорбента или контейнера для хранения. Таким образом, автономный подход может привести к неконтролируемой систематической неопределенности, увеличивая время анализа [11]. Временной фактор становится особенно важным для скрининговых обследований, поскольку инструменты скрининга должны быть легко доступны для широких групп населения. Онлайн-анализ потенциально может обеспечить более надежные результаты из-за отсутствия процедур предварительной обработки образцов и может стать подходящей основой для создания эффективного метода скрининга на РЛ.

1.4. Методы обработки многомерных данных

При использовании ЭН систем, состоящих из неселективных или частично селективных сенсоров и работающих по принципу распознавания образов для решения задач классификации, основная доля работы по извлечению информации ложится на этап обработки данных. В подавляющем большинстве случаев образующийся набор данных имеет высокую размерность, поэтому для извлечения полезной информации используют методы обработки многомерных данных.

1.4.1. Предварительная подготовка данных

Получаемый аналитический сигнал от массива сенсоров можно представить в виде матрицы **X** размерностью I строк и J столбцов. Строки такой матрицы называют образцами, они нумеруются индексом i, меняющимся от 1 до I. Столбцы называются переменными или признаками (например, отклик сенсора), и они нумеруются индексом j, меняющимся от 1 до J. В зависимости от решаемой задачи для измеряемых образцов известна некоторая зависимая переменная, например, метка о принадлежности к определенной группе или его концентрация компонента в образце. Причем таких переменных может быть несколько. Эту информацию можно представить в виде вектора или матрицы **Y** размерностью I строк (количество образцов) и N столбцов (количество зависимых переменных).

Если отклик сенсора представлен в виде одного значения, то мы имеем двухмодальные данные в виде 2D матрицы **X**, но если отклик сенсора представляет собой

набор значений (например, временная зависимость отклика или отклик, просканированный при варьировании одного из параметров), то набор данных является трехмодальным и представляет собой 3D матрицу (Рисунок 2) [80].

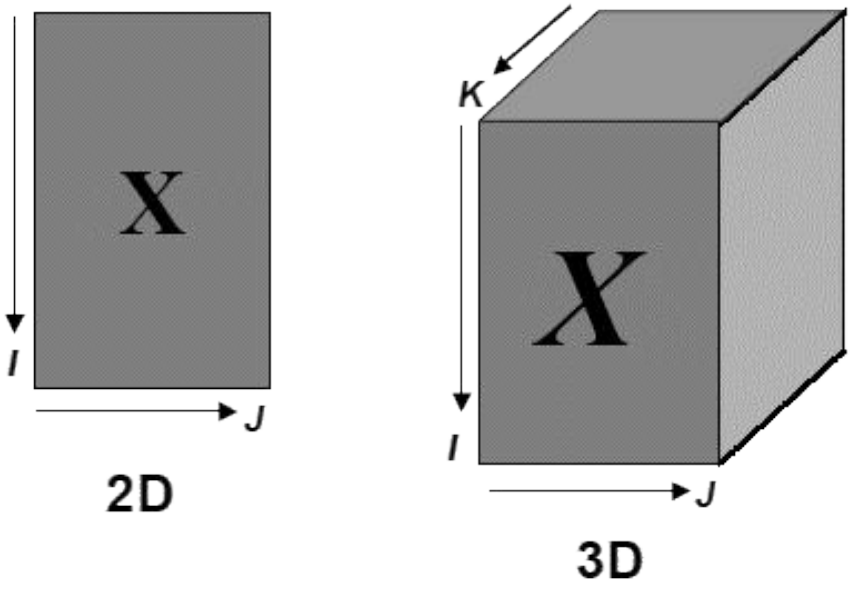

Рисунок 2. Представление двух- и трехмодальных данных

Намного удобнее работать с двухмодальными массивами, поэтому к данным, имеющим структуры 3D матрицы применяют развертку [80]. Таким образом 3D матрица X размерности I×J×K преобразуется в 2D матрицу X размерности I×JK, к которой можно применять методы анализа многомерных данных.

Чтобы заранее избежать трудностей с большим количеством признаков в наборе данных обращаются к различным техникам извлечения признаков, целью которых является получить наиболее информативные признаки с помощью математических преобразований исходной матрицы откликов (сохраняя при этом информацию, относящуюся к целевой величине) [81]. На примере измерения проводимости МО сенсоров в принципе можно использовать набор откликов во времени с последующим применением метода понижения размерности, а можно уже на данном этапе произвести извлечение признаков. Для МО сенсоров чаще всего используют стационарные отклики: R, R/R_0, $(R-R_0)/R_0$. Помимо них также можно использовать интегрирование сигнала или производные 1-го или 2-го порядка. Также в работах в качестве извлеченных признаков

были использованы: время, при котором сигнал достигает определенного отношения, сигнал при определенном времени и другие.

Далее матрица **X** может подвергаться процедурам центрирования и нормирования. При центрировании матрицы **X** из нее вычитается матрицы M, элементы m_{ij} которой равны среднему значению столбца m_j. Данная операция необходима для некоторых проекционных методов таких как метод главных компонент (PCA, principal component analysis).

Нормирование, в отличие от центрирования, не изменяет структуру данных, а просто изменяет вес различных частей данных при обработке. При нормировании по столбцам матрица **X** умножается справа на диагональную матрицу W размерности J×J, диагональные элементы w_{jj} которой равны обратным значениям стандартного отклонения по столбцу x_j. Нормирование данных часто применяют для того, чтобы уравнять вклад в модель от различных переменных [80].

Результаты измерения на мультисенсорных системах часто имеют большое количество переменных (отклики сенсоров и их производные величины), поэтому визуализация данных в простом виде осложнена, если необходимо смотреть на полную картину сразу. Для этих целей применяют многомерный анализ данных с использованием различных методов понижения размерности таких как PCA или линейный дискриминантный анализ (LDA, linear discriminant analysis).

Методы подпространств имеют прочную математическую основу и пользуются популярностью у многих исследователей. Несмотря на то, что PCA и LDA являются наиболее популярными методами, у них есть и свои недостатки. PCA – метод обучения «без учителя». PCA стремится охватить максимальную дисперсию в нескольких измерениях, игнорируя дискриминационную информацию. С другой стороны, LDA – метод обучения с учителем, но предполагает унимодальные нормально распределенные классы с разными средними значениями и равными ковариациями между классами. Кроме того, хорошо известно, что LDA склонен к переобучению, показывая слишком оптимистичные результаты при разделении классов на обучающем наборе для выборок с низким отношением количества образцов к количеству признаков [81].

В методе главных компонент используются новые, формальные переменные t_a (a=1,…A), являющиеся линейной комбинацией исходных переменных x_j (j=1,…J). С

помощью этих новых переменных матрица X разлагается в произведение двух матриц T и P:

$$X = TP^T + E = \sum_{a=1}^{A} t_a p_a^t + E \quad (2)$$

Матрица T называется матрицей счетов. Ее размерность (I×A). Матрица P называется матрицей нагрузок. Ее размерность (J×A). E – это матрица остатков, размерностью (I×J). Новые переменные t_a называются главными компонентами. Число столбцов – t_a в матрице T, и p_a в матрице P, равно A, которое называется числом главных компонент. Эта величина заведомо меньше числа переменных J и числа образцов I. Важным свойством PCA является ортогональность (независимость) главных компонент [80]. Для построения PCA обычно используется алгоритм NIPALS (от англ. nonlinear iterative partial least square) или разложение по сингулярным значениям.

1.4.2. Методы, используемые для решения задач классификации

Принцип решения задач классификации основан на построении моделей, т.е. набора правил, по котором новый образец может быть отнесен к какому-либо классу. Построение или обучение модели проводят на основе обучающего набора образцов с имеющейся априорной информацией о принадлежности к классам (например, класс больных и здоровых). В работах с использованием систем ЭН наиболее часто [82] используют: метод k ближайших соседей (**kNN**, k nearest neighbors) [83], логистическую регрессию (**LR,** logistic regression) [84], метод опорных векторов (**SVM**, support vector machine) [85], метод «случайного леса» (**RF**, fandom forest), состоящий из ансамбля решающих деревьев [86].

kNN. Самый простой метрический метод в задаче классификации — метод k ближайших соседей kNN, суть которого заключается в том, что объект относится к тому классу, к которому принадлежит большинство из его k ближайших соседей. Мера близости задается функцией расстояния. В классическом kNN используется евклидова метрика. Для двух точек $x_1 = (x_{11}, x_{12}, \dots, x_{1j})$ и $x_2 = (x_{21}, x_{22}, \dots, x_{2j})$ евклидово расстояние определяется следующим образом:

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n} (x_{1j} - x_{2j})^2} \quad (3)$$

Также в попытках повысить точность классификации иногда используют взвешенный вариант kNN, при котором во внимание принимается не только количество попавших в область определённых классов, но и их удалённость от нового объекта.

RF. RF — алгоритм машинного обучения [87], заключающийся в использовании комитета (ансамбля) решающих деревьев. Чтобы построить случайный лес из N решающих деревьев, необходимо:

1) Сгенерировать N случайных подвыборок с повторениями X_n, n = 1,…,N.

2) Каждую получившуюся подвыборку X_n использовать в качестве обучающей выборки для построения соответствующего решающего дерева $b_n(x)$. Причем:

- Дерево строится, пока в каждом листе окажется не более n_{min} объектов. Очень часто деревья строят до конца (n_{min} = 1), чтобы получить сложные и переобученные решающие деревья с низким смещением.
- Процесс построения дерева рандомизирован: на этапе выбора оптимального признака, по которому будет происходить разбиение, он ищется не среди всего множества признаков (J), а среди случайного подмножества размера q < J. Причем подмножество размера q выбирается заново каждый раз, когда необходимо разбить очередную вершину. Выбор наилучшего из этих q признаков может осуществляться с помощью критерия информативности. В основном используются критерий информативности Джини или энтропийный информативный критерий.
- Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, и объекту присваивается класс, за который проголосовало наибольшее число деревьев:

$$a(x) = sign\frac{1}{N}\sum_{n=1}^{N} b_n(x) \quad (4)$$

LR. Логистическая регрессия – это метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам. При условии, что метки классов принимают значения $Y = \{-1, +1\}$ в методе LR строится линейным алгоритм классификации a: $X \to Y$ вида:

$$a(x, w) = sign\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = sign\langle x, w\rangle \quad (5)$$

где w_j – вес j-го признака, w_0 – порог принятия решения, $w = (w_0, \dots, w_n)$ – вектор весов, $\langle x, w\rangle$ – скалярное произведение признакового описания объектов на вектор весов.

Предполагается, что искусственно введён нулевой признак: $f_0(x) = -1$. Таким образом, задача обучения линейного классификатора заключается в том, чтобы по выборке X^m настроить вектор весов w. Для этого в методе LR решается задача минимизации эмпирического риска с функцией потерь специального вида:

$$\mathrm{Q}(w) = \sum_{i=1}^{m} \ln\left(1 + \exp(-y_i\langle x_i, w\rangle)\right) \to \min_w \tag{6}$$

SVM. Метод опорных векторов является одним из наиболее популярных методов обучения для решения задач классификации и основан на построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Пусть имеется набор объектов в X пространстве $\mathbb{R}^n$ с соответствующими метками класса $Y = \{-1, +1\}$. Необходимо построить алгоритм классификации $a(x) = X \to Y$. Пусть мы имеем линейно разделимую выборку и существует некоторая гиперплоскость, разделяющая классы -1 и +1. В таком случае в качестве алгоритма классификации воспользуемся линейным пороговым классификатором:

$$a(x) = sign(\langle w, x\rangle - b) = sign(\sum_{i=1}^{l} w_i x_i - b) \tag{7}$$

где $x = (x_1, \dots, x_n)$ – вектор значений признаков объекта, а $w = (w_1, \dots, w_n)\ \epsilon\ \mathbb{R}^n$ и $b\ \epsilon\ \mathbb{R}^n$ – параметры гиперплоскости. Для единственности решения в методе SVM строят ту гиперплоскость, которая максимизирует отступ между классами. Для линейного классификатора отступ определяется уравнением

$$M_i(w, b) = y_i(\langle w, x_i\rangle - b) \tag{8}$$

и характеризует то, насколько объект «близок» своему классу. Чем меньше M_i, тем ближе объект x_i к разделяющей гиперплоскости и тем выше вероятность ошибки. Соответственно, отрицательный отступ M_i говорит о том, что алгоритм $\mathrm{a}(x)$ допускает ошибку на объекте x_i.

Далее для удобства вводится нормировка для уравнения гиперплоскости $\langle cw, x\rangle - cb = 0$ таким образом, чтобы $min\ M_i(w, b) = 1$. Это позволяет ограничить разделяющую полосу между классами $\{x: -1 < \langle w, x_i\rangle - b < 1\}$, внутри которой не может лежать ни один объект обучающей выборки.

Чтобы разделяющая гиперплоскостью отстояла от точек выборки как можно дальше, ширина полосы должна быть максимальной. Пусть x_- и x_+- две произвольные точки классов -1 и $+1$, лежащие на границе полосы, т.е. их отступ равен единице. Тогда

ширину разделяющей полосы можно выразить как проекцию вектора $x_+ - x_-$ на нормаль к гиперплоскости w.

$$\frac{\langle x_+ - x_-, w\rangle}{\|w\|} = \frac{\langle x_+, w\rangle - \langle x_-, w\rangle - b + b}{\|w\|} = \frac{M_+(w,b) - M_-(w,b)}{\|w\|} = \frac{2}{\|w\|} \tag{9}$$

А чтобы разделяющая гиперплоскость находилась на наибольшем расстоянии от точек выборки, ширина полосы должна быть максимальной:

$$\frac{2}{\|w\|} \to \max \quad \Rightarrow \|w\| \to \min \tag{10}$$

Это приводит нас к постановке задачи оптимизации в терминах квадратичного программирования:

$$\begin{cases} \|w\|^2 \to \min\limits_{w,b} \\ M_i(w,b) \geq 1, \;\; i = 1, \dots, l \end{cases} \tag{11}$$

Для обобщения SVM на случай линейно неразделимой выборки, разрешим алгоритму допускать ошибки на обучающих объектах, но так, чтобы их количество было минимальным. Для каждого объекта отнимем от отступа некоторую положительную величину ξ_i, но потребуем, чтобы введенные поправки были минимальны. Данные изменения приведут к следующей постановке задачи, называемом SVM с мягким отступом:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \to \min\limits_{w,b,\xi} \\ M_i(w,b) \geq 1 - \xi_i, \;\; i = 1, \dots, l \\ \\ \xi_i \geq 0, \;\; i = 1, \dots, l \end{cases} \tag{12}$$

Так как у нас нет информации о том, какой из функционалов $\frac{1}{2}\|w\|^2$ и $C\sum_{i=1}^{l}\xi_i$ важнее, вводится коэффициент C, который оптимизируется с помощью перекрестной проверки. В итоге получаем задачу, у которой всегда есть единственное решение.

В случае, когда количество классов больше двух на практике такую задачу обычно разбивают на несколько бинарных задач классификации типа «один против остальных» (One-vs-Rest) или «один против другого» (One-vs-One). Однако мультиклассовый метод опорных векторов (MSVM, multiclass SVM), предложенный в работе Краммера и Зингера [88], позволяет свести задачу мультиклассовой классификации к одной задаче оптимизации, без необходимости ее разбиения на несколько задач бинарной классификации.

1.4.3. Способы оценки результатов классификационных и регрессионных моделей

Для оценки качества проверяемого диагностического теста необходима информация о наличии или отсутствии болезни, полученная эталонным диагностическим тестом, или т.н. «золотым стандартом». Это тест или комбинация тестов, позволяющая надежно определяет наличие или отсутствие болезни у пациента.

Проверяемый диагностический тест может выдавать положительный (пациент имеет болезнь) или отрицательный (пациент здоров) результат для исследуемого пациента. Тогда результат применения бинарного диагностического теста к группе пациентов с учетом проверки золотым стандартом можно представить в виде таблицы, состоящей из 4-х групп исходов: истинно положительные (true positive, TP), истинно отрицательные (true negative, TN), ложноположительные (false positive, FP) и ложноотрицательные (false negative, FN). Такую таблицу еще называют таблицей сопряженности или матрицей ошибок (таблица 4).

Таблица 4. Таблица сопряженности результатов проверяемого диагностического теста

		Результат золотого стандарта	
		1	0
Результат предсказания	1	TP	FP
	0	FN	TN

Диагностическая эффективность теста или **точность (Acc)** определяется как доля истинных результатов среди всех результатов теста:

$$Acc = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{13}$$

Чувствительность (Se) определяется как вероятность получения позитивного исхода для субъекта с болезнью:

$$Se = \frac{TP}{(TP+FN)} \tag{14}$$

Специфичность (Sp) определяется как вероятность получения негативного исхода для субъекта без болезни:

$$Sp = \frac{TN}{(TN+FP)} \tag{15}$$

Оценка чувствительности и специфичности важна при выборе теста для его применения в определенных клинических целях. Чувствительность теста отражает вероятность его положительного результата в присутствии патологии. Высокая чувствительность теста позволяет с его помощью выявлять больных в общей популяции. Специфичность теста отражает вероятность отрицательного результата в отсутствие патологии, что при высокой специфичности позволяет отсеивать здоровых из популяции с предполагаемой патологией. Комбинация клинической чувствительности и клинической специфичности теста характеризует клиническую эффективность теста.

При интерпретации результатов лабораторных тестов вероятность действительного наличия патологии при положительном результате или надежность исключения патологии при отрицательном результате оценивается на основе определения предсказательной ценности положительных или отрицательных результатов тестов.

Прогностичность положительного результата (PPV) определяется как вероятность наличия болезни у субъекта с положительным исходом:

$$PPV = \frac{TP}{(TP+FP)} \tag{16}$$

Прогностичность отрицательного результата (NPV) определяется как вероятность отсутствия болезни у субъекта с положительным исходом:

$$NPV = \frac{TN}{(TN+FN)} \tag{17}$$

Если рассматривать в качестве выходного значения классификатора не метку класса, а вероятность принадлежности, например к классу 1, то при варьировании порога, по которому определяется принадлежность пациента к группе здоровых или больных, можно получить множество матриц сопряженности с различными значениями чувствительности и специфичности. Для установления оптимального порога, а также для сравнительного анализа эффективности алгоритмов классификации используется кривая оперативной характеристики (receiver operating characteristic, ROC-кривая), т.е. кривая взаимной зависимости вероятностей истинно положительных результатов равных чувствительности и ложноположительных результатов равных единице минус специфичность при всевозможных значения порога классификации. ROC-кривая является графическим представлением полного спектра чувствительности и

специфичности, поскольку на ней могут быть отображены все возможные пары "чувствительность-специфичность" для конкретного теста (Рисунок 3).

В зависимости от порогового значения и от распределения предсказанных классификационным алгоритмом вероятностей для исследуемой выборки пациентов ROC-кривая имеет разную форму и разное положение. Желательное соотношение между чувствительностью и специфичностью теста достигается выбором точки разделения. Наиболее четкое разграничение между больными и здоровыми обследуемыми достигается при использовании тестов, которые имеют характеристическую кривую результатов, сдвинутую в сторону левого верхнего угла графика.

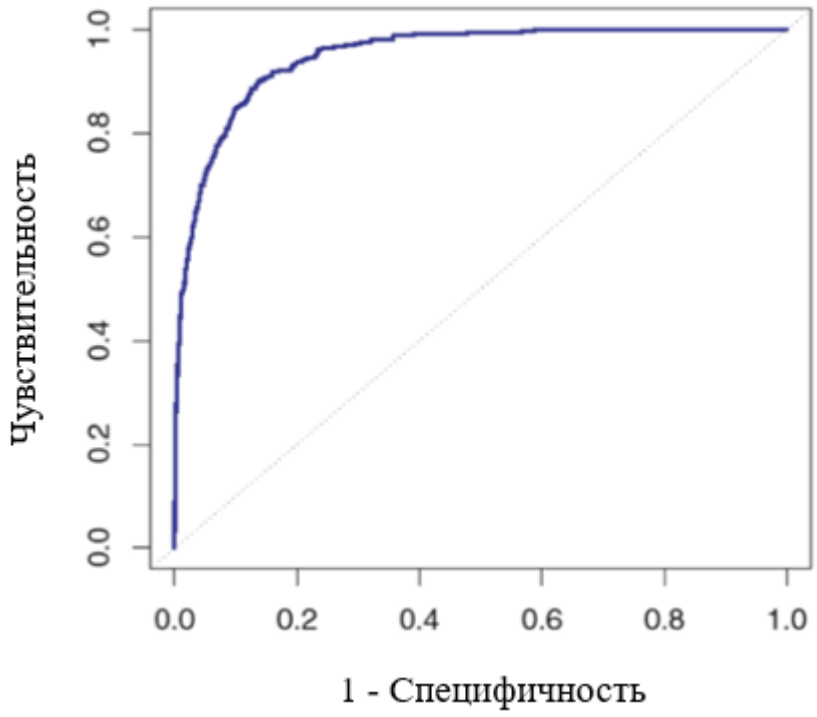

Рисунок 3. Пример ROC-кривой

Для идеального теста график проходит через верхний левый угол, где доля истинно положительных тестов составляет 100% или 1 (идеальная чувствительность), а доля ложноположительных равна 0 (идеальная специфичность). Поэтому чем ближе кривая к верхнему левому углу, тем выше диагностическая эффективность (точность) теста, и наоборот, чем меньше изгиб кривой и чем ближе она расположена к прямой, проходящей под углом 45°, тем менее эффективно диагностическое исследование. Точки на такой диагонали соответствуют отсутствию диагностической эффективности.

Методом оценки ROC-кривых является оценка площади под кривыми (ROC AUC, area under curve ROC). Теоретически площадь изменяется от 0 до 1.0, однако поскольку диагностически полезные тесты характеризуются кривой, расположенной выше положительной диагонали (Рисунок 3), то обычно говорят об изменениях от 0.5 (отсутствие диагностической эффективности теста) до 1.0 (максимальная эффективность теста). Эта оценка может быть получена непосредственно вычислением площади под многогранником, ограниченным справа и снизу осями координат и слева вверху - экспериментально полученными точками. При визуальной оценке ROC-кривых расположение их относительно друг друга указывает на их сравнительную эффективность. Кривая, расположенная выше и левее, свидетельствует о большей диагностической эффективности соответствующего теста.

Основным критерием, использующимся для оценки регрессионной модели, является **среднеквадратичная ошибка** (root mean square error, RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (18)$$

где $\hat{y}_i$- прогнозируемое значение параметра с помощью регрессионной модели, y_i – априорно известное значение параметра, n – число образцов. Заметим, что единицы измерения RMSE и оцениваемого параметра *y* совпадают.

Оценка прогнозирующей силы математических методов с помощью вышеупомянутых метрик качества оптимально проводить методом **проверки на тестовом наборе**. В таком варианте начальный набор данных разделяется на обучающий и тестовый наборы. Первый используется для настройки параметров модели, а второй – для оценки качества полученной модели. Однако в случаях, когда набор данных невелик, используют метод **перекрестной проверки**. Исходный набор данных разбивается на k непересекающихся одинаковых по объему частей. Далее производится k итераций, на каждой из которых модель обучается на k-1 части начального набора (обучающая выборка), а модель тестируется на одной части начального набора (контрольная выборка), которая еще не участвовала в обучении. Результатом является среднее значение метрики качества по всем разбиениям на контрольных выборках. При k равным числу образцов n можно получить вариант **перекрестной проверки по отдельным объектам**.

Отметим важность способа оценки прогнозирующей силы математических моделей, будь это классификационная и регрессионная модель. Все вышеупомянутые

метрики качества необходимо оценивать на тестовом наборе, который не был использован для обучения модели.

1.4.4. Методы переноса градуировочных зависимостей для стандартизации пары сенсорных систем

Помимо достижения высоких метрик клинической информативности теста для внедрения сети таких систем существует еще одно препятствие, заключающееся в ограниченном сроке работы многомерных градуировочных моделей. Перестроение полноценной градуировки является затратной и трудоемкой процедурой из-за большого количества стандартных образцов. Причины, по которой градуировка может становиться непригодной для использования, могут быть разные: временной дрейв сигнала или изменение характеристик сенсора в процессе эксплуатации, матричные эффекты, изменение внешних параметров окружающей среды (температура, влажность). Также случаются ситуации, при которых необходима замена сенсора или перенос градуировочной модели с одного массива сенсоров на другой. В качестве решения описанных проблем используются методы переноса градуировочных зависимостей [89]. В данной работе основное внимание уделено переносу градуировочной зависимости именно между двумя массивами МО сенсоров т.к. параметры окружающей среды (температура и влажность) в принципе можно контролировать с помощью дополнительных инструментов (термостатирование газовой ячейки и установки увлажнителя), а влияние временного дрейфа для МО сенсоров находится на уровне воспроизводимости извлекаемых сигналов.

Методы переноса градуировочной зависимости направлены на корректировку нового измеренного набора данных устраняя новую дисперсию данных, связанных с различием характеристик соответствующих сенсоров. Для этого устанавливается взаимосвязь между двумя экспериментальными условиями и данные, измеренные в новых условиях, корректируются в соответствии с установленным соотношением и используются для предсказания концентраций или в классификационных задачах. Небольшой набор стандартных образцов, измеряемых на обоих массивах, называется набор для переноса градуировки и используется для установления связи между двумя мультисенсорными системами. Существует несколько подходов по реализации переноса градуировочной зависимости: через коррекцию выходного сигнала (концентрации, метки

класса), предсказанных на новом массиве сенсоров, через корректировку измерений, полученных на новом массиве сенсоров (стандартизация сигнала) или корректировки предсказательной модели нового массива. [67]. Полноценный обзор методов переноса градуировочных зависимостей проделаны в следующих работах [90,91].

Стандартизация сигнала является наиболее распространенным среди методов переноса градуировок. При этом так как конечной задачей данной работы является именно классификация, то наиболее предпочтительным будет использование коррекции набора данных, полученных на другом массиве. Отметим, что данный подход позволит объединить результаты измерений образцов, получаемых на нескольких приборов, в единую базу данных и на ее основе строить классификационную модель для диагностики [67].

Стандартизацию осуществляют, используя отношение между откликами сенсора, полученных для стандартизационных образцов на главном и на втором массиве (который необходимо стандартизировать) сенсоров для последующей корректировки измерений для неизвестных образцов. Методы стандартизации данных можно разделить по способу установления связи между двумя наборами данных сигналов сенсоров. Одномерная прямая стандартизация (UDS, Univariate Direct Standardization,) использует отношение между каждым каналом отдельно. Фрагментарно-прямая стандартизация (PDS, Piecewise Direct Standardization) использует отношение между группой сигналов и прямая стандартизация (DS, Direct Standardization) – между всеми сигналами. UDS и PDS - линейные методы и учитывают линейный сдвиг сигнала сенсора. PDS был предложен как улучшение варианта UDS с предположением, что спектральные сигналы для смежных длин волн имеют высокую корреляцию. Однако отклики сенсоров могут быть независимы или частично скоррелированы в зависимости от наполнения и расположения сенсоров в массиве, а также аналитов, что вызывает сомнение в эффективности использования PDS для МО сенсоров. Учитывая данный факт, для МО сенсоров наиболее распространено использование метода прямой стандартизации, учитывающего отклики всех сенсоров. Связь между двумя наборами данных может быть установлена с помощью различных многомерных подходов: множественной линейной регрессии, метода регрессии на латентные структуры – 2 (PLS2, Projection to Latent Structures Regression) и др. [67].

В работе [92] для переноса градуировочной зависимости между двумя идентичными массивами, состоящих из 6 МО сенсоров каждый, была применена робастная регрессия, устойчивая к наличию выбросов. В качестве градуировочной модели была использована искусственная нейронная сеть с обратным распространением ошибки на полученном наборе данных при измерении газовых образцов, отобранных на предприятиях целлюлозно-бумажной промышленности. Для переноса использовали набор данных для стандартизации, состоящий из 27 газовых смесей из сероводорода, диметилсульфида, диметилдисульфида и метилмеркаптана.

В работе [93] на 5 идентичных массивах с 8 МО сенсорами проводили перенос градуировочной зависимости с основного массива на остальные. Сначала были измерены сигналы для 4 соединений (этанол, этилен, оксид углерода (II) и метан) на 10 концентрационных точках. Для построения градуировки и проверки было использовано по 20 образцов. Набор для стандартизации состоял из 8 образцов: по 2 концентрационные точки для 4 соединений. Для переноса градуировочных зависимостей использовали методы DS и PDS на основе PLS2, ортогональной коррекции сигнала и взвешенном методе наименьших квадратов.

Исследователи отмечают, что эффективность стандартизации данных во многом зависит от самих данных и от набора стандартных образцов [67,94]. При стандартизации данных желательно использовать небольшое количество стандартных образцов для переноса ввиду трудоемкости процедуры их приготовления и измерения. С другой стороны количества стандартных образцов должно быть достаточно для описания дисперсии между двумя наборами данных для эффективного переноса градуировки. В некоторых случаях выбор образцов проводят на основе имеющихся знаний об аналите и поставленной задаче. Помимо ручного отбора также используют алгоритм Кеннарда-Стоуна [95]. Данный алгоритм в основном используется для равномерно распределенных образцов в пространстве признаков и заключается в последовательном отборе образца, который наиболее удален от выбранных ранее. В качестве начального состояния отбираются 2 образца наиболее удаленных друг от друга. В качестве меры в основном используется Евклидово расстояние. Также для исследовательских целей по оценке результативности методов стандартизации откликов в среднем может быть использован многократный случайный отбор образцов для стандартизации.

В данной работе для оценки эффективности переноса градуировки были использованы следующие методы: классический метод однофакторной стандартизации (UDS), метод однофакторной стандартизации без использования коэффициента свободного члена регрессии (UDSwoi), метод прямой стандартизации на базе регрессии на латентные структуры (DS-PLS2) и метод прямой стандартизации с использованием L1 регуляризатора (DS-L1R).

Глава 2. Используемые методы исследования и приборы

2.1. Описание характеристик сенсоров

В данной работе были использованы три мультисенсорные системы (МС 1, МС 2.1, МС 2.2). Система МС 1 была использована для задачи оптимизации набора сенсоров, процедуры проведения анализа ВВ на модельных газовых смесях и в дальнейшем в медицинском исследовании. Системы МС 2.1 и МС.2.2 были использованы для исследования возможности совмещения откликов, то есть объединения базы данных и использования модели классификации одного прибора.

Согласно проведенному обзору научных работ, в которых проводились медицинские исследования на пациентах было обнаружено, что для разработки моделей диагностики заболевания используется в среднем не более 10 сенсоров. Предположительно, увеличение количества сенсоров в системе как минимум не увеличивает, а то и понижают информативность из-за высокой коллинеарности в откликах сенсоров. Поэтому для достижения максимальной информативности мультисенсорной системы необходимо использовать сенсоры с как можно большим различием в перекрестной чувствительности. Каждая мультисенсорная система состояла из 6 МО газочувствительных полупроводниковых сенсоров, изготовленных золь-гель методом в лаборатории прикладной химической физики НИЦ «Курчатовский институт» Описание сенсоров приведено в таблице 5. Каждый сенсор представляет собой многослойную систему, состоящую из полупроводникового газочувствительного слоя (сенсорный слой), диэлектрической основы из Al_2O_3 и нагревательного слоя (нагреватель). Сенсорный слой и нагреватель нанесены по разные стороны подложки методом трафаретной печати. Полупроводниковый газочувствительный слой содержал наночастицы SnO_2 с различными допантами Pt, Pd или La [96].

Таблица 5. Состав сенсоров для МС 1, МС 2.1 и МС 2.2

Обозначение сенсора	**Состав чувствительного слоя**
S1, S2	SnO_2 с добавками Pt (3%) и Pd (1%), Cd
S3, S4	SnO_2 с добавками Pt (3%) и Pd (1%)
S5, S6	SnO_2 с добавками Pt (3%) и Pd (1%), La

* - S1 и S2, S3 и S4, S5 и S6 обладают одинаковым качественным составом, но имеют различные сопротивления нагревателей, вследствие чего при равном поданном напряжении нагреватели имеют различные температуры

В зависимости от температуры нагревателя, и, как следствие, температуры сенсорного слоя, относительная чувствительность сенсоров к разным группам соединений различна [97]. Таким образом, информативность данных об анализируемых пробах можно повысить с использованием нескольких температурных режимов. Известно [98], что реакция сенсора на определенный газ имеет максимум при определенной температуре. Например, сенсор на основе SnO_2 с добавкой Pd имеет максимальную чувствительность к H_2 при температуре около 200°C, к пропану – при 350°C, к метану – при 450°C и т.д. Эти максимумы не ярко выраженные, и чувствительность к определенному газу наблюдается в относительно большом диапазоне температур. В экспериментах, проводимых в данной работе, использовались три температурных режима, подобранных эмпирическим путем и распределенные по рабочему температурному диапазону. Также выбранные температуры располагаются на участках, характерных для максимумов чувствительности к веществам, которые легко окисляются (например, спирты, кетоны), относительно легко окисляются (длинноцепочечные алканы), а также к газам, окисляющимся с относительной трудностью (пропан). Поверхность сенсоров не загрязнена продуктами разложения некоторых соединений, поскольку взаимодействие этих продуктов с кислородом на каталитической поверхности при высокой температуре приводит к полному окислению адсорбированного соединения и продуктов разложения. Фактически, разложение целевого соединения на каталитической поверхности является частью отклика сенсора (то есть процесса каталитического окисления хемосорбированным кислородом). Например, длинноцепочечные углеводороды, которые окисляются на поверхности с

разрывом углеродной цепи, дают чувствительность сенсора при гораздо более низкой температуре (ниже 300°C), чем метан (около 450°C).

Для достижения требуемой чувствительности к исследуемым аналитам (биомаркерам легочной онкопатологии) были выбраны 3 температурных режима работы сенсоров T1, T2 и T3. Для нагревания сенсорного слоя в состав микрочипа входит слой микронагревателя, сформированный из платиносодержащей суспензии. Нагреватель использовался одновременно как термистор ввиду строго линейной зависимости его сопротивления RH от температуры. Температурный коэффициент нагревателей α = 0.0027 ºC^{-1} являлся постоянной величиной на всем диапазоне рабочих температур от 150 до 600 ºC. Для нагрева каждого i-го сенсора, на его нагреватель подавалось постоянное напряжение Ut_0 через ограничительный резистор R_{0i}. Измеряя напряжение на нагревателе UH_i, можно рассчитать его сопротивление RH_i по уравнению:

$$RH_i = \frac{R_{0i} \times UH_i}{Ut_0 - UH_i} \quad (19)$$

Если сопротивление нагревателя i-го сенсора при температуре 20 ºC обозначить RH_i^{20}, то температурный коэффициент нагревателя: $\alpha = (RH_i - RH_i^{20}) \,/\, (RH_i^{20} \times (T - 20)$, где RH_i – сопротивление нагревателя при температуре T. Тогда температура нагревателя рассчитывается по следующей формуле:

$$T_i\,(°C) = \frac{RH_i - RH_i^{20}}{RH_i^{20} \times \alpha} + 20 \quad (20)$$

В таблицах 6, 7 и 8 представлены расчетные значения температур нагревателей для 6 сенсоров при температурных режимах T1, T2 и T3, соответствующих трем различным напряжениям нагрева Ut_0 4.48, 4.98 и 5.48 В для МС 1 и 2.67, 3.01 и 3.35В для МС 2.1 и МС 2.2 соответственно. Измерения сопротивления и напряжения проводили с использованием мультиметра модели DT-832 (Дадитс, Китай).

Таблица 6. Температуры нагревателей сенсоров для трех температурных режимов T1, T2 и T3 для МС 1 (максимальная относительная погрешность определения температуры нагревателя составила 13.8%, 13.7% и 12.6% для T1, T2 и T3 соответственно)

Сенсор	R_{0i}, Ом	RH_i^{20}, Ом	Ut_0, мВ		
			4480	4980	5480
			T, ^{0}C		
S1	33.0	7.9	360	409	464
S2	32.6	8.3	325	377	428
S3	32.6	10.2	360	411	462
S4	32.9	14.1	473	534	602
S5	32.8	11.4	444	502	560
S6	32.8	12.6	392	448	502
			T1	**T2**	**T3**

Таблица 7. Температуры нагревателей сенсоров для трех температурных режимов T1, T2 и T3 для МС 2.1 (максимальная относительная погрешность определения температуры нагревателя составила 14.3%, 12.1% и 11.4% для T1, T2 и T3 соответственно)

Сенсор	R_{0i}, Ом	RH_i^{20}, Ом	Ut_0, мВ		
			2670	3010	3350
			T, ^{0}C		
S1	10.0	10.4	334	385	436
S2	10.0	10.6	335	387	438
S3	10.0	11.5	362	417	473
S4	10.0	10.8	337	390	442
S5	10.0	10.5	344	399	452
S6	10.0	11.2	345	401	455
			T1	**T2**	**T3**

Таблица 8. Температуры нагревателей для трех температурных режимов T1, T2 и T3 для МС 2.2 (максимальная относительная погрешность определения температуры нагревателя составила 14.2%, 12.0% и 11.3% для T1, T2 и T3 соответственно)

Сенсор	R_{0i}, Ом	RH_i^{20}, Ом	Ut_0, мВ		
			2670	3010	3350
			T, ^{0}C		
S1	10.0	11.1	349	402	455
S2	10.0	10.6	337	390	442
S3	10.0	11.5	343	397	450
S4	10.0	11.1	352	408	463
S5	10.0	11	350	407	461
S6	10.0	10.7	347	403	457
			T1	**T2**	**T3**

Как видно из таблиц 6, 7 и 8 сенсоры одинакового состава (S1 и S2, S3 и S4, S5 и S6) за счет различий в нагревателях имели разные температуры при одинаковых порядковых температурных режимах, и из-за этого обладали разными газочувствительными свойствами. Таким образом, при измерении показаний откликов 6 сенсоров S1, …, S6 при трех температурных режимах T1, T2, T3 получалось 18 информативных признаков: S1_T1, S2_T1, …, S5_T3, S6_T3.

На рисунке 4 представлена электрическая схема, по которой измеряли проводимость G (См) каждого сенсора $G_i = 1/RS_i$, где RS_i – сопротивление i-го сенсора, Ом. На каждый сенсор подавалось стабилизированное напряжение $US_0 = 5В$ (4950 мВ) и измерялось выходное напряжение преобразователя ток-напряжение US_i, пропорциональное току сенсора: $US_i = US_0 \times RB_i/RS_i$, где RB_i – сопротивление резистора смещения. Проводимость рассчитывалась по следующей формуле:

$$G_i = \frac{US_i}{US_0 \cdot RB_i} \tag{21}$$

На рисунке 5 представлена пример интерфейса программы, отображающего зависимость выходных напряжений сенсоров от времени при последовательном пропускании пробы через ячейку сенсоров.

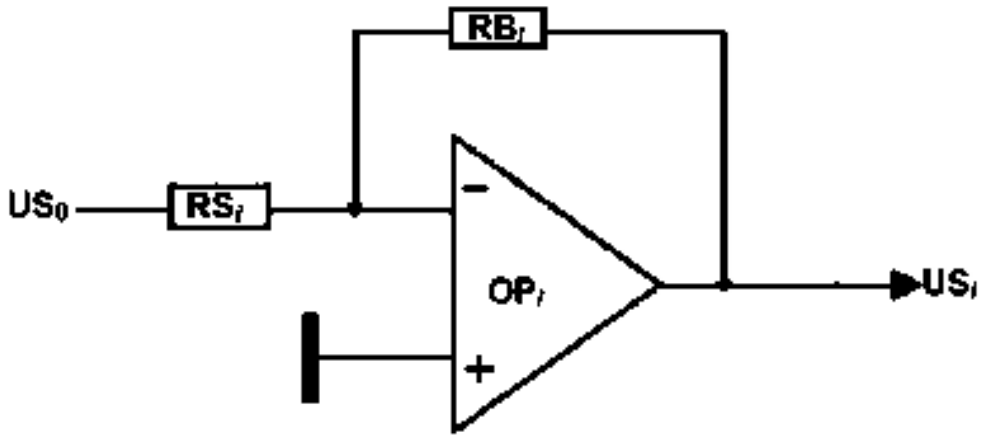

Рисунок 4. Схема измерения проводимости каждого сенсора

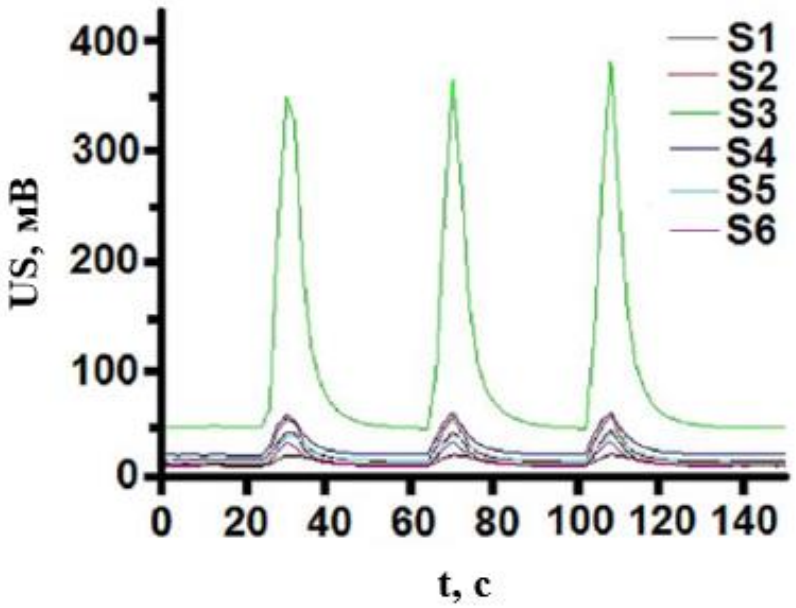

Рисунок 5. Зависимость выходных напряжений US_i для 6 сенсоров при последовательной подаче 3-х проб

Напряжения US_i и UH_i для всех 6 сенсоров преобразовывались с помощью многоканального 24 бит сигма-дельта АЦП и регистрировались на компьютере с частотой 0.25 Гц для МС 1 и 1.0 Гц для МС 2.1 и МС 2.2 в виде таблицы и графиков. Время установления стабильных показаний при переключении температурного режима составляло от 3 до 7 минут.

2.2. Методика приготовления модельных газовых смесей

В данной работе для проведения экспериментов были использованы однокомпонентные градуировочные воздушные газовые смеси воздух – н-гептан, воздух – пропан-1-ол, воздух – этилбензол, воздух – о-ксилол, а также трехкомпонентные смеси (воздух – н-гептан, пропан-1-ол, о-ксилол), которые были приготовлены в соответствии с подходами, описанными в ГОСТ Р ИСО 6144-2008. Газовые смеси готовили с использованием шприца путем введения известного объема жидких целевых компонентов в специальный пакет с известным объемом, наполненный фоновым газом (комнатный воздух).

Пакеты

Для приготовления газовой смеси в качестве смесительной камеры и одновременно сосуда для хранения были использованы тедларовые пакеты фирмы Restek® (США) на 1, 3 и 5 литров. Пакеты оборудованы специальной мембраной, позволяющей ввести жидкость целевого компонента внутрь. Клапан пакета выполнен из инертного полипропилена и имеет как удобный вход для соединения с трубкой пробоотборника, так и возможность ввода пробы с помощью шприца через мембрану.

Шприц

Для отбора жидких целевых компонентов были использованы калиброванные мерные шприцы фирмы Hamilton (США), 1 и 10 мкл, с газонепроницаемыми прокладками для обеспечения герметичности во избежание существенных утечек газа или жидкости.

Подготовка пакетов

Перед приготовлением газовой смеси для каждого пакета проводилась процедура чистки пакета с помощью троекратного наполнения и высвобождения комнатным воздухом. Дополнительные испытания показали, что аналитический сигнал пробы комнатного воздуха после такой процедуры отчистки находится на уровне шумового сигнала базовой линии сенсоров. После процедуры чистки пакет вакуумируют с помощью насоса.

Наполнение пакетов окружающим воздухом

Пакет заполняют фоновым газом (комнатным воздухом) с помощью наноса 12 В (Alldoo Micropump Co., Китай) с известной и постоянной скоростью нагнетания до необходимого объема пакета. Скорость прокачки комнатного воздуха насосом контролируется напряжением от источника питания постоянного тока Б5-47 (Завод Измеритель, Армения) и регистрируется ротаметром VA-10414 (Dwyer Instruments Inc., США).

Ввод вещества

Необходимый объем целевого компонента вычисляется на основе требуемого состава окончательной газовой смеси и объема пакета. После того, как объем жидкости в шприце достигнет заданного значения, иглу шприца сразу вводят в пакет через мембрану медленно надавливая на поршень шприца и одновременно извлекая при этом иглу из мембраны. После введения вещества газовую смесь выдерживали в течение заранее определенного времени для гомогенизации и достижения температурного равновесия

между смесью и окружающей средой. Дополнительные эксперименты показали, что для всех исследуемых ЛОС для достижения равновесия и получения воспроизводимого сигнала сенсора достаточно 20 минут.

Исследуемые ЛОС

В данном исследовании использовались следующие вещества: н-гептан (≥ 99%, для ВЭЖХ), этилбензол (≥ 99.5%, аналитический стандарт), пропан-1-ол (≥ 99.9%, для ВЭЖХ), о-ксилол (≥ 99.0%, чистый для анализа) фирмы Sigma Aldrich (Merck KGaA, Дармштадт, Германия).

Расчет концентрации аналита A в конечной газовой смеси производили по следующей формуле (объемом внесенного вещества порядка нескольких мкл в сравнении с используемыми объемами пакета порядка 1-5 л можно пренебречь):

$$c_A\ (ppm) = \frac{V_A * \rho_A * R * T}{M_A * V_{\text{п}} * p} * 1000 \quad (22)$$

где

V_A – объем вещества А в виде жидкости, мкл;

ρ_A – плотность вещества А, г/мл;

M_A – молярная масса вещества А, г/моль;

$V_{\text{п}}$ – объем пакета перед вводом вещества, л;

R – газовая постоянная, равная 8.314 Дж/(моль∗К)

T- температура, К;

p – давление, кПа.

2.3. Анализ модельных газовых смесей и проб выдыхаемого воздуха в медицинском исследовании с использованием МС 1

2.3.1. Схема экспериментальной установки для анализа модельных газовых смесей и проб выдыхаемого воздуха

В отличие от достаточно часто используемых подходов с предварительным концентрированием ЛОС или любыми дополнительными процедурами хранения, в данной работе использовался прямой онлайн-анализ для анализа проб выдыхаемого воздуха. Использование дополнительных промежуточных этапов в процессе пробоподготовки может отрицательно повлиять на результаты: например, потеря аналита

или появление артефактов, деградация адсорбированных аналитов при хранении [38], термическое разложение или изомеризация некоторых соединений в процессе термодесорбции [39,40], и деградация материала сорбента [41,42]. Для устранения «эффекта памяти», связанного с сорбцией ЛОС на стенках транспортировочных трубок, в экспериментальную схему онлайн-анализа ВВ (рисунок 6) был установлен насос для постоянной продувки магистралей схемы и ячейки (3,5 л×мин$^{-1}$). Это позволило обеспечить стабильный сигнал и, как следствие, стабильную базовую линию без дрейфа. При этом установлено, что дальнейшее увеличение скорости приводит к значимому уменьшению стабильности сигналов сенсоров.

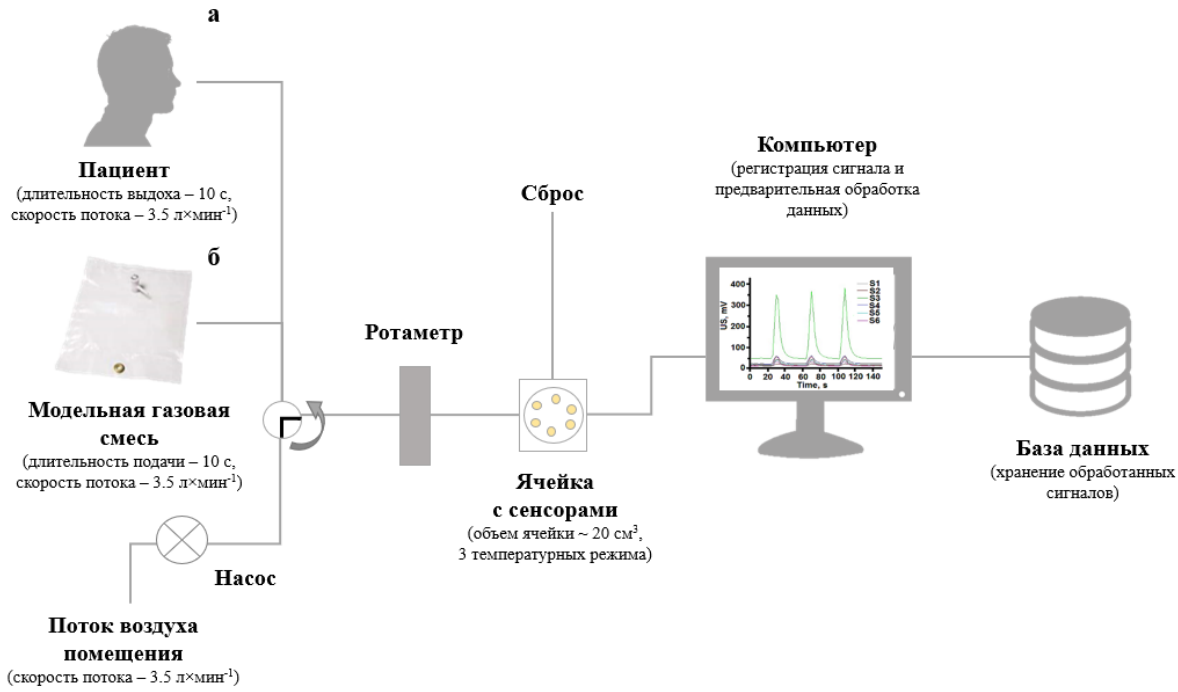

Рисунок 6. Экспериментальная схема анализа для МС 1 (а – схема онлайн-анализа ВВ при медицинском исследовании, б – схема онлайн-анализа модельных газовых смесей для оценки перекрестных чувствительностей)

Для учета присутствующих экзогенных ЛОС в окружающем воздухе помещения, ячейку с сенсорами продували комнатным воздухом, таким образом, не было необходимости в «очистке» легких испытуемого чистым воздухом в течение 3–5 минут для уменьшения влияния экзогенных ЛОС на результат анализа. Отметим, что предварительная очистка дыхания пациента медицинским воздухом, которая использовалась в похожих исследованиях, может приводить как к значительному

изменению профиля ЛОС пациента, так и потере ЛОС. Подход, предложенный в данной работе, упрощает анализ, уменьшая влияние экзогенных ЛОС. Перед началом анализа каждый испытуемый ожидал в течение 10 мин до первого измерения ВВ. Более того, дополнительное исследование с использованием угольного фильтра показало отсутствие изменений в откликах сенсоров. Стоит пояснить, что в качестве первичного критерия отбора помещений, пригодных для анализа ВВ была использована оценка изменения проводимости сенсоров между базовой экспериментальной схемой анализа и той же схемой с подключением на входе угольного фильтра. Критерий был определен следующим образом: в случае если максимальное относительное изменение сенсора составляет более 5%, то помещение определялось как непригодное для проведения анализа ВВ пациентов. На рисунке 7 представлен пример такой проверки для одного помещения.

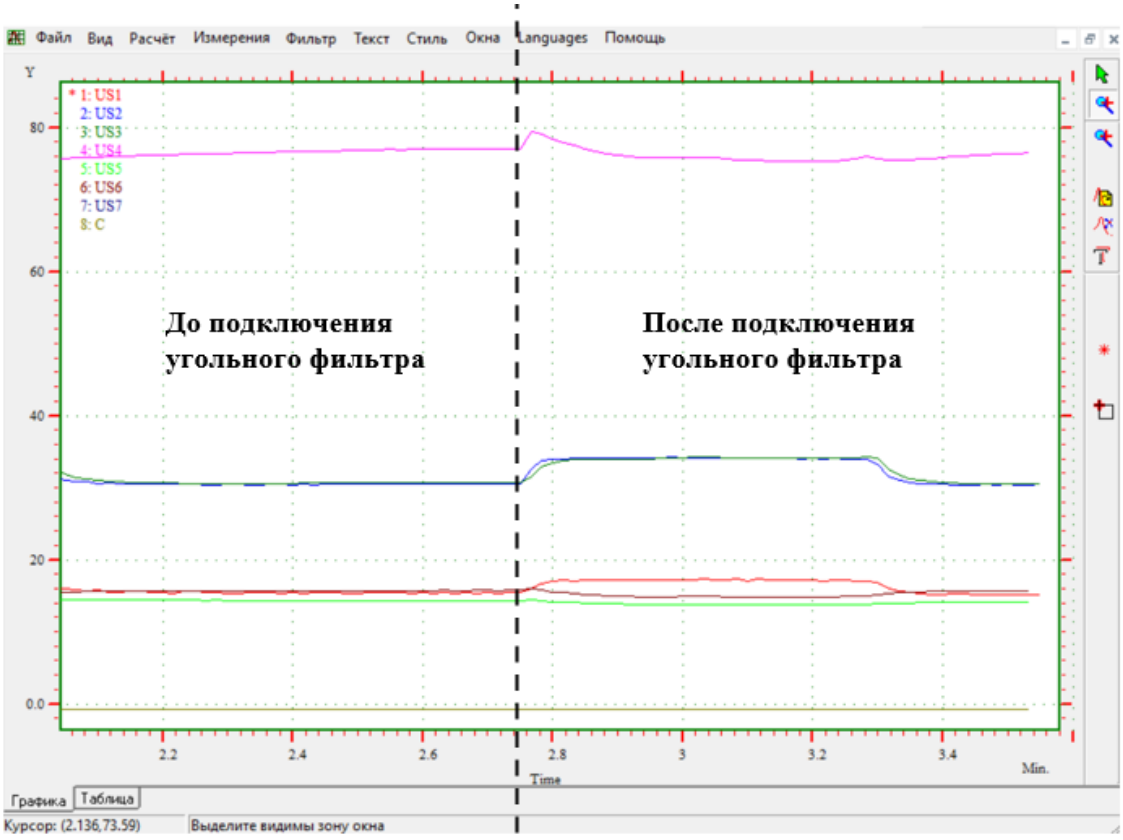

Рисунок 7. Изменение проводимостей сенсоров при подключении ко входу экспериментальной схемы анализа угольного фильтра. На рисунке изображен пример интерфейса приложения для отображения проводимостей сенсоров в относительных единицах (ось ординат) от времени в минутах (ось абсцисс)

Процедуру анализа как ВВ пациента, так и модельных газовых смесей можно условно разделить на три последовательных этапа. На первом этапе клапан насоса открыт и через ячейку с сенсорами прокачивается окружающий воздух со скоростью 3,5 л×мин$^{-1}$, в то время как клапан для анализируемой пробы закрыт (как показано на рисунке 6). Данная скорость была выбрана как оптимальная, что позволяет пропустить через ячейку сенсоров достаточный объем ВВ пациента (~600 мл), для которого приемлемая продолжительность выдоха составила 10 секунд, и получить информативный отклик в режиме онлайн-анализа. На втором этапе клапан насоса закрывается и открывается клапан для подачи пробы. Затем подается проба продолжительностью 10 секунд через ячейку с сенсорами со скоростью 3,5 л×мин$^{-1}$, либо в случае медицинского исследования пациент делает выдох через стерильный одноразовый мундштук продолжительностью 10 секунд через ячейку со скоростью 3,5 л×мин$^{-1}$, что соответствует объему порядка 600 мл ВВ. На третьем этапе клапан подачи пробы закрывается, открывается клапан насоса и снова прокачивается окружающий воздух через ячейку с сенсорами с той же скоростью 3,5 л×мин$^{-1}$. Скорость потока контролируется с помощью ротаметра. Скорость выдоха контролировалась пациентом также с отслеживанием показаний ротаметра. В данной работе использовались микровакуумный насос постоянного тока 12 В для медицинских целей (Alldoo Micropump Co., Ltd, Юэцин Чжэцзян, Китай) и ротаметр VA-0414 (Dwyer Instruments Inc, Индиана, США). Аналогичные действия производились на всех температурных режимах после выхода на стабильный сигнал сенсоров.

В качестве аналитического сигнала для каждого сенсора на каждом температурном режиме из кривой проводимости извлекался интеграл проводимости по времени за вычетом площади, образуемой базовой линией (рисунок 8).

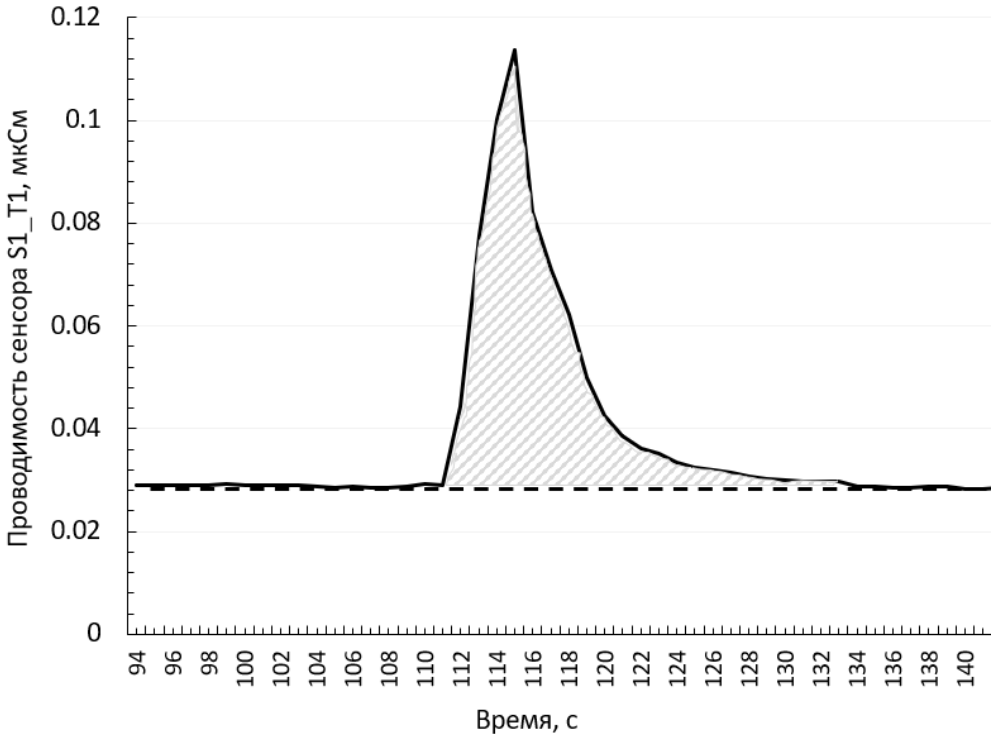

Рисунок 8. Принцип извлечения аналитического сигнала (в качестве аналитического сигнала выступает значение площади под кривой проводимости сенсора за вычетом базовой линии)

2.3.2. Определение относительных чувствительностей летучих органических соединений для используемых сенсоров

Для оптимизации параметров измерения и оценки свойств перекрестной чувствительности мультисенсорной системы использовали однокомпонентные модельных газовые смеси следующих соединений: н-гептан, этилбензол и пропан-1-ол. Все эти вещества являются потенциальными биомаркерами РЛ, которые часто использовались в аналогичных исследованиях [17,18]. Все реагенты использовали для оценки чувствительности и линейности градуировочных зависимостей.

Для оценки характера чувствительности были построены градуировочные зависимости для пропан-1-ола, этилбензола и н-гептана. Диапазон исследуемых концентраций для градуировочных зависимостей составлял 0.5–500 ppm. Для каждого концентрационного уровня использовалось по 6 параллельных измерений (n=6). В выбранном концентрационном диапазоне хорошо работает линейное приближение для построения градуировки. Параметры градуировочных зависимостей представлены в

таблице 9. Исходные данные для построения градуировочных зависимостей расположены в репозитории [99]. Стоит отметить, что в диапазоне заданных концентраций зависимости для гептана и пропан-1-ола являлись линейными, в то время как для этилбензола нелинейность аналитического сигнала возникает при очень высоких концентрациях (начиная с 250 ppm).

Коэффициенты относительной чувствительности (relative sensitivity factor, RSF) были рассчитаны относительно пропан-1-ола по уравнению:

$$RSF_{A(i)} = \frac{k_{A(i)}}{k_{Prop(i)}} \quad (23)$$

где $k_{A(i)}$– коэффициент наклона градуировочной зависимости для i-го сенсора для соединения А, а $k_{Prop(i)}$- коэффициент наклона градуировочной зависимости для i-го сенсора для пропан-1-ола. Коэффициенты были рассчитаны для трех температурных режимов (таблица 10).

Таблица 9. Параметры градуировочных зависимостей, построенных для н-гептана, пропан-1-ола и этилбензола для первого температурного режима

	н-гептан				
Сенсор	наклон	оффсет	R^2	RMSEC	RMSEP
S1	99.3	-4.93	0.9955	12.5	13.3
S2	24.07	-3.13	0.9970	10.1	10.6
S3	0.06	-12.1	0.9961	11.6	12.5
S4	0.05	-9.95	0.9912	17.4	17.1
S5	0.07	-15.58	0.9938	14.6	15.5
S6	0.02	-18.79	0.9913	17.3	18.1
	пропан-1-ол				
Сенсор	наклон	оффсет	R^2	RMSEC	RMSEP
S1	82.4	0.94	0.9981	7.2	7.3
S2	43.42	-2.05	0.9951	11.4	12.5
S3	0.62	-7.82	0.9951	11.4	11.4
S4	2.24	1.87	0.9932	13.6	14.3
S5	5.23	-4.24	0.9960	10.4	11.2
S6	1.94	-0.87	0.9933	13.4	14.3
	этилбензол				
Сенсор	наклон	оффсет	R^2	RMSEC	RMSEP
S1	29.21	-13.53	0.9787	19.2	19.8
S2	10.88	-12.35	0.9683	23.4	23.4
S3	0.15	-17.09	0.9805	18.3	18.6
S4	0.56	-19.51	0.9692	23	23
S5	0.56	-10.78	0.9936	10.5	10.9
S6	0.17	-22.42	0.9519	28.8	29.2

Таблица 10. Представление коэффициентов относительных чувствительностей относительно пропан-1-ола для трех температурных режимов (максимальная относительная погрешность определения RSF составила 11.6%)

Соединение	**S1_T1**	**S2_T1**	**S3_T1**	**S4_T1**	**S5_T1**	**S6_T1**
н-гептан	1.205	0.554	0.097	0.022	0.013	0.01
этилбензол	0.354	0.251	0.242	0.25	0.107	0.088
пропан-1-ол	1.0	1.0	1.0	1.0	1.0	1.0

Соединение	**S1_T2**	**S2_T2**	**S3_T2**	**S4_T2**	**S5_T2**	**S6_T2**
н-гептан	1.406	0.604	0.054	0.008	0.007	0.004
этилбензол	0.148	0.059	0.08	0.091	0.108	0.066
пропан-1-ол	1.0	1.0	1.0	1.0	1.0	1.0

Соединение	**S1_T3**	**S2_T3**	**S3_T3**	**S4_T3**	**S5_T3**	**S6_T3**
н-гептан	0.904	0.478	0.027	0.028	0.005	0.004
этилбензол	0.103	0.078	0.046	0.072	0.04	0.036
пропан-1-ол	1.0	1.0	1.0	1.0	1.0	1.0

Как видно из таблицы 8 относительные чувствительности сенсора для исследуемых веществ могут различаться на несколько порядков, что говорит о хорошей перекрестной чувствительности выбранных для исследования сенсоров.

2.3.3. Обработка данных и обучение классификаторов

Для анализа, визуализации и обработки данных, оценки распределения с помощью статистического критерия, применения PCA, а также для обучения математических моделей-классификаторов (kNN, RF, SVM, LR) было использовано программное

обеспечение Python 3.6 (Python Software Foundation, США) и библиотеки pandas, scipy, matplotlib, numpy и scikit-learn.

На рисунке 9 изображена схема обработки массива откликов от МС 1 в рамках проведенного медицинского исследования.

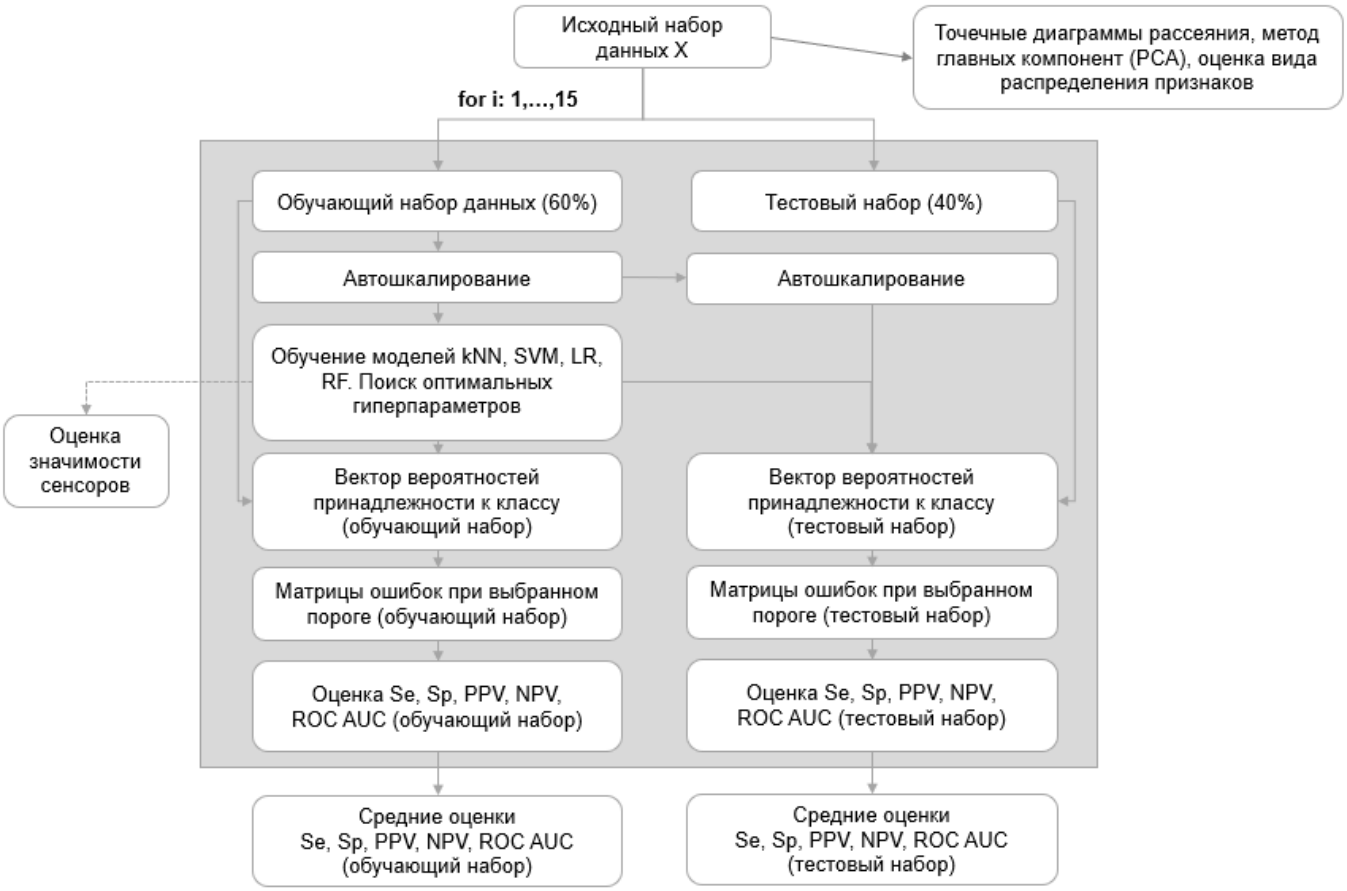

Рисунок 9. Схема обработки данных с помощью МС 1, полученных в ходе медицинского исследования, и представления итоговых результатов разработанного диагностического теста

Детальный скрипт обработки и исходные данные: массив откликов с соответствующими атрибутами пациента (группа, возраст, пол и др.) содержатся в репозитории [99]. На каждой итерации для каждого алгоритма выполняется внутренняя перекрестная проверка на обучающем наборе с числом блоков равным трем. Такая проверка выполняется для всех возможных комбинаций гиперпараметров алгоритма, заданных изначально. В таблице 11 представлено описание набора значений гиперпараметров для каждого классификатора. В качестве оптимальной выбирается та

комбинация гиперпараметров, при которой усредненная по трем блокам метрика качества:

$$\text{сбалансированная точность} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (24)$$

будет максимальна.

Таблица 11. Описание сетки гиперпараметров для поиска оптимальной комбинации на обучающем наборе данных (для каждой итерации разбиения исходных данных)

Метод классификации	**Гиперпараметр**	**Исследованные значения**
kNN	Количество соседей (n_neighbors)	[1, 2, …, 5]
kNN	Тип взвешивания (weights)	[равномерное взвешивание, взвешивание по евклидовой метрике]
LR	Обратный коэффициент регуляризации (C)	[0.01, 0.02, …, 1.00]
LR	Тип регуляризации (penalty)	[L1, L2]
RF	Количество базовых деревьев (n_estimators)	[10, 20, …, 50]
RF	Максимальная глубина дерева (max_depth)	[1, 3, 5, …, 13]
SVM	Параметр регуляризации (C)	[0.0001, 0.001, 0.01, …, 10]
SVM	Ядро (kernel)	[линейное, ядро с гауссовой радиальной базисной функцией]

2.4. Анализ модельных газовых смесей для переноса градуировочных зависимостей с использованием МС 2.1 и МС 2.2.

2.4.1. Схема экспериментальной установки для анализа модельных газовых смесей

Схема установки для оценки возможности переноса градуировочной зависимости по принципу схожа со схемой, изображенной на рисунке 6. Отличие текущей схемы (рисунок 10) от вышеуказанной заключается лишь в значении нескольких параметров, а именно: скорость потока газовой смеси или потока комнатного воздуха для формирования базовой линии – 0.4 л/мин, продолжительность подачи пробы – 90 с, временной период интегрирования сигнала – 300 с.

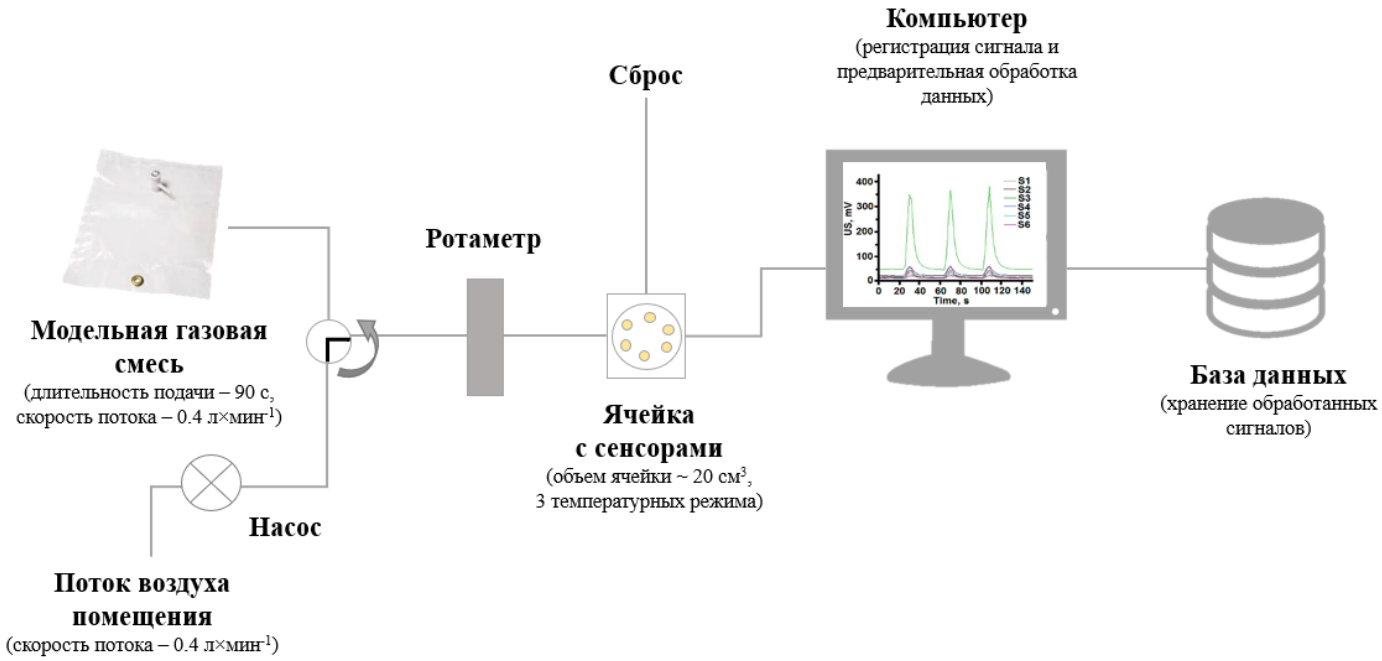

Рисунок 10. Экспериментальная схема анализа модельных газовых смесей для оценки возможности переноса градуировочных зависимостей для МС 2.1 и МС 2.2

В таблице 12 приведены основные значения параметров проведения анализа для МС 1, МС 2.1 и МС 2.2.

Таблица 12. Параметры измерения проводимости сенсоров для мультисенсорных систем при анализе модельных газовых смесей

Параметр	**МС 1**	**МС 2.1 и МС 2.2**
Продолжительность подачи пробы	10 с	90 с
Скорость потока воздушной смеси при подаче пробы и формировании базовой линии	3.5 л/мин	0.4 л/мин
Временной период интегрирования сигнала начиная от времени подачи пробы	90 с	300 с

2.4.2. Обработка данных

Для анализа, визуализации и обработки данных, применения PCA, а также для обучения математических моделей-классификаторов (MSVM и SVM) и регрессионных моделей (UDS, UDSwoi, DS-PLS2, DS-L1R) было использовано программное обеспечение Python 3.6 (Python Software Foundation, США) и библиотеки pandas, matplotlib, numpy и scikit-learn.

Для моделирования и проведения экспериментов по оценке возможности проведения переноса градуировочных зависимостей с помощью методов стандартизации отклика были собраны данные, полученные при анализе градуировочных образцов однокомпонентных газовых смесей трех ЛОС и образцов двух газовых смесей с идентичным качественным составом по набору входящих ЛОС, но различных по суммарному количественному составу. В таблице 13 и 14 приведены концентрации вышеуказанных образцов. В случае образцов однокомпонентных смесей для каждой концентрации и каждого компонента готовилось по 2 образца и итоговое количество проанализированных образцов для каждой МС составило 42. В случае с образцами газовых смесей, состоящих из трех компонентов ЛОС, для каждой смеси готовилось по 8 образцов. Итоговое количество проанализированных образцов для каждой МС составило 16.

Таблица 13. Состав однокомпонентных градуировочных газовых смесей для моделирования задачи классификации

	Концентрация, ppm		
Номер образца	**пропан-1-ол**	**н-гептан**	**о-ксилол**
1	1.2	1.2	1.5
2	2.4	2.4	3.0
3	6.0	6.1	7.4
4	11.9	12.1	14.9
5	23.9	24.3	29.7
6	59.6	60.7	74.3
7	119.3	121.5	148.5

Таблица 14. Состав газовых смесей (ГС 1 и ГС 2) для моделирования задачи классификации

№ смеси	**Компонент смеси**	**Концентрация, ppm**
1	пропан-1-ол	33
	н-гептан	17
	о-ксилол	20
2	пропан-1-ол	26
	н-гептан	17
	о-ксилол	24

Исходные матрицы откликов сенсоров с информацией о качественном составе образцов, полученных при анализе однокомпонентных градуировочных газовых смесей и трехкомпонентных смесей, расположены в репозитории [99].

Схема проведения экспериментов по оценке возможности проведения переноса градуировочных зависимостей с помощью методов стандартизации приведена на рисунке 11. По итогу эксперимента сравнивали усредненные значения точности мультиклассовой классификации (в случае с однокомпонентными газовыми смесями) и точности бинарной классификации (в случае с многокомпонентными смесями) для

каждой комбинации «модель – тестовый набор для обучения» по 15 случайным разбиениям набора данных на обучающий и тестовый наборы в соотношении 70% к 30% и случайным отбором стандартизационных образцов. На этапе обучения на каждой итерации для каждого алгоритма выполняется внутренняя перекрестная проверка на обучающем наборе данных с числом блоков равным трем. В таблице 15 представлено описание набора значений гиперпараметров для используемых классификаторов MSVM и SVM. С подробным алгоритмом эксперимента можно ознакомиться в скрипте, расположенном в репозитории [99].

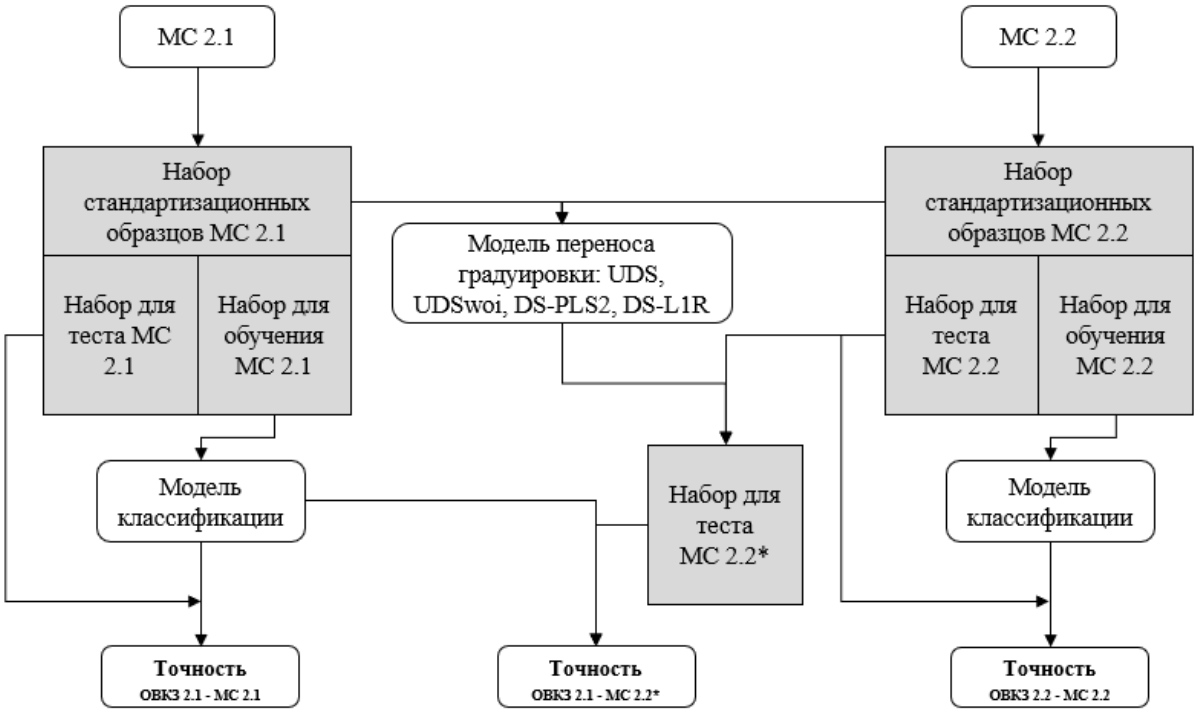

Рисунок 11. Схема обработки данных для двух массивов сенсоров и проведения корректировки откликов МС 2.2 с помощью методов переноса градуировочных зависимостей

Таблица 15. Описание сетки гиперпараметров для поиска оптимальной комбинации на обучающем наборе данных

Метод классификации	**Гиперпараметр**	**Исследованные значения**
MSVM	Параметр регуляризации (C)	$[10^{-2}, 10^{-1}, \ldots, 10^{-2}]$
SVM	Параметр регуляризации (C)	$[10^{-8}, 10^{-7}, \ldots, 10^{-4}]$

Глава 3. Разработка метода онлайн-анализа выдыхаемого воздуха для диагностики рака легких с использованием мультисенсорной системы

3.4. Описание медицинского исследования

Все исследования с участием как пациентов, так и здоровых добровольцев проводились в соответствии с этическими нормами и Хельсинкской декларацией от 1964 года. Все эксперименты проводились с разрешения Национального медицинского исследовательского центра онкологии им. Н.Н. Петрова комитетом по этике No. 15/83 от 15 марта 2017 г [1,100]. Мультисенсорная система МС 1 была установлена в исследовательском центре. Были использованы следующие критерии включения в исследование: пациенты в возрасте 20 лет и старше; подозрение на РЛ на основании клинических симптомов или рентгенологического исследования (пациентам была предоставлена возможность провести все необходимые дополнительные диагностические исследования такие как компьютерная томография, фибробронхоскопия и/или трансторакальная трепан-биопсия); активный курильщик или курильщик в прошлом, бросивший не более 10 лет до участия в исследовании. Критерии исключения из исследования: пациенты с тяжелыми сопутствующими нарушениями здоровья (включая декомпенсированную патологию сердечно-сосудистой, эндокринной системы, например, сахарный диабет, или легочной системы; декомпенсированная или субкомпенсированная органная недостаточность; некорректируемые коагулопатии; цереброваскулярные нарушения; нестабильная стенокардия); онкологические заболевания в анамнезе за 5 лет до исследования, за исключением рака кожи или рака шейки матки in situ; РЛ или операции на легких в анамнезе; ожидаемая выживаемость 1 год или меньше.

В ходе исследования [100] были проанализированы образцы ВВ 118 пациентов (49 – женщины (42%), 69 – мужчины (58%)), среди них 65 пациентов с РЛ и 53 здоровых пациента, которые добровольно приняли участие в исследовании и подписали согласие на участие . У всех пациентов, которые входили в группу РЛ, диагноз был подтвержден с помощью морфологического исследования после анализа ВВ. Были диагностированы следующие формы РЛ: немелкоклеточный РЛ – 59, мелкоклеточный РЛ – 6. У 30

пациентов была ранняя стадия заболевания (I / II), а у 35 – поздняя стадия (III / IV). Кроме того, в контрольную группу вошли 54 здоровых добровольца без признаков заболевания легких по клиническим симптомам и рентгенологическому обследованию. Данные по полу и возрасту контрольной группы также представлены в таблице 16.

Таблица 16. Характеристики исследуемых групп (группа РЛ и группа здоровых)

Группа	**Группа РЛ**	**Группа здоровых**
Количество	65	53
Возраст, среднее ± ско*	65 ± 9	56 ± 12
Мужчины	42 (65%)	27 (51%)
Женщины	23 (35%)	26 (49%)

*ско – среднеквадратичное отклонение

Пациенты были проинформированы о необходимости воздержаться от еды и курения в течение часа перед анализом ВВ с помощью МС 1. Непосредственно перед анализом пациент прополаскивал полость рта чистой теплой водой и ожидал в помещении, где проводились измерения, не менее 10 минут перед первым измерением.

3.4. Описание процедуры проведения анализа ВВ пациентов

Процедура анализа ВВ для пациента состояла из трех последовательных этапов. На **первом этапе** клапан насоса открыт и продувается окружающий воздух через газовую ячейку со скоростью 3,5 л×мин$^{-1}$, в то время как клапан для пациента закрыт (как показано на рисунке 12). На **втором этапе** клапан насоса закрывается и открывается клапан для пациента. Пациент делает выдох через стерильный одноразовый мундштук продолжительностью 10 секунд через газовую камеру со скоростью 3,5 л×мин$^{-1}$, что соответствует объему порядка 600 мл ВВ. Скорость контролировалась пациентом с отслеживанием показаний ротаметра. На **третьем этапе** клапан пациента закрывается, открывается клапан насоса и снова продувается окружающий воздух через газовую

ячейку с той же скоростью 3,5 л×мин$^{-1}$. Скорость потока также отслеживается с помощью ротаметра.

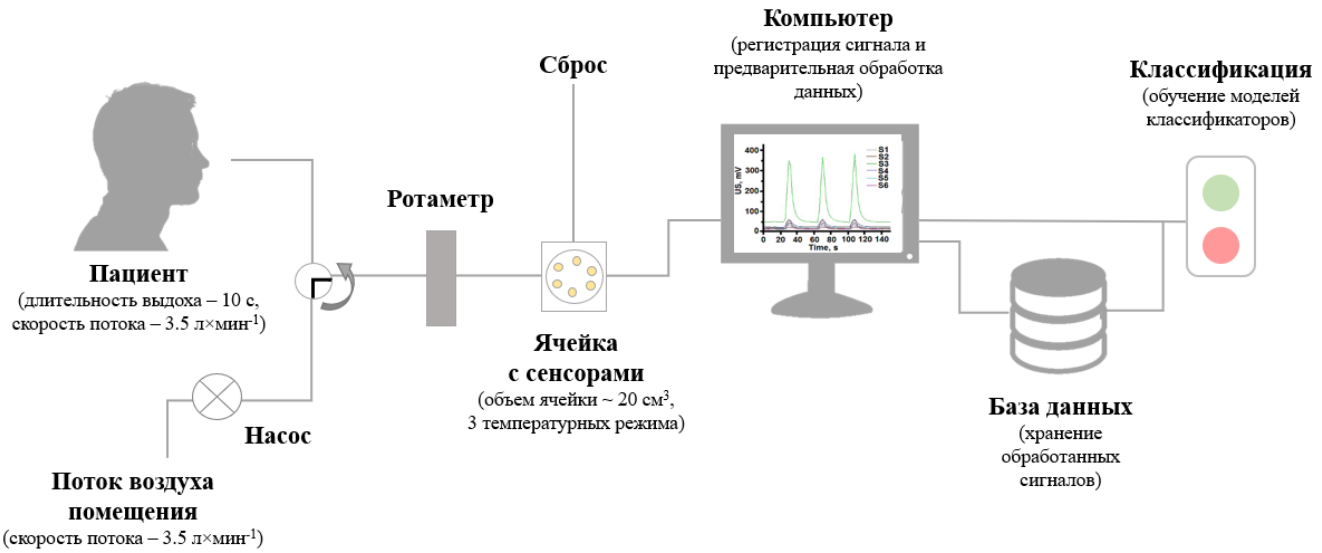

Рисунок 12. Схема онлайн-анализа ВВ

В качестве аналитического сигнала для каждого сенсора на каждом температурном режиме из кривой проводимости извлекался интеграл проводимости по времени за вычетом площади, образуемой базовой линией. Время между последовательными измерениями при одной и той же температуре сенсора составляло 1-2 минут для одного пациента.

Воспроизводимость извлекаемого аналитического сигнала для всех сенсоров составляла порядка 2–15%. Изначально проводилось три последовательных измерения ВВ для каждого из трех температурных режимов. В процессе проведения исследования была выявлена возможность проводить только одно измерение на каждом температурном режиме, в результате чего общее время анализа для одного пациента варьировалось от 15 до 25 минут.

Чем выше уровень влажности, тем меньше на МО сенсоры влияют ее флуктуации. Поэтому все измерения проводились в помещениях с относительно высоким уровнем влажности (60% ± 5%). Ввиду использования высоких рабочих температур влияние колебаний температуры воздуха, проходящего через ячейку, незначительно; поэтому для ячейки предварительный нагрев не использовался.

3.4. Выбор наиболее эффективного алгоритма обработки данных и классификационной модели

Итоговая матрица признаков **X** имеет размерность 118×18, где 118 – общее число пациентов, а 18 – количество извлеченных признаков (интеграл пика проводимости от времени) из 6 сенсоров на 3 температурных режимах. Вектор **y** имеет размерность 118×1 с значением 0 (пациент принадлежит к группе здоровых) или 1 (пациент относится к группе РЛ).

Начнем с того, что классы не сбалансированы (53 против 65 для группы здоровых группы и группы РЛ соответственно), поэтому при разбиениях на перекрестной проверке использовалась стратификация по классу, таким образом соотношение группы РЛ и контрольной группы была одинаковым между разбиениями.

Для исследования и визуализации данных в двумерном пространстве были построены 3 матрицы для каждого температурного режима с точечными диаграммами рассеяния. В большей степени межгрупповое разделение наблюдается для первого температурного режима (рисунок 9). Стоит отметить, что визуально наибольшее разделение между группой РЛ и группой здоровых наблюдается для пар с присутствием сенсора №4.

Для проверки распределения признаков на нормальность использовали критерий Шапиро-Уилка [101], являющийся одним из наиболее эффективных критериев проверки нормальности. Гипотеза о нормальности распределения была отвергнута для всех признаков (α=0.05). Из рисунка 9 можно сделать предположении о логнормальном распределении данных в признаках. Действительно, при логарифмировании матрицы признаков гипотеза о нормальном распределении была отвергнута уже только для 6 из 18 признаков (α=0.05) (таблица 17). Тем не менее большинство упомянутых моделей-классификатор не требуют, чтобы признаки были нормально распределены. Так как данные распределены не нормально, то стоит исключить из рассмотрения часто используемый классификатор на основе LDA.

Для анализа PCA данные были предварительно автошкалированы. На рисунке 14 представлена зависимость объясненной дисперсии от количества главных компонент, используемых в модели. На рисунке 15 изображены представления образцов ВВ в пространстве 3 главных компонент.

Рисунок 13. Матрица точечных диаграмм рассеяния для 6 сенсоров при температурном режиме T1. В ячейке (i,j) находится точечная диаграмма рассеяния для интеграла проводимости (мкСм×с) i и j сенсоров. В диагональных ячейках (i,i) изображено сглаженное распределение для i сенсора. Принадлежность образца ВВ к группе отображена цветом (синий – контрольная группа (0), оранжевый – группа РЛ (1))

Таблица 17. Проверка на нормальность исходных данных и преобразованных с помощью логарифмирования по основанию **e** с использованием критерия Шапиро-Уилка (α=0.05)

	Исходные данные			**После логарифмирования по основанию e**		
#	**W-value**	**p-value**	**H0 отклоняется**	**W-value**	**p-value**	**H0 отклоняется**
S1_T1	0.930	1×10^{-5}	да	0.982	0.116	нет
S2_T1	0.872	1×10^{-8}	да	0.990	0.512	нет
S3_T1	0.917	2×10^{-6}	да	0.986	0.259	нет
S4_T1	0.875	2×10^{-8}	да	0.978	0.051	нет
S5_T1	0.917	2×10^{-6}	да	0.990	0.513	нет
S6_T1	0.900	2×10^{-7}	да	0.983	0.154	нет
S1_T2	0.853	2×10^{-9}	да	0.995	0.933	нет
S2_T2	0.823	1×10^{-10}	да	0.993	0.844	нет
S3_T2	0.897	2×10^{-7}	да	0.976	0.034	да
S4_T2	0.847	1×10^{-9}	да	0.977	0.037	да
S5_T2	0.704	4×10^{-14}	да	0.987	0.299	нет
S6_T2	0.897	2×10^{-7}	да	0.987	0.347	нет
S1_T3	0.679	1×10^{-14}	да	0.974	0.021	да
S2_T3	0.692	2×10^{-14}	да	0.970	0.009	да
S3_T3	0.785	7×10^{-12}	да	0.978	0.047	да
S4_T3	0.733	2×10^{-13}	да	0.969	0.009	да
S5_T3	0.806	4×10^{-11}	да	0.987	0.307	нет
S6_T3	0.822	1×10^{-10}	да	0.985	0.199	нет

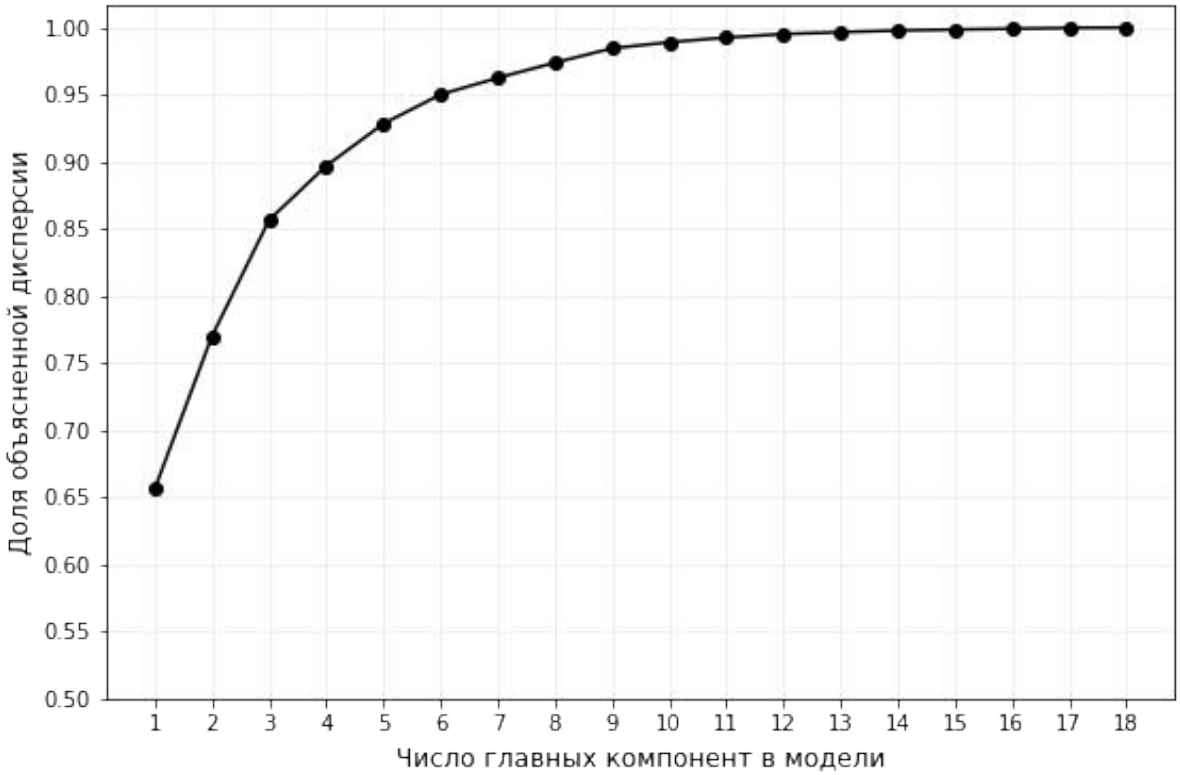

Рисунок 14. Зависимость доли объясненной дисперсии данных от количества главных компонент в модели

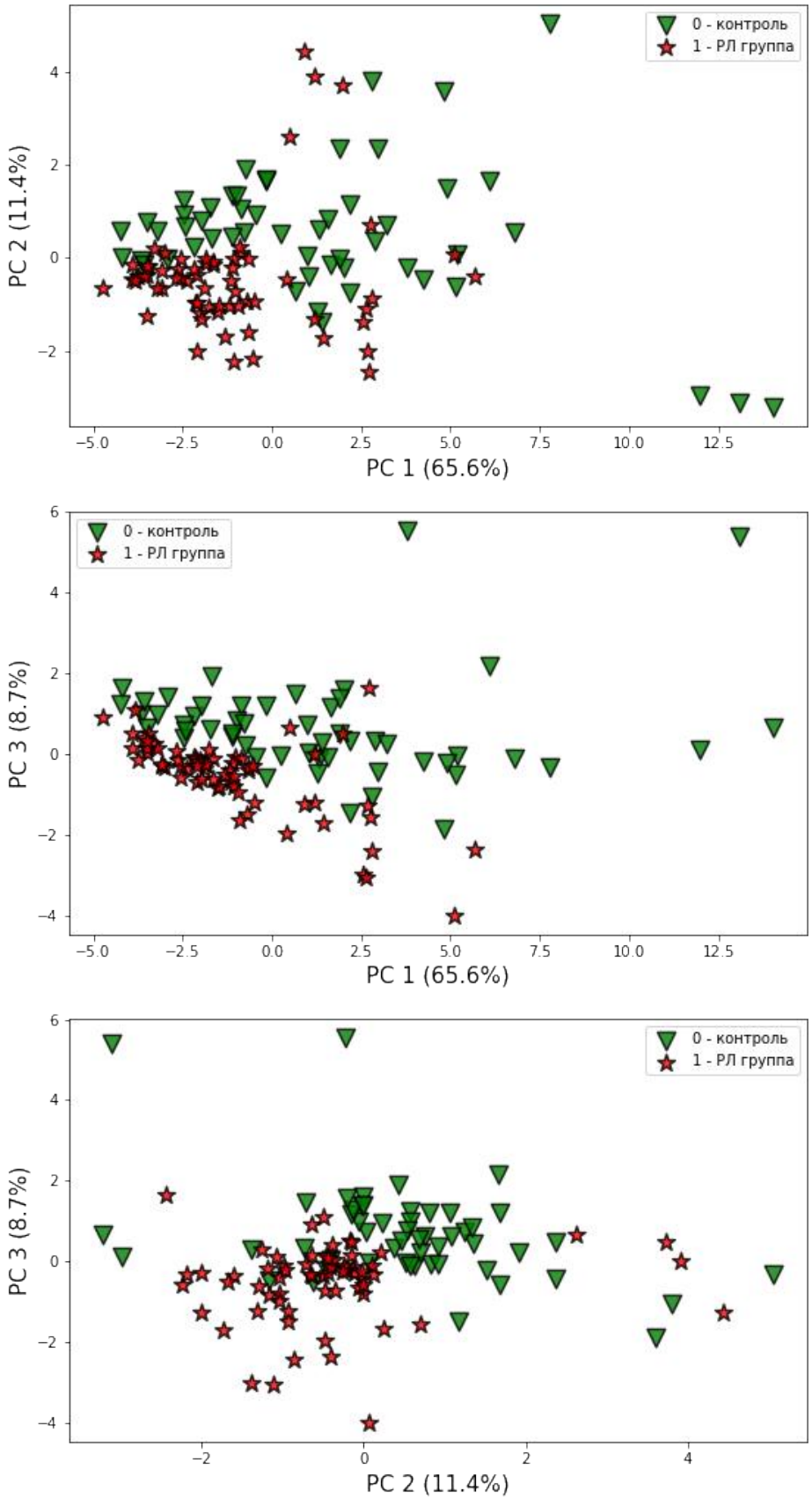

Рисунок 15. Образцы ВВ в пространстве пар трех главных компонент (PC1-PC2, PC1-PC3, PC2-PC3)

В соответствии со схемой обработки данных, представленной на рисунке 9, для каждой итерации (всего 15) производилось случайное разбиение матрицы откликов сенсоров и вектора меток класса на обучающий и тестовый наборы данных в соотношении 60 к 40. Затем, на тренировочных наборах были обучены рассматриваемые модели классификации с учетом внутренней перекрестной проверки для выбора оптимальных значений гиперпараметров для каждой модели. Далее обученные модели использовали для предсказания вероятностей и меток класса как для образцов тренировочного набора, так и для тестового.

На рисунке 16 представлены построенные ROC кривые для моделей классификации для одной из итераций. После построения ROC кривых рассчитывались метрики ROC AUC.

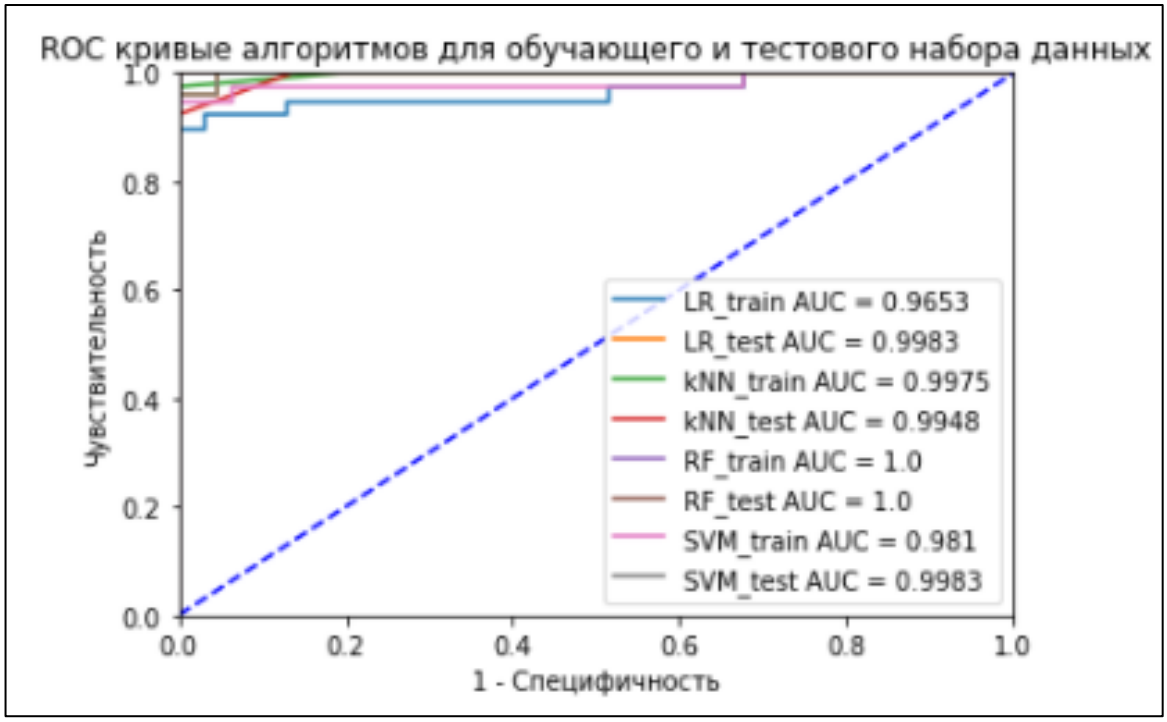

Рисунок 16. ROC кривые алгоритмов классификации, полученные при предсказании меток классов для одного разбиения данных на обучающий и тестовые наборы

При использовании стандартного порога классификации (p=0.5) каждый образец был классифицирован как истинно положительный (TP), ложноположительный (FP), истинно отрицательный (TN) или ложноотрицательный (FN). Далее производили построение матриц ошибок. В таблицах 18 и 19 представлены примеры построенных

матриц ошибок для одной из итераций для обучающего и тестового набора соответственно.

Таблица 18. Матрицы ошибок предсказания классов для обучающего набора данных по одному случайному разбиению

kNN		Предсказанный класс	
		1	0
Истинный класс	1	36	0
	0	3	31

LR		Предсказанный класс	
		1	0
Истинный класс	1	36	3
	0	2	29

RF		Предсказанный класс	
		1	0
Истинный класс	1	39	0
	0	0	31

SVM		Предсказанный класс	
		1	0
Истинный класс	1	37	2
	0	0	31

Таблица 19. Матрицы ошибок предсказания классов для тестового набора данных по одному случайному разбиению

kNN		Предсказанный класс	
		1	0
Истинный класс	1	24	2
	0	0	22

LR		Предсказанный класс	
		1	0
Истинный класс	1	26	0
	0	2	20

RF		Предсказанный класс	
		1	0
Истинный класс	1	26	0
	0	0	22

SVM		Предсказанный класс	
		1	0
Истинный класс	1	25	1
	0	1	21

Далее для каждой модели классификации и типа использованного набора данных рассчитывались метрики диагностического теста, которые в дальнейшем были усреднены по всем 15 итерациям. В таблице 20 содержится сводная информация по усредненным метрикам качества (чувствительности, специфичности и ROC AUC) как при классификации образцов из обучающего набора, так и для тестового набора данных.

Таблица 20. Средние значения метрик с доверительным интервалом (при уровне значимости α=0.05) при классификации алгоритмов на обучающем и тестовом наборе данных при 15-кратном случайном разбиении набора данных

Метрика	**Набор данных**	**Алгоритм классификации**			
		LR	**kNN**	**SVM**	**RF**
ROC AUC	обучение	0.981±0.007	0.998±0.001	0.993±0.005	0.999±0.000
	тест	0.959±0.012	0.961±0.018	0.972±0.014	0.960±0.014
Acc	обучение	0.956±0.015	0.970±0.012	0.990±0.007	0.988±0.007
	тест	0.931±0.018	0.940±0.016	0.936±0.016	0.898±0.023
Se	обучение	0.957±0.011	0.947±0.022	0.986±0.011	0.981±0.012
	тест	0.938±0.021	0.905±0.026	0.920±0.025	0.853±0.041
Sp	обучение	0.954±0.025	1.000±0.000	0.995±0.006	0.997±0.004
	тест	0.924±0.036	0.981±0.015	0.954±0.030	0.951±0.026
PPV	обучение	0.964±0.018	1.000±0.000	0.996±0.004	0.998±0.003
	тест	0.939±0.027	0.983±0.013	0.962±0.024	0.956±0.023
NPV	обучение	0.946±0.014	0.939±0.025	0.983±0.013	0.977±0.014
	тест	0.929±0.023	0.899±0.027	0.913±0.025	0.851±0.035

3.4. Анализ полученных результатов

Из таблицы 12 видно, что при классификации на тестовых наборах данных результаты несколько хуже, чем на обучающих наборах. Стоит отметить, что это общая ситуация, которая связана с переобучением моделей. Все же в реальных условиях к достигаемым результатам диагностического теста наиболее близки будут те оценки метрик, которые получены на тестовом наборе данных. Таким образом, для выбора

оптимального классификатора будем использовать именно результаты на тестовых наборах (таблица 21). В данной работе были достигнуты высокие значения метрик ROC AUC. Здесь сложно выделить какой-то определенный алгоритм в силу того, что различия между средними значениями статистически не значимы (при уровне значимости α=0.05). Отметим, что в данной работе данный показатель ROC AUC можно интерпретировать как вероятность, с которой случайно выбранному пациенту с РЛ будет присвоен вес больший, чем случайному здоровому пациенту. Переходя к основным метрикам, оцениваемым в пилотных работах по разработке диагностических тестов на основе МС, а именно, Se и Sp, реже PPV и NPV, стоит заметить, что в реальных условиях проведения скринингового обследования соотношение количества пациентов с заболеванием к количеству здоровых пациентов будет заметно отличаться от 1:1. Поэтому рассмотрение диагностического теста в первую очередь начинается с Sp и NPV. Соответствующие величины должны быть не менее 0.98-0.99, в противном случае резко возрастает количество неоправданных биопсий, сопряжённых с риском осложнений, увеличивается использование дополнительных методов обследования, а также, стоимость обследования. Требования к Se и PPV менее жесткие – для Se не менее 0.90, для прогностичности отрицательного результата – 0.85. По полученным результатам можно увидеть, что вышесказанным требованием удовлетворяет модель на основе kNN классификатора с Se – 0.905±0.026, Sp – 0.981±0.015, PPV – 0.983±0.013, NPV – 0.899±0.027.

Таблица 21. Результаты классификации для моделей на тестовых наборах данных

Метрика	**Алгоритм классификации**			
	LR	**kNN**	**SVM**	**RF**
ROC AUC	0.959±0.012	0.961±0.018	0.972±0.014	0.960±0.014
Acc	0.931±0.018	0.940±0.016	0.936±0.016	0.898±0.023
Se	0.938±0.021	0.905±0.026	0.920±0.025	0.853±0.041
Sp	0.924±0.036	0.981±0.015	0.954±0.030	0.951±0.026
PPV	0.939±0.027	0.983±0.013	0.962±0.024	0.956±0.023
NPV	0.929±0.023	0.899±0.027	0.913±0.025	0.851±0.035

Тем не менее набор используемых классификаторов позволяет нам оценить важность признака в классификационной модели. Так, например, важность признаков в LR можно представить в виде абсолютных значений коэффициентов модели (рисунок 17). Так как в данной модели используется L1 регуляризация, то для наименее важных признаков модель при обучении обнуляет коэффициенты.

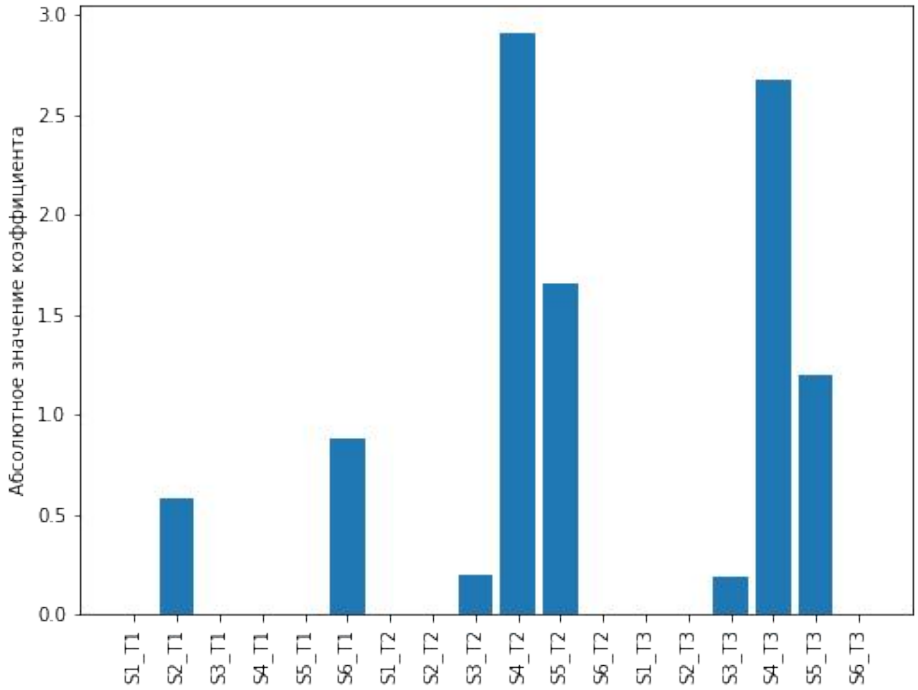

Рисунок 17. Важность признаков в LR на основе абсолютных значений коэффициентов модели

Поскольку каждый сенсор используется при различных температурах, мы ожидаем некоторой коллинеарности между данными одного сенсора при этих температурах. Таким образом, эффективное количество датчиков меньше 18. Это также можно наблюдать при оценке важности признаков (рисунок 14) в модели LR. Используя подобные подходы в исследовании вклада признаков, можно оптимизировать качественный состав сенсоров в рабочем наборе. Также из рисунка 14 можно выделить сенсор S4 как наиболее важный при классификации, что было отмечено ранее при анализе попарных графиков рассеяния.

Результаты данного пилотного исследования показали применимость онлайн-анализа ВВ с помощью массива МО газочувствительных сенсоров в диагностических

целях. Разделение пациентов группы РЛ и группы здоровых наблюдалось с приемлемыми уровнями чувствительности и специфичности. В таблице 22 для сравнительной оценки к проведенным пилотным исследованиям по разделению групп РЛ и группы здоровых добавлены результаты по текущей работе с использованием модели на основе kNN.

Таблица 22. Сравнение полученных результатов со другими пилотными работами с использованием газочувствительных МО сенсоров (Se – чувствительность, Sp – специфичность, Acc – точность)

Ссылка	**Характеристика выборки**	**Se**	**Sp**	**Acc**
Настоящая работа (kNN)	N=118 (65 РЛ, 53 контроль)	90.5 ± 2.6%	98.1 ± 1.5%	94.0 ± 1.6%
[70]	N=101 (43 РЛ, 58 контроль)	95.3%	90.5%	92.6%
[72]	N=18 (9 РЛ, 9 контроль)	100%	88.9%	94.4%
[73]	N=89 (47 РЛ, 42 контроль)	93.6%	83.3%	-
[74]	N=76 (31 РЛ, 45 контроль)	-	-	88%
[75]	N=37 (12 РЛ, 25 контроль)	83%	88%	-
[76]	N=84 (32 РЛ, 52 контроль)	85%	84%	-
[77]	N=290 (144 РЛ, 146 контроль)	94.4%	32.9%	-
[78]	N=145 (52 РЛ, 93 контроль)	83%	84%	-
[79]	N=16 (6 РЛ, 10 контроль)	85.7%	100%	93.8%

Реализация онлайн-режима дает возможность избежать дополнительных процедур отбора и пробоподготовки, которые могут негативно повлиять на качество результатов из-за потери или загрязнения ЛОС. Поскольку использование пробоотборных пакетов может привести к загрязнению или неконтролируемой сорбции важных ЛОС, это сильно влияет на результаты. А такие процедуры, как предварительная очистка легких пациента медицинским воздухом, могут приводить к неконтролируемому изменению профиля ЛОС пациента и, как следствие, некорректному разделению пациентов.

Выводы

Эффекты памяти также значительно уменьшаются в онлайн-режиме с прокачиванием окружающего воздуха при повышенной скорости. Разработанный подход обеспечивает адекватные результаты анализа ВВ даже при использовании ограниченного количества сенсоров. С другой стороны, за счет использования трех температурных режимов – количество датчиков увеличивается в три раза, поскольку переключение с одного режима на другой значительно изменяет относительную чувствительность каждого сенсора.

Результаты, полученные в данной работе, показывают возможность внедрения представленной МС системы в диагностическую практику. В частности, значения основных критериев информативности разработанного диагностического теста (чувствительность, специфичность, PPV, NPV, ROC AUC), которые имеют наиболее важное значение для медицинского персонала при оценке диагностических систем, являются одними из наиболее высоких по сравнению с другими исследованиями (таблица 14), упомянутыми выше [70-79].

Глава 4. Разработка метода переноса градуировочной зависимости и стандартизации откликов между двумя мультисенсорными системами

4.1. Описание дизайна исследования

Как уже упоминалось ранее, существующие технологии изготовления МО сенсоров не позволяют получить такие сенсоры, которые имели бы идентичные характеристики и, следовательно, и идентичный характер отклика к аналиту. Данный факт ограничивает использование разработанной модели классификации для других МС, состоящих из набора соответствующих по составу сенсоров. И полученные критерии информативности для разработанной модели классификации на новых образцах можно ожидать лишь в том случае, если образцы были измерены на том же наборе сенсоров.

Для проверки возможности работы двух МС без каких-либо корректировок был поставлен следующий эксперимент. Регистрировали отклики градуировочных образцов для двух МС (МС 2.1 и МС 2.2). Информация о наборе сенсоров МС представлена в разделе 2.1, информация о составе и концентрациях градуировочных образцов представлена в разделе 2.4.2 в таблицах 13 и 14. На обучающем наборе откликов градуировочных образцов массивов МС 2.1 и МС 2.2 были обучены модели мультиклассовой классификации на основе метода опорных векторов Краммера-Зингера (MSVM) MSVM 2.1 и MSVM 2.2 соответственно. Далее на тестовом наборе откликов градуировочных образцов проводили классификацию для массива МС 2.1 с помощью модели MSVM 2.1 и для массива МС 2.2 с помощью моделей MSVM 2.1 и MSVM 2.2. В итоге получили точность классификации для следующих комбинаций «модель – источник тестового набора»: MSVM 2.1 - МС 2.1, MSVM 2.1 - МС 2.2, MSVM 2.2 - МС 2.2. На рисунке 18 представлены полученные результаты на тестовых наборах откликов для 15 случайных разбиений данных на обучающий и тестовый наборы. Разбиение данных проводилось в соотношении 70%/30% соответственно.

Рисунок 18. Оценка изменения точности классификации при использовании для тестового набора МС 2.2 модели классификации, обученной на образцах МС 2.1 (MSVM 2.1 – МС 2.2). На графике представлены средние значения точности для 15 случайных разбиений исходного набора откликов на обучающих набор и тестовый в соотношении 70/30. Стандартное отклонение для каждой из трех серий равно нулю.

По полученным расчетам видно, что модели с обучающим и тестовым наборами образцов, измеренных на одной и той же МС, почти безошибочно классифицируют тестовые образцы (1.000 для MSVM 2.1 – МС 2.1 и 0.985 для MSVM 2.2 – МС 2.2); при использовании разных источников данных для обучения модели и теста результаты классификации несколько хуже (0.892 для MSVM 2.1 - МС 2.2). Полученные результаты еще раз подтверждают, что для модели, обученной на данных одного набора сенсора, не стоит ожидать сравнимых результатов точности классификации на данных тех же образцов другого массива сенсоров с аналогичными, но физически различными сенсорами. Поэтому для решения поставленной задачи был рассмотрен вариант работы с применением математических преобразований для стандартизации данных между МС.

Для изучения возможности совместной работы двух массивов аналогичных сенсоров посредством корректировки откликов одного из наборов с помощью методов

переноса градуировочных зависимостей были поставлены эксперименты на модельных газовых смесях (воздух – ЛОС) в соответствии со схемой, приведенной на рисунке 11.

Принцип методов переноса градуировочной зависимости, основанных на коррекции откликов массива сенсоров, описаны в главе 1.4.1 настоящей работы. В данной работе для корректировки откликов использовали четыре метода:

Метод однофакторной стандартизации (**UDS**) отклика. Для каждой пары соответствующих сенсоров из двух систем (МС 2.1 и МС 2.2) с помощью метода наименьших квадратов (МНК) определяли коэффициенты линейной регрессии вида:

$$X_j^{\text{МС 2.1}} = k_j * X_j^{\text{МС 2.2}} + b_j \quad (25)$$

Далее тестовый набор МС 2.2 корректировали с учетом найденной связи.

Метод однофакторной стандартизации отклика без использования свободного члена регрессии (**UDSwoi**). Принцип аналогичен методу UDS, за исключением того, что найденное значение свободного члена b_j не используется в уравнении связи:

$$X_j^{\text{МС 2.1}} = k_j * X_j^{\text{МС 2.2}} \quad (26)$$

Метод прямой стандартизации с L1 регуляризацией (**DS-L1R**). Данный тип регуляризации позволяет настроить регуляризационный коэффициент таким образом, чтобы получить требуемое количество ненулевых членов в построенной регрессионной модели. Такая особенность позволяет использовать процедуру стандартизации отклика с количеством стандартизационных образцов меньшим, чем количество независимых переменных в системе. В оптимизируемый функционал модели вводится дополнительное слагаемое регуляризации, и оценка параметров модели $\hat{\beta}$ выражается следующей формулой:

$$\hat{\beta} = arg\ min\left(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}\beta_j x_{ij})^2 + \lambda|\beta|\right) \quad (27)$$

Прямая стандартизация с использованием регрессии на латентные структуры (**DS-PLS2**). Вкратце, в данном подходе производится одновременная декомпозиция матриц откликов стандартизационных образцов для МС 2.1 и МС 2.2:

$$\mathrm{X}^{\text{МС 2.2}} = TP^T + E; \qquad X^{\text{МС 2.1}} = UQ^T + F; \qquad T = XW + G; \quad (28)$$

Построение проекций происходит согласованно с максимизацией корреляции между соответствующими векторами $X^{MC\ 2.2}$-счетов t_a и $X^{MC\ 2.1}$-счетов u_a с учетом заданного количества главных компонент.

Эффективность переноса градуировочных зависимостей и возможность масштабирования модели классификации с одного массива сенсоров на другие в данном эксперименте оценивали по точности мультиклассовой классификации.

Отдельно стоит уделить внимание способу отбора стандартизованных образцов. Напомним, что под этим термином понимают некоторое количество идентичных образцов, измеренных на всех приборах, для которых необходима коррекция отклика. В аналитической химии для решения такого рода задач используют в первую очередь стандартные образцы, но при анализе проб сложного состава, для которых нет аттестованных образцов, используется вышеуказанный термин. Помимо рандомизированного отбора также обращаются к математическим алгоритмам. Наиболее часто в научных работах для отбора минимального по количеству образцов, но близкого по репрезентативности ко всей выборке образцов используют алгоритм Кеннарда-Стоуна [95]. В данной работе была поставлена задача оценить эффективность стандартизации в среднем, т.е. без учета влияния качественного состава набора образцов для стандартизации и учета разбиения данных для обучения модели и теста, поэтому для получения устойчивых результатов проводили многократные вычисления точности классификации при рандомизированных выборках образцов для стандартизации и рандомизированном разбиение набора данных на обучающую и тестовую выборку соответственно.

Чтобы не скомпрометировать результаты эксперимента было обеспечено отсутствие стандартизационных образцов в тестовых наборах данных.

4.2. Оценка результатов стандартизации при классификации индивидуальных образцов ЛОС

В результате многократного проведения экспериментов в соответствии со схемой, представленной на рисунке 11 для каждого алгоритма стандартизации отклика получены усредненные значения точности мультиклассовой классификации образцов для каждого

количества стандартизационных образцов в исследуемом диапазоне (от 2 до 11 включительно). Данные результаты представлены на рисунке 19.

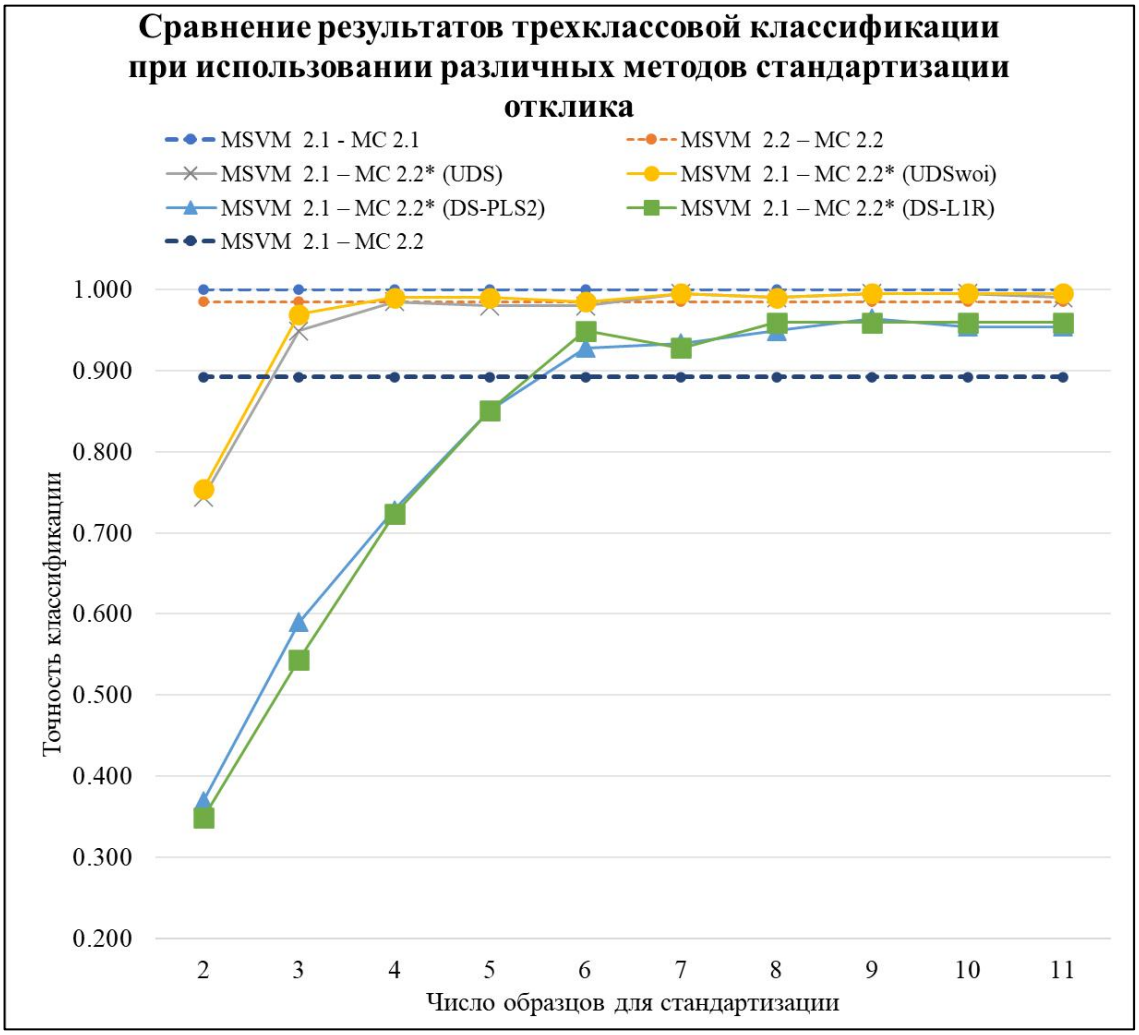

Рисунок 19. Зависимость средней точности классификации на тестовом наборе данных для различных методов стандартизации откликов от количества использованных стандартизационных образцов. (Каждая точка на графике соответствует среднему значению точности классификации по 15 повторениям эксперимента с рандомизированным отбором стандартизационных образцов и разбиением данных на обучающую и тестовую выборки в соотношении 70/30)

Для изучения воспроизводимости результатов для каждого значения точности классификации были оценены доверительные интервалы (показано на рисунке 20). Можно заметить, что для большинства моделей с использованием методов стандартизации при малом количестве образцов доверительные интервалы достаточно широкие и уменьшаются с ростом числа стандартизационных образцов.

Предположительно, что с увеличением числа стандартизационных образцов уменьшается вклад тех стандартизационных образцов, которые плохо описывают общее различие между сенсорами. Исключение составляет модель с регуляризатором DS-L1R, где ширина доверительных не имеет какой-либо тенденции к увеличению или сужению и находится на одном уровне внутри рассматриваемого диапазона числа стандартизационных образцов.

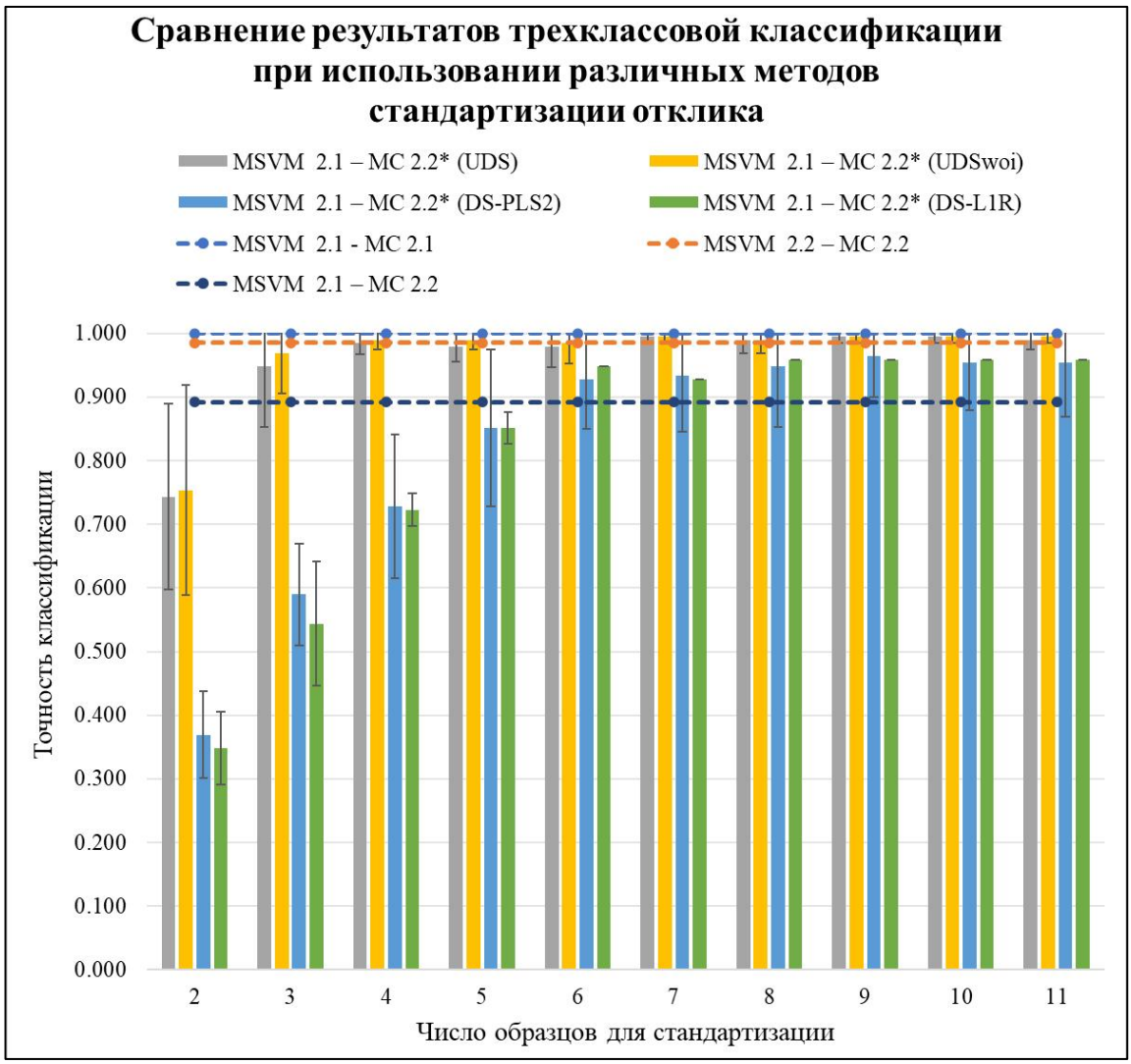

Рисунок 20. Зависимость средней точности классификации на тестовом наборе данных для различных методов стандартизации откликов от количества использованных стандартизационных образцов. (Каждая точка на графике соответствует среднему значению точности классификации по 15 повторениям эксперимента с рандомизированным отбором стандартизационных образцов и разбиением данных на обучающую и тестовую выборки в соотношении 70/30) Также на графике представлены планки погрешностей, соответствующие доверительным интервалам.

На основании полученных результатов можно выделить методы UDS и UDSwoi как наиболее приемлемые для стандартизации откликов МС 2.2 так как именно для этих методов удается достичь уровень точности классификации близкий к уровню самой модели MSVM 2.2 – МС 2.2 за наименьшее количество образцов для стандартизации, равное 4.

4.3. Оценка результатов стандартизации при классификации смесей ЛОС

В исследовательских работах, связанных с количественным многокомпонентным анализом ВВ авторы часто указывают на то, что различие профиля ЛОС между группой здоровых людей и пациентов с патологией заключается не в отсутствии или наличии определённых ЛОС, а диапазоном концентраций. Поэтому в данной работе была также смоделирована задача по воспроизводству двух групп образцов с идентичным качественным составом, но различающимися по количественному составу для некоторых компонентов смеси. В таблице 15 представлены составы смесей с концентрациями ЛОС.

Таблица 15. Состав газовых смесей (ГС 1 и ГС 2) для моделирования задачи классификации

№ смеси	**Компонент смеси**	**Концентрация, ppm**
1	пропан-1-ол	33
	н-гептан	17
	о-ксилол	20
2	пропан-1-ол	26 (-20%)
	н-гептан	17 (+ 0%)
	о-ксилол	24 (+20%)

Отметим, что концентрации каждого из аналитов для смесей 1 и 2 были подобраны таким образом, чтобы извлекаемый отклик для большинства сенсоров был близким к отклику, получаемому при анализе образцов ВВ с точностью до порядка.

Для каждой системы МС 2.1 и МС 2.2 было приготовлено и измерено по 8 образцов каждой смеси ГС 1 и ГС 2. В соответствии со схемой эксперимента, представленной на

рисунке 16, для каждого алгоритма стандартизации отклика получены усредненные значения точности бинарной классификации образцов для каждого количества стандартизационных образцов в исследуемом диапазоне (от 2 до 7 включительно). Полученные результаты представлены на рисунке 21.

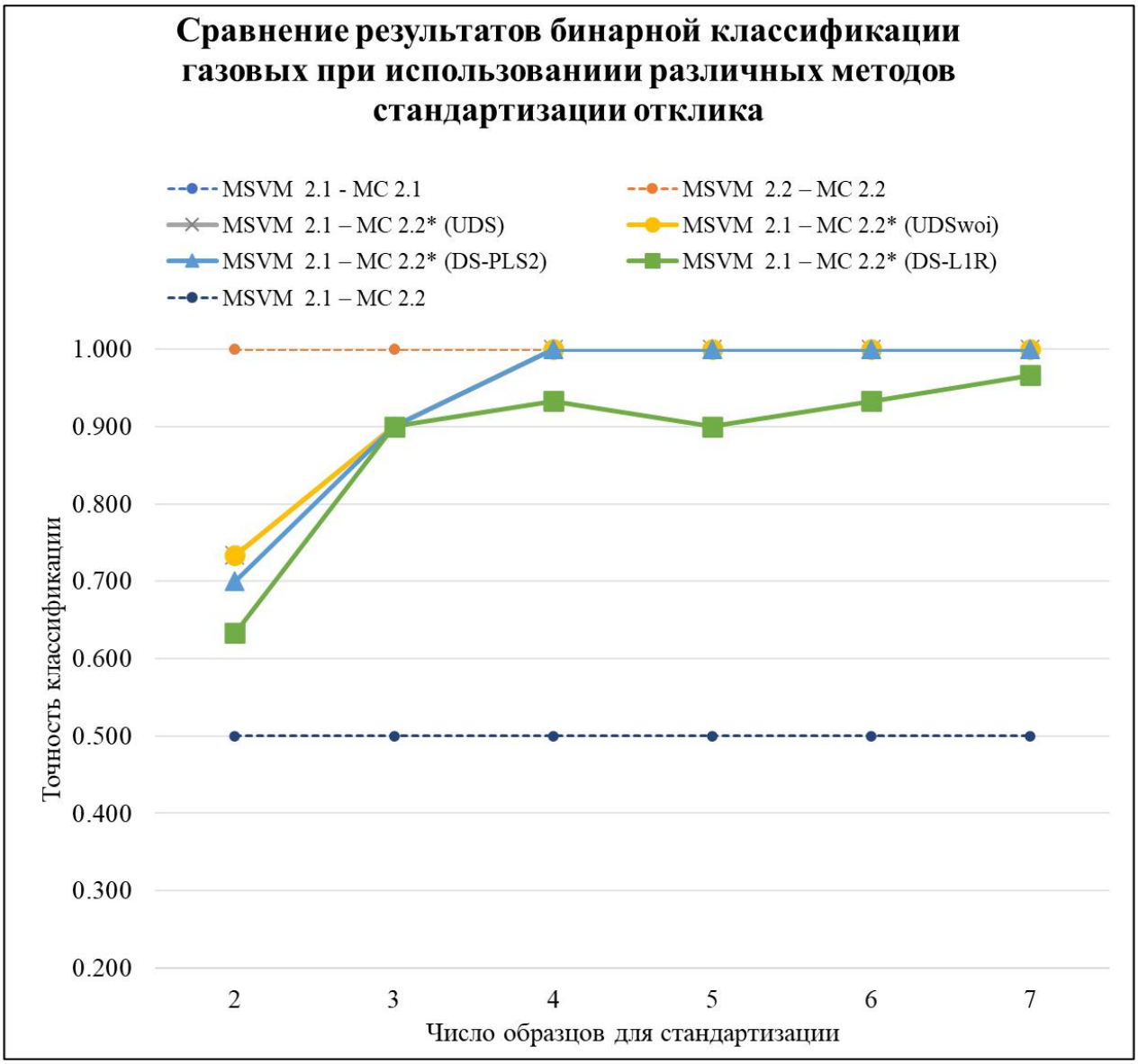

Рисунок 21. Зависимость средней точности классификации на тестовом наборе образцов (газовые смеси 1 и 2) для различных методов стандартизации откликов от количества использованных стандартизационных образцов. (Каждая точка на графике соответствует среднему значению точности классификации по 15 повторениям эксперимента с рандомизированным отбором стандартизационных образцов и разбиением данных на обучающую и тестовую выборки в соотношении 70/30)

Для изучения воспроизводимости результатов для каждого значения точности классификации были оценены доверительные интервалы (Рисунок 22). В данном случае максимальная воспроизводимость результатов достигается для всех методов стандартизации, кроме DS-L1R, для которого доверительные интервалы остаются на одном уровне в рассматриваемом диапазоне стандартизационных образцов. Как и в предыдущей задаче минимально необходимое количество образцов для стандартизации откликов МС 2.2 составило 4.

Для качественной оценки результата стандартизации откликов был использован метод PCA. А именно, по данным МС 2.1 построена PCA модель. Относительная объясненная дисперсия для первой компоненты составила 80.5%, для второй – 8.1%. Далее данные МС 2.2 были спроецированы в пространство главных компонент PCA модели. Также были получены данные МС 2.2* (UDS): данные МС 2.1, подвергнутые корректировке с помощью однофакторной стандартизации по трем стандартным образцам после проекции на главные компоненты. На рисунке 23 представлены отклики градуировочные образцов пропан-1-ола, н-гептана и о-ксилола тестового набора для МС 2.1, МС 2.2 и МС 2.2* в пространстве первых двух главных компонент.

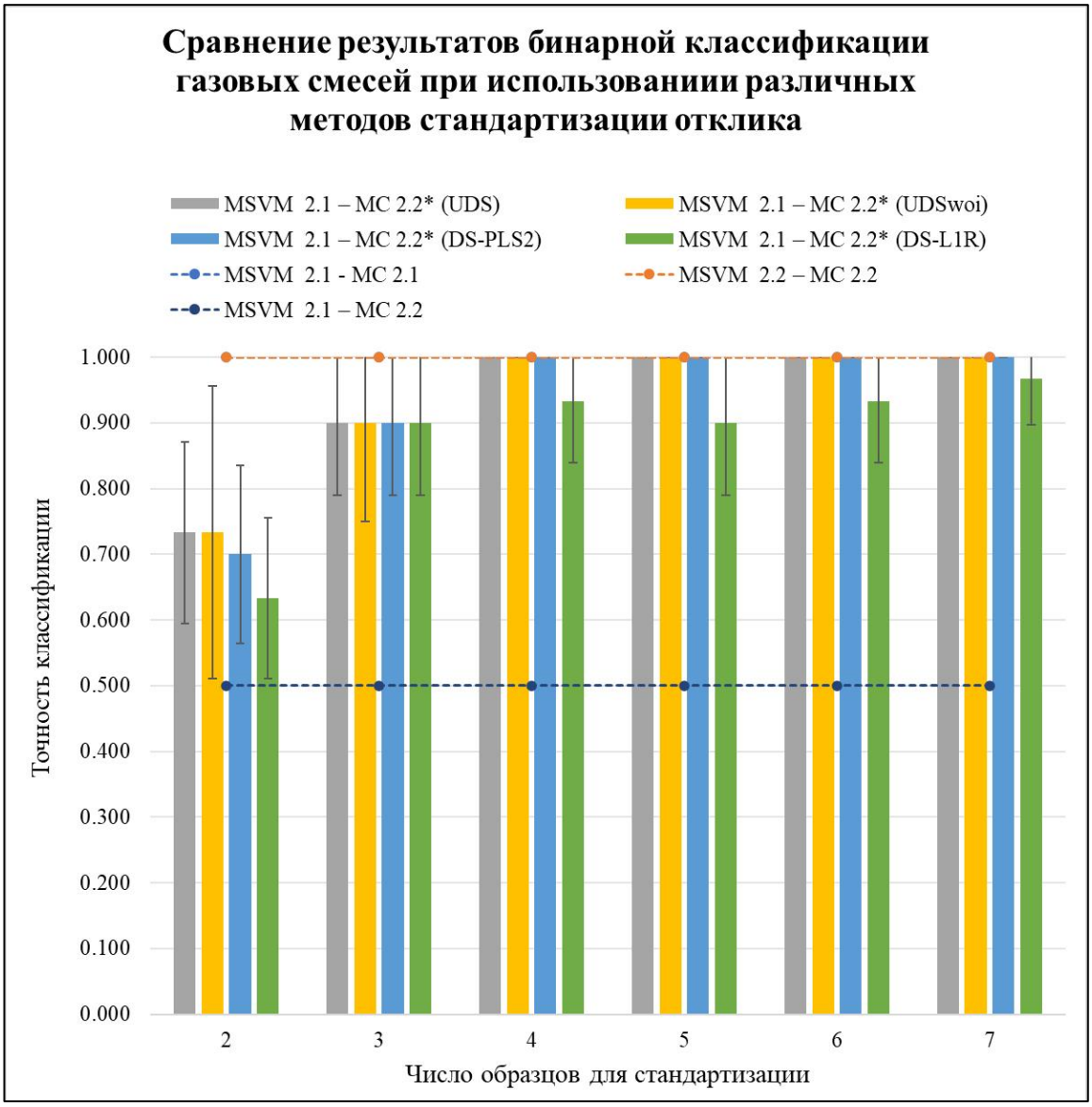

Рисунок 22. Зависимость средней точности классификации на тестовом наборе образцов (газовые смеси 1 и 2) для различных методов стандартизации откликов от количества использованных стандартизационных образцов. (Каждая точка на графике соответствует среднему значению точности классификации по 15 повторениям эксперимента с рандомизированным отбором стандартизационных образцов и разбиением данных на обучающую и тестовую выборки в соотношении 70/30) Также на графике представлены планки погрешностей, соответствующие доверительным интервалам.

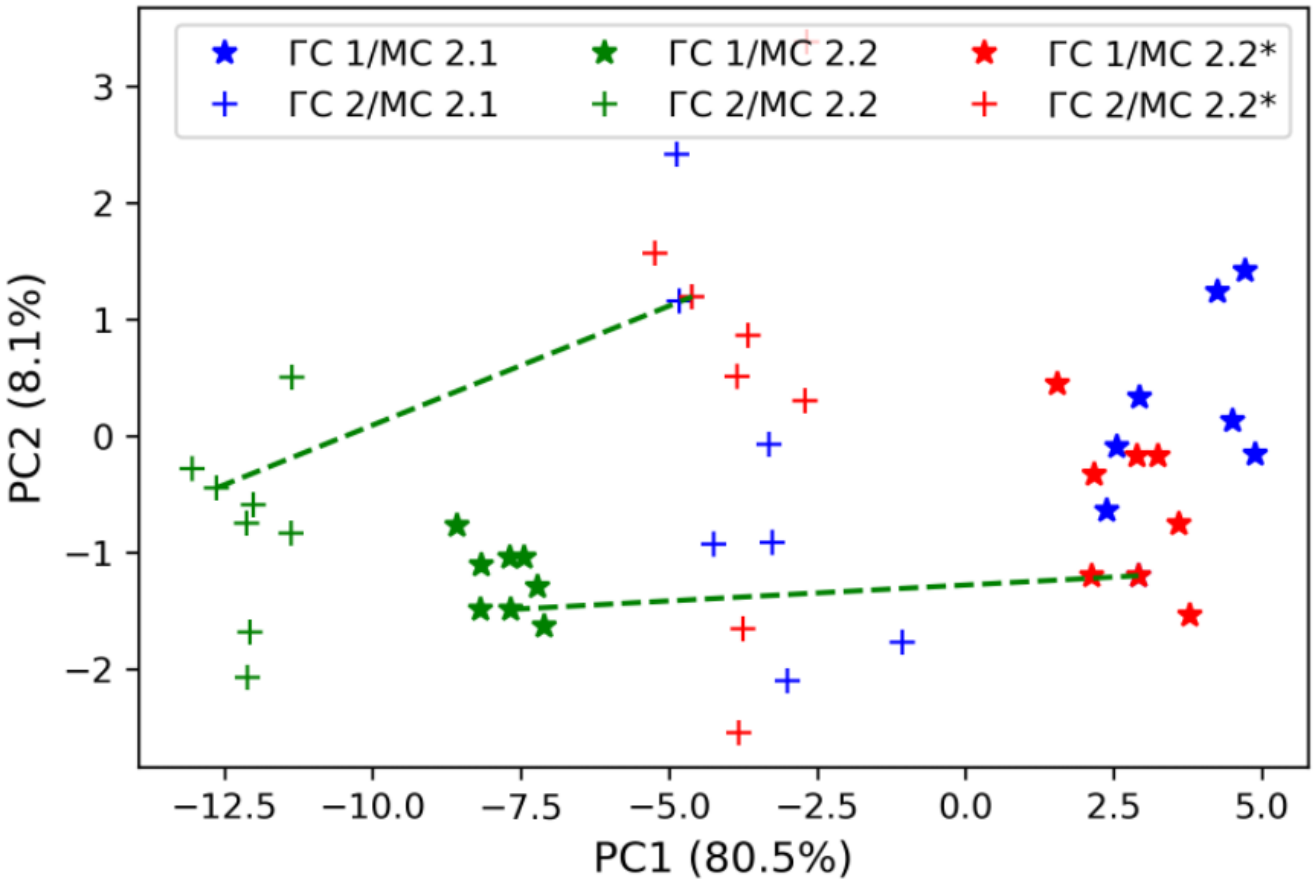

Рисунок 23. Визуальная интерпретация процесса корректировки откликов и взаимного расположения образцов в пространстве главных компонент (в скобках указана доля объясненной дисперсии для каждой главной компоненты). Синим цветом отмечены образцы смесей, измеренные на МС 2.1, зеленым – образцы смесей на МС 2.2, красным – образцы МС 2.2, скорректированные методом UDS по двум стандартизационным образцам (для этих двух образцов с помощью зеленой пунктирной линии показано отображение после корректировки откликов). PC1 – ось первой главной компоненты, PC2 – ось второй главной компоненты.

Выводы

Результаты, полученные в главе 4 демонстрируют возможность проведения стандартизации отклик одной МС для совместного использования одной модели классификации, обученной на откликах другой МС. Методы однофакторной стандартизации продемонстрировал наибольшую и приемлемую точность классификации при минимальном количестве стандартизационных образцов, равном четырем.

Заключение

Предложена, создана и апробирована схема онлайн-анализа ВВ с помощью системы газочувствительных металлооксидных сенсоров не требующая дополнительной пробоподготовки.

С помощью разработанной системы проведено сравнительное обследование группы здоровых (53) и больных РЛ людей (65). На основании полученных результатов разработан и апробирован алгоритм обработки экспериментальных данных, позволяющий эффективно выделять больных РЛ с чувствительностью (90.5 ± 2.6)%, специфичностью (98.1 ± 1.5)%, точностью (94.0 ± 1.6)%, ROC AUC 0.961 ± 0.018, прогностичностью положительного результата (98.3 ± 1.3)% и прогоностичностью отрицательного результата (89.9 ± 2.7)% безотносительно внешних факторов состояния пациента (возраст, пол, курение и др.) и основываясь исключительно на откликах мультисенсорной системы.

Проведены эксперименты по исследованию возможности переноса градуировочных зависимостей между двумя мультисенсорными системами, показавшие необходимость использования переноса градуировочной зависимости для одной из систем посредством стандартизации отклика. На модельных задачах мультиклассовой классификации смесей (воздух – ЛОС) и бинарной классификации 2 смесей с идентичным качественным составом, но различающимся по концентрациям индивидуальных компонентов не более чем 20% показана эффективность использования метода однофакторной стандартизации, для которого достигается максимальная точность при наименьшем количестве образцов для стандартизации, равном четырем.

Список сокращений и условных обозначений

ВВ – выдыхаемый воздух

ЛОС – летучие органические соединения

МО – металлооксидный

МС – мультисенсорная система

ПАВ – поверхностные акустические волны

РЛ – рак легкого

ско – среднеквадратичное отклонение

ТФМЭ – твердофазная микроэкстракция

ХОБЛ – хроническая обструктивная болезнь легких

Acc – точность теста

DS – прямая стандартизация

DS-L1R – прямая стандартизация с L1 регуляризатором

DS-PLS2 – прямая стандартизация с регрессией на латентные структуры

FN – ложноотрицательный результат

FP – ложноположительный результат

GS-MS – газовая хромато-масс-спектрометрия

kNN – метод k ближайших соседей

LDA – линейный дискриминантный анализ

LR – логистическая регрессия

MCC-IMS – поликапиллярная спектрометрия ионной подвижности

MSVM – мультиклассовый метод опорных векторов

NPV – прогностичность отрицательного результата

PC – главная компонента

PCA – метод главных компонент

PCA – метод главных компонент

PDS – фрагментарно-прямая стандартизация

PPV – прогностичность положительного результата

PTR-MS – масс-спектрометрия с реакцией переноса протона

RF – метод «случайного леса»

RMSE – среднеквадратичная ошибка

ROC – кривая операционной характеристики

ROC AUC – площадь под кривой операционной характеристики

Se – чувствительность теста

SIFT-MS – масс-спектрометрия выбранных ионов в потоке

Spe – специфичность теста

SVM – метод опорных векторов

TN – истинно отрицательный результат

TP – истинно положительный результат

UDS – одномерная прямая стандартизация

UDSwoi – одномерная прямая стандартизация без использования коэффициента свободного члена

Список литературы

1. Arseniev A. et al. Combined diagnostics of lung cancer using exhaled breath analysis and sputum cytology // Probl. Oncol. 2020. Vol. 66, № 4. P. 381–384.
2. Ganeev A.A. et al. Analysis of exhaled air for early-stage diagnosis of lung cancer: opportunities and challenges // Russ. Chem. Rev. 2018. Vol. 87, № 9. P. 904–921.
3. Saalberg Y., Wolff M. VOC breath biomarkers in lung cancer // Clin. Chim. Acta. Elsevier B.V., 2016. Vol. 459. P. 5–9.
4. Rattray N.J.W. et al. Taking your breath away: Metabolomics breathes life in to personalized medicine // Trends Biotechnol. Elsevier Ltd, 2014. Vol. 32, № 10. P. 538–548.
5. Xu F., Zou L., Ong C.N. Multiorigination of chromatographic peaks in derivatized GC/MS metabolomics: A confounder that influences metabolic pathway interpretation // J. Proteome Res. 2009. Vol. 8, № 12. P. 5657–5665.
6. Zhou J. et al. Review of recent developments in determining volatile organic compounds in exhaled breath as biomarkers for lung cancer diagnosis // Anal. Chim. Acta. Elsevier Ltd, 2017. Vol. 996. P. 1–9.
7. Atkinson A.J. et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework // Clin. Pharmacol. Ther. 2001. Vol. 69, № 3. P. 89–95.
8. Hakim M. et al. Volatile organic compounds of lung cancer and possible biochemical pathways // Chem. Rev. 2012. Vol. 112, № 11. P. 5949–5966.
9. Rocco G. et al. Breathprinting and Early Diagnosis of Lung Cancer // Journal of Thoracic Oncology. International Association for the Study of Lung Cancer, 2018. Vol. 13, № 7. 883–894 p.
10. Amann A. et al. Methodological issues of sample collection and analysis of exhaled breath // Exhaled Biomarkers. 2010.
11. Turner C. Techniques and issues in breath and clinical sample headspace analysis for disease diagnosis // Bioanalysis. 2016. Vol. 8, № 7.
12. Pleil J.D., Lindstrom A.B. Collection of a single alveolar exhaled breath for volatile organic compounds analysis // Am. J. Ind. Med. 1995. Vol. 28, № 1. P. 109–121.
13. Miekisch W. et al. Impact of sampling procedures on the results of breath analysis // J. Breath Res. 2008. Vol. 2, № 2.
14. Miekisch W., Schubert J.K. From highly sophisticated analytical techniques to life-saving diagnostics: Technical developments in breath analysis // TrAC - Trends Anal. Chem. 2006. Vol. 25, № 7. P. 665–673.

15. McCafferty J.B. et al. Effects of breathing pattern and inspired air conditions on breath condensate volume, pH, nitrite, and protein concentrations // Thorax. 2004. Vol. 59, № 8. P. 694–698.
16. Соодаева С.К., Климанов И.А. Нарушения окислительного метаболизма при заболеваниях респираторного тракта и современные подходы к антиоксидантной терапии. 2009. P. 34–37.
17. Horváth I. et al. Exhaled breath condensate: Methodological recommendations and unresolved questions // Eur. Respir. J. 2005. Vol. 26, № 3. P. 523–548.
18. Kubáň P., Foret F. Exhaled breath condensate: Determination of non-volatile compounds and their potential for clinical diagnosis and monitoring. A review // Anal. Chim. Acta. 2013. Vol. 805. P. 1–18.
19. Buszewski B. et al. Human exhaled air analytics: Biomarkers of diseases // Biomed. Chromatogr. 2007. Vol. 21. P. 553–566.
20. Krilaviciute A. et al. Detection of cancer through exhaled breath: A systematic review // Oncotarget. 2015. Vol. 6, № 36. P. 38643–38657.
21. US EPA. Method TO-15: Compendium of methods for the determination of toxic organic compounds in ambient air // EPA Methods. 1999. № January. P. 1–32.
22. Beauchamp J. et al. On the use of Tedlar® bags for breath-gas sampling and analysis // J. Breath Res. 2008. Vol. 2, № 4.
23. Schmekel B., Winquist F., Vikström A. Analysis of breath samples for lung cancer survival // Anal. Chim. Acta. Elsevier B.V., 2014. Vol. 840. P. 82–86.
24. Trabue S.L., Anhalt J.C., Zahn J.A. Bias of Tedlar Bags in the Measurement of Agricultural Odorants // J. Environ. Qual. 2006. Vol. 35, № 5. P. 1668–1677.
25. Mieth M. et al. Multibed Needle Trap Devices for on Site Sampling and Preconcentration of Volatile Breath Biomarkers. 2009. Vol. 81, № 14. P. 5851–5857.
26. Hyšpler R. et al. Determination of isoprene in human expired breath using solid-phase microextraction and gas chromatography-mass spectrometry // J. Chromatogr. B Biomed. Sci. Appl. 2000. Vol. 739, № 1. P. 183–190.
27. Mutlu G.M. et al. Collection and analysis of exhaled breath condensate in humans // Am. J. Respir. Crit. Care Med. 2001. Vol. 164, № 5. P. 731–737.
28. Dyne D., Cocker J., Wilson H.K. A novel device for capturing alveolar breath samples for solvent analysis // J. Automat. Chem. 1997. Vol. 19, № 2. P. 59.
29. Poli D. et al. Exhaled volatile organic compounds in patients with non-small cell lung cancer: Cross sectional and nested short-term follow-up study // Respir. Res. 2005. Vol. 6. P. 1–10.
30. Kusano M., Mendez E., Furton K.G. Development of headspace SPME method for analysis of

volatile organic compounds present in human biological specimens // Anal. Bioanal. Chem. 2011. Vol. 400, № 7. P. 1817–1826.

31. Van Den Velde S. et al. Differences between alveolar air and mouth air // Anal. Chem. 2007. Vol. 79, № 9. P. 3425–3429.
32. Amann A. et al. Applications of breath gas analysis in medicine // Int. J. Mass Spectrom. 2004. Vol. 239, № 2–3. P. 227–233.
33. Alonso M., Castellanos M., Sanchez J.M. Evaluation of potential breath biomarkers for active smoking: Assessment of smoking habits // Anal. Bioanal. Chem. 2010. Vol. 396, № 8. P. 2987–2995.
34. Alonso M. et al. Capillary thermal desorption unit for near real-time analysis of VOCs at sub-trace levels. Application to the analysis of environmental air contamination and breath samples // J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 2009. Vol. 877, № 14–15. P. 1472–1478.
35. Scheepers P.T.J. et al. Determination of exposure to benzene, toluene and xylenes in Turkish primary school children by analysis of breath and by environmental passive sampling // Sci. Total Environ. Elsevier B.V., 2010. Vol. 408, № 20. P. 4863–4870.
36. Gordon S.M. et al. Volatile organic compounds as breath biomarkers for active and passive smoking // Environ. Health Perspect. 2002. Vol. 110, № 7. P. 689–698.
37. Woolfenden E. Sorbent-based sampling methods for volatile and semi-volatile organic compounds in air. Part 1: Sorbent-based air monitoring options // J. Chromatogr. A. Elsevier B.V., 2010. Vol. 1217, № 16. P. 2674–2684.
38. Helmig D. Artifact-free preparation, storage and analysis of solid adsorbent sampling cartridges used in the analysis of volatile organic compounds in air // J. Chromatogr. A. 1996. Vol. 732, № 2. P. 414–417.
39. Calogirou A. et al. Decomposition of terpenes by ozone during sampling on tenax // Anal. Chem. 1996. Vol. 68, № 9. P. 1499–1506.
40. Dewulf J., Van Langenhove H. Anthropogenic volatile organic compounds in ambient air and natural waters: a review on recent developments of analytical methodology, performance and interpretation of field measurements // J. Chromatogr. A. 1999. Vol. 843, № 1–2. P. 163–177.
41. Peng C.Y., Batterman S. Performance evaluation of a sorbent tube sampling method using short path thermal desorption for volatile organic compounds // J. Environ. Monit. 2000. Vol. 2, № 4. P. 313–324.
42. Cao X.L., Nicholas Hewitt C. Build-up of artifacts on adsorbents during storage and its effect on passive sampling and gas chromatography-flame ionization detection of low concentrations of volatile organic compounds in air // J. Chromatogr. A. 1994. Vol. 688, № 1–2. P. 368–374.
43. Materić D. et al. Methods in Plant Foliar Volatile Organic Compounds Research // Appl. Plant

Sci. 2015. Vol. 3, № 12. P. 1500044.

44. Wang C. et al. Noninvasive detection of colorectal cancer by analysis of exhaled breath // Anal. Bioanal. Chem. 2014. Vol. 406, № 19. P. 4757–4763.
45. Wang C., Sahay P. Breath analysis using laser spectroscopic techniques: Breath biomarkers, spectral fingerprints, and detection limits // Sensors. 2009. Vol. 9, № 10. P. 8230–8262.
46. Schwarz K. et al. Breath acetone - Aspects of normal physiology related to age and gender as determined in a PTR-MS study // J. Breath Res. 2009. Vol. 3, № 2.
47. Smith D. et al. Mass spectrometry for real-time quantitative breath analysis // J. Breath Res. 2014. Vol. 8, № 2.
48. Smith D. et al. Quantification of acetaldehyde released by lung cancer cells in vitro using selected ion flow tube mass spectrometry // Rapid Commun. Mass Spectrom. 2003. Vol. 17, № 8. P. 845–850.
49. Rutter A. V. et al. Quantification by SIFT-MS of acetaldehyde released by lung cells in a 3D model // Analyst. 2013. Vol. 138, № 1. P. 91–95.
50. Sulé-Suso J. et al. Quantification of acetaldehyde and carbon dioxide in the headspace of malignant and non-malignant lung cells in vitro by SIFT-MS // Analyst. 2009. Vol. 134, № 12. P. 2419–2425.
51. Baumbach J.I. et al. Significant different volatile biomarker during bronchoscopic ion mobility spectrometry investigation of patients suffering lung carcinoma // Int. J. Ion Mobil. Spectrom. 2011. Vol. 14, № 4. P. 159–166.
52. Lamote K. et al. Detection of malignant pleural mesothelioma in exhaled breath by multicapillary column/ion mobility spectrometry (MCC/IMS) // J. Breath Res. IOP Publishing, 2016. Vol. 10, № 4. P. 46001.
53. Bessa V. et al. Detection of volatile organic compounds (VOCs) in exhaled breath of patients with chronic obstructive pulmonary disease (COPD) by ion mobility spectrometry // Int. J. Ion Mobil. Spectrom. 2011. Vol. 14, № 1. P. 7–13.
54. Arasaradnam R.P. et al. Non-invasive exhaled volatile organic biomarker analysis to detect inflammatory bowel disease (IBD) // Dig. Liver Dis. Editrice Gastroenterologica Italiana, 2016. Vol. 48, № 2. P. 148–153.
55. Wilson A.D. Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath // Metabolites. 2015. Vol. 5, № 1. P. 140–163.
56. Behera B. et al. Electronic nose: A non-invasive technology for breath analysis of diabetes and lung cancer patients // Journal of Breath Research. 2019. Vol. 13, № 2.
57. McWilliams A. et al. Sex and smoking status effects on the early detection of early lung cancer in high-risk smokers using an electronic nose // IEEE Trans. Biomed. Eng. 2015. Vol. 62, № 8.

P. 2044–2054.

58. Chen X. et al. A study of an electronic nose for detection of lung cancer based on a virtual SAW gas sensors array and imaging recognition method // Meas. Sci. Technol. 2005. Vol. 16, № 8. P. 1535–1546.
59. Gasparri R. et al. Volatile signature for the early diagnosis of lung cancer // J. Breath Res. IOP Publishing, 2016. Vol. 10, № 1. P. 16007.
60. Mazzone P.J. et al. Exhaled breath analysis with a colorimetric sensor array for the identification and characterization of lung cancer // J. Thorac. Oncol. International Association for the Study of Lung Cancer, 2012. Vol. 7, № 1. P. 137–142.
61. Shehada N. et al. Silicon Nanowire Sensors Enable Diagnosis of Patients via Exhaled Breath // ACS Nano. 2016. Vol. 10, № 7. P. 7047–7057.
62. Meixner H., Lampe U. Metal oxide sensors // Sensors Actuators, B Chem. 1996. Vol. 33, № 1–3. P. 198–202.
63. Marikutsa A. V. et al. Active sites on the surface of nanocrystalline semiconductor oxides ZnO and SnO 2 and gas sensitivity // Russian Chemical Bulletin. 2017. Vol. 66, № 10.
64. Волькенштейн Ф.Ф. Электронные процессы на поверхности полупроводников при хемосорбции. Наука. Гл. ред. физ.-мат. лит., 1987.
65. Мясников И.А, Сухарев В.Я., Куприянов Л.Ю. З.С.А. Полупроводниковые сенсоры в физико-химических исследованиях. Москва: Наука, 1991. 327 p.
66. Pijolat C. et al. Gas detection for automotive pollution control // Sensors Actuators, B Chem. 1999. Vol. 59, № 2. P. 195–202.
67. Rudnitskaya A. Calibration update and drift correction for electronic noses and tongues // Front. Chem. 2018. Vol. 6, № September.
68. А.В. Ш. Селективное определение газов полупроводниковыми сенсорами. 2005.
69. Baldini C. et al. Electronic nose as a novel method for diagnosing cancer: A systematic review // Biosensors. 2020. Vol. 10, № 8. P. 1–21.
70. Blatt R. et al. Lung cancer identification by an electronic nose based on an array of MOS sensors // IEEE Int. Conf. Neural Networks - Conf. Proc. 2007. P. 1423–1428.
71. Tran V.H. et al. Breath analysis of lung cancer patients using an electronic nose detection system // IEEE Sens. J. 2010. Vol. 10, № 9. P. 1514–1518.
72. Yu K. et al. A portable electronic Nose intended for home healthcare based on a mixed sensor array and multiple desorption methods // Sensor Letters. 2011. Vol. 9, № 2.
73. Wang D. et al. A hybrid electronic noses' system based on MOS-SAW detection units intended for lung cancer diagnosis // J. Innov. Opt. Health Sci. 2012. Vol. 5, № 1. P. 1–7.
74. De Vries R. et al. Integration of electronic nose technology with spirometry: Validation of a

new approach for exhaled breath analysis // J. Breath Res. IOP Publishing, 2015. Vol. 9, № 4. P. 46001.

75. Tan J.L., Yong Z.X., Liam C.K. Using a chemiresistor-based alkane sensor to distinguish exhaled breaths of lung cancer patients from subjects with no lung cancer // J. Thorac. Dis. 2016. Vol. 8, № 10. P. 2772–2783.
76. van Hooren M.R.A. et al. Differentiating head and neck carcinoma from lung carcinoma with an electronic nose: a proof of concept study // Eur. Arch. Oto-Rhino-Laryngology. Springer Berlin Heidelberg, 2016. Vol. 273, № 11. P. 3897–3903.
77. Kort S. et al. Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis // Lung Cancer. Elsevier Ireland Ltd, 2018. Vol. 125. P. 223–229.
78. van de Goor R. et al. Training and Validating a Portable Electronic Nose for Lung Cancer Screening // J. Thorac. Oncol. International Association for the Study of Lung Cancer, 2018. Vol. 13, № 5. P. 676–681.
79. Marzorati D. et al. A Metal Oxide Gas Sensors Array for Lung Cancer Diagnosis Through Exhaled Breath Analysis // Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. 2019.
80. Родионова О.Е., Померанцев А.Л., Ран Н.Н.С. Хемометрика в аналитической химии [Electronic resource]. 2006.
81. Marco S., Gutierrez-Galvez A. Signal and data processing for machine olfaction and chemical sensing: A review // IEEE Sens. J. 2012. Vol. 12, № 11. P. 3189–3214.
82. Leopold J.H. et al. Comparison of classification methods in breath analysis by electronic nose // J. Breath Res. 2015. Vol. 9, № 4. P. 046002.
83. Wlodzimirow K.A. et al. Exhaled breath analysis with electronic nose technology for detection of acute liver failure in rats // Biosens. Bioelectron. Elsevier, 2014. Vol. 53. P. 129–134.
84. Benedek P. et al. Exhaled biomarker pattern is altered in children with obstructive sleep apnoea syndrome // Int. J. Pediatr. Otorhinolaryngol. 2013. Vol. 77, № 8. P. 1244–1247.
85. Hakim M. et al. Diagnosis of head-and-neck cancer from exhaled breath // Br. J. Cancer. Nature Publishing Group, 2011. Vol. 104, № 10. P. 1649–1655.
86. Mazzone P.J. et al. Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array // Thorax. 2007. Vol. 62, № 7. P. 565–568.
87. Breiman L. Random Forests // Mach. Learn. 2001. Vol. 45, № 1. P. 5–32.
88. Crammer K., Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines // J. Mach. Learn. Res. - JMLR. 2002. Vol. 2, № 2. P. 265–292.
89. Rudnitskaya A. et al. Measurements of the effects of wine maceration with oak chips using an electronic tongue // Food Chem. Elsevier Ltd, 2017. Vol. 229. P. 20–27.

90. Pillonel L., Bosset J.O., Tabacchi R. Data transferability between two MS-based electronic noses using processed cheeses and evaporated milk as reference materials // Eur. Food Res. Technol. 2002. Vol. 214, № 2. P. 160–162.
91. Pérez Pavón J.L. et al. Strategies for qualitative and quantitative analyses with mass spectrometry-based electronic noses // TrAC - Trends Anal. Chem. 2006. Vol. 25, № 3. P. 257–266.
92. Deshmukh S. et al. Calibration transfer between electronic nose systems for rapid In situ measurement of pulp and paper industry emissions // Anal. Chim. Acta. 2014. Vol. 841. P. 58–67.
93. Fonollosa J. et al. Evaluation of calibration transfer strategies between Metal Oxide gas sensor arrays // Procedia Eng. Elsevier B.V., 2015. Vol. 120. P. 261–264.
94. Panchuk V. et al. Extending electronic tongue calibration lifetime through mathematical drift correction: Case study of microcystin toxicity analysis in waters // Sensors Actuators, B Chem. Elsevier B.V., 2016. Vol. 237. P. 962–968.
95. Kennard R.W., Stone L.A. Computer Aided Design of Experiments // Technometrics. 1969. Vol. 11, № 1. P. 137–148.
96. Vasiliev A.A. et al. Reducing humidity response of gas sensors for medical applications: Use of spark discharge synthesis of metal oxide nanoparticles // Sensors (Switzerland). 2018. Vol. 18, № 8.
97. Hierlemann A., Gutierrez-Osuna R. Higher-order chemical sensing // Chem. Rev. 2008. Vol. 108. P. 563–613.
98. Malyshev V. V., Pislyakov A. V. Dynamic properties and sensitivity of semiconductor metal-oxide thick-film sensors to various gases in air gaseous medium // Sensors Actuators, B Chem. 2003. Vol. 96, № 1–2. P. 413–434.
99. Alexander Kononov. Breath Analysis [Electronic resource]. 2021. URL: https://github.com/camberbatch/Breath_analysis.
100. Kononov A. et al. Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer // J. Breath Res. 2020. Vol. 14, № 1.
101. Shapiro A.S.S., Wilk M.B. An Analysis of Variance Test for Normality (Complete Samples) Published by : Biometrika Trust Stable URL : http://www.jstor.org/stable/2333709 // Biometrika. 1965. Vol. 52, № 3/4. P. 591–611.

SAINT PETERSBURG STATE UNIVERSITY

Alexander S. Kononov

DEVELOPMENT OF A METHOD FOR LUNG CANCER DIAGNOSIS BASED ON ONLINE ANALYSIS OF EXHALED BREATH USING METAL OXIDE GAS SENSITIVE SENSORS

Dissertation is submitted for the degree
of Candidate of Chemical Sciences

Scientific Specialty 1.4.2. Analytical chemistry

Translation from Russian

Academic supervisor
Doctor of Physics and Mathematics, Professor
Alexander A. Ganeev

Saint Petersburg
2021

Contents

Contents **99**
Introduction **101**
Chapter 1. Literature review **105**
1.1. Potential biomarkers of lung cancer in exhaled breath 105
1.2. Sampling and sample preparation methods in exhaled air analysis 107
1.3. Methods of exhaled breath analysis suitable for the diagnostics of lung cancer 110
1.4. Methods of multivariate data processing 119
Chapter 2. Experimental details **132**
2.1. Description of sensor characteristics 132
2.2. Technique for the preparation of model gas mixtures 137
2.3. Analysis of model gas mixtures and exhaled breath samples in medical research using MS1 138
2.4. Analysis of model gas mixtures for calibration transfer using MS 2.1 and MS 2.2 146
Chapter 3. Development of an online analysis of exhaled breath method for the diagnosis of lung cancer using a multisensory system **151**
3.4. Description of the medical study 151
3.4. Description of patient exhaled breath analysis procedure 152
3.4. Selection of the most effective data processing algorithm and classification model 153
3.4. Analysis of the results 161
Conclusions 165
Chapter 4. Development of a calibration transfer and response standardization method between two multisensory systems **166**
4.1. Description of the study design 166
4.2. Evaluation of standardization results in the classification of individual VOC samples 169
4.3. Evaluation of standardization results in the classification of VOC mixtures 172
Conclusions 175
Conclusions **176**
List of abbreviation **177**
References **179**

The **author's personal contribution** consisted in collecting and analyzing the scientific literature, active participation in the formulation of tasks, research, planning, preparing and conducting experiments, studying the physical and chemical properties of the sensors and processing the data obtained, as well as in the analysis, interpretation and generalization of the results, preparation of reports and publications.

Acknowledgements. I sincerely thank everyone who contributed to the fulfillment of this work. I express special gratitude to Alexander A. Ganeev for his mentoring at all stages of the research work, experimental expertise, teaching critical thinking and the ability to identify the essentials. I thank Dzhagatspanyan E. Igor for the multiple productive discussions on the details of semiconductor sensors and gas analyzers, as well as for his help in preparing articles for publication.

I express my gratitude to my co-authors and colleagues: Boris A. Korotetsky, Anna R. Gubal, Victoria A. Chuchina, Andrey O. Nefedov, Alexey A. Vasilyev, Andrey I, Arseniev.

Finally, I thank my spouse, parents, brother, friends and loved ones for their support.

Introduction

Early detection of lung cancer (LC) is usually associated with a significant improvement in the efficiency of its treatment. However, currently used methods of early detection of lung cancer are not effective enough which leads to late detection of the disease and consequently to a high mortality rate. In this regard the development of a high-throughput and reliable diagnostic method is an important task that needs to be solved as soon as possible. Exhaled breath (EB) analysis to determine a number of organic compounds which are recognized biomarkers of LC is becoming a promising method for early detection of LC. This research area is attracting more and more interest which is confirmed by the increasing number of scientific publications on this topic every year. In this field it is possible to create not only a screening method for early detection of LC but also a method allowing to monitor the condition of an LC patient both before and after treatment. However, in creating such a method a number of often conflicting conditions must be met. The method should have short sampling and analysis times, be relatively cheap and noninvasive and if possible, method should work in online mode. The most important requirement not only for the considered but also for any other methods of LC diagnosis is high levels of specificity and positive predictive value [1]. The corresponding values must be at least 98-99% otherwise the number of unjustified biopsies associated with the risk of complications increases dramatically, the use of additional methods of examination increases, and cost of analysis increases. The requirements for sensitivity and negative predictive value are less stringent – at least 90% for sensitivity and 85% for negative predictive value [2].

At this stage of development of EB analysis all methods of analysis that used for LC diagnosis gas chromatography-mass spectrometry (GC-MS), mass spectrometry with proton transfer reaction (PTR-MS), polycapillary ion mobility spectrometry (MCC-IMS) do not fully meet the requirements: GC-MS has low productivity, high workload and the possibility of using this method only in offline mode. Methods that can be used in online mode: PTR-MS, MCC-IMS – are not sensitive enough. However, all these methods are quite difficult to use for direct control of the state of a patient with LC. It is much easier to use a multisensory system (MS) for recognizing EB images that called "electronic nose" (EN) for which acceptable levels of specificity and positive predictive value can be achieved although they need to be improved as well. It should be noted that the currently existing EN systems do not allow us to fully solve the

problems of diagnostics LC which is mainly due to the properties of the sensors used for this purpose. The disadvantages of the currently used EN include insufficient cross sensitivity for the main biomarkers of LC and insufficient long-term stability of their analytical characteristics. These disadvantages are common to many types of sensors but less to metal oxide sensors that however have another disadvantage – manufacturing variability. Existing manufacturing techniques do not allow sensors with identical characteristics and consequently an identical character of response to an analyte. This prevents large-scale production of multisensory systems in which data could be collected in a common database and a single classification model could be used for all instruments. To resolve this problem there are a number of methods to eliminate instrumental variation often known as calibration transfer. This approach consists of adjustment of the data from additional instruments (where test samples are measured) to correspond to the master or main instrument (where data is produced for training of prediction model). The set of samples used for standardization is measured on both the main device and on the device to be standardized. Regression algorithms are then applied to establish the relationship between the variables.

In this context the development of a new direct diagnostic method which includes both the development of new sensors and multisensory system with the possibility of operating in a common database, to create a system of diagnostic LC on exhaled breath is a very important and actual task.

The **aim of this thesis** is to develop a methodology for the online analysis of EB using a system of gas-sensitive metal oxide sensors for the diagnosis of LC. In connection with this goal the following **tasks** were solved:

6. Development of a scheme of online analysis of exhaled breath using a system of gas sensitive metal oxide sensors without need of additional sample preparation;

7. Determination of the relative sensitivities of VOCs for preliminary selection of sensors;

8. Conducting a comparative medical study and analysis of EB of patients in the group of LC patients and healthy group;

9. Selection of an efficient algorithm for data processing allowing to effectively separate groups of LC patients and healthy group with high sensitivity and specificity without regard to external factors of the patient's condition (age, gender, smoking, etc.) and based solely on the responses of the multisensory system;

10. Conducting a study of VOC analysis on two sensor systems with identical sensor groups and developing an approach to standardize multisensory systems.

The scientific novelty:

4. A scheme for online analysis of EB using a system of gas-sensitive metal-oxide sensors without need of additional sample preparation has been proposed, designed and approbated. This system combines online measurement and integration ща signal over constant time, high purge rates and as a consequence high performance with minimization of memory effects;

5. An algorithm for processing experimental data has been developed and validated to effectively separate LC patients and healthy subjects with high sensitivity (90.5 ± 2.6)%, specificity (98.1 ± 1.5)%, accuracy (94.0 ± 1.6)%, ROC AUC 0.961 ± 0.018, positive predictive value (98.3 ± 1.3)% and negative predictive value (89.9 ± 2.7)%;

6. A data processing algorithm was developed and validated to estimate the performance of the calibration transfer between two multisensory systems by response standardization on model classification problems.

The practical significance:

4. An online EB analysis system using a cell of 6 gas-sensitive MO sensors has been developed allowing a single patient's EB to be analyzed in 25-30 minutes at 3 temperature modes;

5. Developed an online analysis scheme and data processing algorithm to efficiently separate groups of LC patients and healthy individuals with high sensitivity (90.5 ± 2.6)%, specificity (98.1 ± 1.5)%, accuracy (94.0 ± 1.6)%, ROC AUC 0.961 ± 0.018, positive predictive value (98.3 ± 1.3)% and negative predictive value (89.9 ± 2.7)%;

6. Methodological approaches for sensor systems standardization with identical sensors using the method of graduation dependence transfer were developed which allows using and processing the results of EB analysis from several multisensory systems in a single database.

Statements to be defended:

3. System of online analysis of EB using an array of gas sensitive MO sensors for LC diagnostics;

4. An algorithm for processing experimental data that effectively separates groups of LC patients and healthy subjects with high sensitivity (90.5 ± 2.6)%, specificity (98.1 ± 1.5)%,

accuracy (94.0 ± 1.6)%, ROC AUC 0.961 ± 0.018, positive predictive value (98.3 ± 1.3)% and negative predictive value (89.9 ± 2.7)%.

Work approbation:

The results of the thesis were presented and discussed at the following conferences and competitions: International Student Conference "Science and Progress - 2018". (St. Petersburg, 2018), competition of interdisciplinary student and postgraduate projects "Start-up SPbSU - 2018". (St. Petersburg, 2018), VI St. Petersburg International Cancer Forum "White Nights 2020", National (All-Russian) Conference on Natural Sciences and Humanities with International Participation "Science SPbU - 2020" (St. Petersburg, 2020), International Conference on Natural Sciences and Humanities "Science SPbU - 2020", St. Petersburg International Cancer Forum "White Nights 2021" (St. Petersburg, 2020).

The main results of the thesis were reported in in prominent topical journals indexed in the scientometric databases WoS and Scopus:

1. A.A. Ganeev, A.R. Gubal, G.N. Lukyanov, A.I. Arseniev, A.A. Barchuk, I.E. Jahatspanian, I.S. Gorbunov, A.A. Rassadina, V.M. Nemets, A.O. Nefedov, B.A. Korotetsky, N.D. Solovyev, E. Iakovleva, N.B. Ivanenko, A.S. Kononov, M. Sillanpaa and T. Seeger. Analysis of exhaled air for early-stage diagnosis of lung cancer: opportunities and challenges // Russian Chemical Reviews (2018) 87 (9), pp. 904-921, DOI: 10.1070/RCR4831;
2. A. Kononov, B. Korotetsky, I. Jahatspanian, A. Gubal, A. Vasiliev, A. Arsenjev, A. Nefedov, A. Barchuk, I. Gorbunov, K. Kozyrev, A. Rassadina, E. Iakovleva, M. Sillanpaa, Z. Safaei, N. Ivanenko, N. Stolyarova, V. Chuchina, A.Ganeev. Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer // Journal of breath research (2019) 14 (1), 016004, DOI: 10.1088/1752-7163/ab433d;
3. A. Arseniev, A. Nefedova, A. Ganeev, A. Nefedov, S. Novikov, A. Barchuk, S. Kanaev, I. Jahatspanian, A. Gubal, A. Kononov, S. Tarkov, N. Aristidov. Combined diagnostics of lung cancer using exhaled breath analysis and sputum cytology // Problems in oncology (2020) 66 (4), pp. 381-384, DOI: 10.37469/0507-3758-2020-66-4-381-384.

The work was performed at the Institute of Chemistry of the Federal State Budgetary Educational Institution of Higher Education "St. Petersburg State University" (2017-2021).

Chapter 1. Literature review

1.1. Potential biomarkers of lung cancer in exhaled breath

The analysis of exhaled breath particularly for diagnostic purposes is currently an actively developing area of research. [3]. The possibility of using EB analysis for lung cancer (LC) detection has been studied for many years and now is getting more and more attention of researchers due to the rapid development of metabolomics [4]. Metabolomics analysis of EB is usually focused on quantitative determination of metabolites with low molecular weight (less than 1000 a.m.u.) [5]. Changes in concentrations of such compounds can be caused by various pathophysiological processes, genetic modifications or environmental factors which affect on living systems [5]. Such changes in EB can be preventive signs of such diseases as LC [6].

Volatile organic compounds (VOCs) contained in EB are formed during metabolic reactions occurring both in the human body and in the microbiota. Under pathological conditions in the symbiosis of the microbiota metabolic shifts inevitably occur and as a result there are changes in the produced substances including low-molecular-weight ones. Such compounds can be found in the human EB. In case of pathology the changes in the spectrum of low-molecular metabolites of microflora in theory can be detected with following diagnosis of LC at the early stages.

There are several hundred compounds in human exhalation but only some of them can be useful for detecting LC at an early stage of the disease [2]. A reliable diagnosis requires the identification of certain compounds which presence or concentration definitely correlates with the disease. According to the World Health Organization: a biomarker is any substance, structure, or process that can be measured in the body or its products and also affects or predicts the rate of outcome or disease [7]. Note that biomarkers for healthy and sick people differ usually not in their presence/absence but in their concentration ranges. The mechanisms of generation of potential LC biomarkers in human exhalation are discussed in detail in this paper [8].

It is possible to identify a few compounds whose informativity has been shown in a number of works. Table 1 presents the LC biomarkers for which a significant separation between the LC group and the healthy group (control group) has been shown and which are found in at

least two works [3]. Biomarkers are grouped into classes with indication of their possible nature of origin [9].

Table 1. Informative biomarkers of LC in human exhalation (the number of studies in which a biomarker is marked as informative is given in parentheses)

Class of compounds	**Potential endogenous source**	**Basic compounds and/or derivatives**	**Exogenous source**
Alkanes/ Alkenes/ Alkadienes	Oxidative stress (peroxidation of polyunsaturated fatty acids)	Isoprene (4), decane (3), butane (3), pentane (3), undecane (2), methylcyclopentane (2), 4-methyloctane (2), propane (2), 2-methylpentane (2), heptane (2)	Environment, plastic or fuel
Alcohols	Metabolism of hydrocarbons absorbed through the gastrointestinal tract	Propan-1-ol (5), propan-2-ol (3)	Environment, food, disinfectants
Aldehydes	Metabolism of alcohols; Lipid peroxidation	Hexanal (4), heptanal (3), propanal (3), butanal (2), pentanal (2), octanal (2), nonanal (2)	Environment, food, food waste, cigarette smoke
Ketones	Oxidation of fatty acids; Protein metabolism	Butane-2-on (5), aceton (3), pentane-2-on (2)	Environment, food, food waste, drugs, fragrances, paints
Carboxylic acids	Amino acid metabolism	Acetic acid (2), propionic acid (2)	Food preservatives, solvents, polymers
Aromatic compounds	-	Ethylbenzene (4), styrene (4), benzaldehyde (2), benzene (3), propylbenzene (2), 1,2,4-trimethylbenzene (2), o-xylene (2)	Gasoline, cigarette smoke, fuel, tar, oils

To date a significant number of papers have been published with particularly conflicting results: the average of biomarker's concentration in the EB of patients with LC may be significantly higher in one study and significantly lower in another than the average of biomarker's concentration in a group of healthy individuals [6]. Also note that different groups of researchers used different methods of sampling and sample preparation and biomarker detection. The absence of a standard EB analysis procedure is the main reason for the differences in the results obtained.

In one review of potential LC biomarkers, it was shown that the use of a single substance was not enough to successfully discrimination between the LC group and the healthy person group [3]. On the opposite researchers note that it is the set of substances that forms the EB profile of the patient that is necessary for a diagnostic test [3,6].

1.2. Sampling and sample preparation methods in exhaled air analysis

Sampling is one of the important steps in the analysis of exhaled air. There are a number of parameters that need to be paid attention to in order to avoid incorrect supposition about the nature of any of the identified compounds. These parameters include type of EB (volume of breathing used), breathing technique, multiplicity of sampling, method of sampling, effects of VOCs that are present in the environment, sample storage and transportation conditions. All these parameters are in detail considered and discussed in works [10-12]. When the composition of the EB is analyzed online or in real time the sampling and preconcentration stages can be skipped.

1.2.1. Specific features of exhaled air sampling

Mixed expiratory air or alveolar air only can be sampled to analyze the EB composition. When using the first option there is a high risk of sample contamination by exogenous compounds from the oral cavity and dead space (nasopharynx, trachea, bronchi and bronchioles up to their transition to alveoli) which may compromise the analysis result [10]. Alveolar air is rich in volatile blood compounds so the alveolar sampling method is considered more precise ensuring representativeness and consistency of sample quality [13,14].

Using of various breathing techniques such as breath-holding, hyperventilation, breathing against resistance, etc., is usually aimed either at accumulation of released gases or at separation of EB fractions in one exhalation [13,14].

Sampling can be achieved in one or more complete exhalations. Multiple exhalation composition analysis is more reproducible in terms of sample composition [10] but single exhalation tends to take less time and is more acceptable to patients.

The problem of condensation of water vapor contained in EB and redistribution of EB components between the condensate and the gaseous phase should be mentioned separately. Water vapor which is saturated in EB participates in the transfer of many volatile and non-volatile compounds by dissolving molecules (according to distribution coefficients) inside an aerosol particle [15,16]. All non-volatile compounds such as hydrogen peroxide, adenosine, leukotrienes, isoprostanes, peptides, and cytokines accumulate in water vapor [17]. In addition, polar organic and inorganic compounds such as alcohols, ketones, carboxylic acids, ammonia, and nitrogen oxides can partly concentrate in the EB condensate [18]. To obtain the most complete information on the composition of EB sometimes analyze not only the exhalation but also individually the condensate of the EB. To combat uncontrolled condensation of water vapor in the sampling devices and communications all elements of the system are thermostatted at 37-40ºC.

1.2.2. Methods of exhaled air sample storage

Storage of EB samples can be realized in various ways [10,19]. The most widespread and recommended method of EB sampling is the use of tedlar sampling bags [20,21]. The bags are made of such chemically inert polymeric materials as polyvinyl fluoride, perfluoroalkoxide polymers, polytetrafluoroethylene and polyvinylidene chloride [22]. Such bags have a number of advantages: they are impermeable to gas diffusion (if they are additionally covered with aluminum foil) [23], convenient in use (they can be used more than once, if carefully blown out with purified air, nitrogen, or argon after the previous sample). Despite all the bags have disadvantages: plasticizers and solvents used in polymer production such as phenol and N,N-dimethylacetamide can be released in comparatively high concentrations contaminating the sample [24]. The packages are vulnerable to punctures. Some components such as hexane-1-al and 2-methylbuta-1,3-diene cannot be stored in bags for more than a few hours [25,26].

Another method is using gas-tight syringes. A 50 ml syringe is connected to a mouthpiece

into which the patient exhales. During exhalation approximately 20-30 ml of EB is taken with the syringe which is then transferred to vacuumed glass tubes where the sample is stored until analysis [13]. Another form of EB sample storage is EB condensate [27].

A comparatively recent development is the Bio-VOC breath sampler [28]. This device allows the collection of alveolar air and after the sample collection is completed the VOCs are concentrated using a solid-phase microextraction (SPME) system [29,30]. The main disadvantage is the small volume of collected air (100-150 ml) [29-31].

1.2.3. Preconcentration

The VOC content of EB can vary from a few $\mu mol \times l^{-1}$ to several $fmol \times l^{-1}$ [13,32]. Therefore, depending on the method used to analyze the composition of EB it is necessary to resort to an intermediate step between sampling and analysis to increase the content of the target components.

Concentration on solid sorbents followed by thermal desorption is most often used as a preconcentration method for EB analysis [33,34]. It allows to reach detection limits at a level of ppt at a volume of a sample up to 1 l [35,36]. Despite a large variety of solid sorbents with different retention strengths, operating temperatures and hydrophobicity, one sorbent is not enough to adsorb all the compounds contained in the EB sample due to the wide range of volatility of the VOCs detected. Therefore, multi-component sorption tubes are used where different sorbents are consecutively packed with increasing retention strength [37]. Concentration is conducted under room temperature or lower while practically complete thermal desorption of analytes from the sorbent surface at 250-300°C.

Key sources of error (loss of analyte or appearance of artifacts) in this preconcentration method are degradation of adsorbed analytes during storage [38], thermal decomposition or isomerization of some compounds during thermal desorption [39,40], and degradation of sorbent material [41,42].

1.3. Methods of exhaled breath analysis suitable for the diagnostics of lung cancer

1.3.1. Methods of exhaled breath analysis with quantification of volatile organic compounds

Gas chromatography-mass spectrometry (GC-MS) is probably the most universal and sensitive method for the determination of VOCs in exhalation allowing the analysis of a large number of compounds in the range from ppb to ppt. Therefore, we can say that GC-MS is the gold standard in the determination of low VOC contents in human exhalation [43].

Despite its universality and low detection limits the GC-MS method has a number of disadvantages associated first of all with sampling and sample preparation. The procedure of analysis and processing of the results is usually not very difficult at the current level of automation, availability of selective detectors and a variety of chromatographic columns. But the implementation of GC-MS in clinical settings has a number of limits due to high costs, difficulties in use, and the need for highly skilled analytical chemists to operate the equipment and to interpret the results.

In addition, GC-MS analysis is time-consuming and not an online analysis method. Note that loss and degradation of analytes, in particular reactive or thermally labile metabolites, and possible contamination are important problems that have not been completely solved yet and that have to be overcome to improve the data quality of this type of analysis [10,44,45].

Proton Transfer Reaction Mass Spectrometry (PTR-MS) involves the preformation of the reactant ion H_3O^+ in a low-pressure discharge in water vapor in a hollow cathode and a short drift tube. These ions then enter a drift tube with a constant axial field, at the entry of which the measured sample is introduced. At the end of the tube is a collision cell in which the protonation reaction of the analyte (M) takes place:

$$H_3 0^+ + M \Leftrightarrow MH^+ + H_2O \quad (1)$$

Then the ions are fed into a mass spectrometer usually a quadrupole mass spectrometer. The ratio of the analyte signal intensity to the precursor H_3O^+ signal intensity is used to determine the analyte quantitatively.

The PTR-MS method has high sensitivity: the detection limits in some cases are at the ppt level [46]. The advantages of the PTR-MS method as well as other online systems are especially evident in the determination of unstable compounds in particular aldehydes [46].

The main problems of this method are partial fragmentation of analytes, numerous

interferences, and, as a result, the difficulty of interpreting mass spectra and quantitative determination of a number of compounds. In addition, the humidity of the analyzed air significantly affects the sensitivity of the method and the relative signal intensities of the protonated analyte fragments [47]. Note that for a number of compounds, such as propan-1-ol, it is impossible to use its protonated form MH^+ because it is unstable although for many other compounds the detection of protonated MH^+ components is possible. One of the main disadvantages of the PTR-MS method is the limitation of the range of detected compounds only those for which the proton affinity for MH^+ is greater than that of the H_3O^+ ion.

One of the methods allowing to determine the VOC content in EB is the mass spectrometry of selected ions in flux (SIFT-MS). This method is based on the pre-excretion of one of the reactant ions – H_3O^+, O_2^+ or NO^+ (using a quadrupole mass filter) – from a mixture of components excited in moist air in a radiofrequency discharge followed by chemical ionization of a wide range of compounds in a drift tube and detection of ions by a mass spectrometer. The SIFT-MS method is similar in principle to the PTR-MS method, the difference being that the SIFT-MS method uses a pre-excretion of one of the reactant ions using a mass filter while the PTR-MS ion source forms only one reactant ion - H_3O^+, but the conditions are chosen such that the intensities of other molecular ions are much lower. Such an approach not only simplifies the system but also allows higher sensitivities and lower detection limits in the PTR-MS method than in the SIFT-MS method [47].

At the same time, SIFT-MS is among the few methods that are used to quantify a number of potential LC biomarkers, in particularly acetaldehyde, propan-1-ol, propan-2-ol, acetic acid, methylformate, ethylbenzene, isoprene, etc. [8,48]. Special attention was paid to the determination of acetaldehyde in exhalation and in the gas environment in which growing cancer cells are found [8,49,50].

The detection limits of a number of low-atomic VOCs are at the level of ppb units enough for determination of potential LC markers. At the same time the SIFT-MS method has not obtained results with acceptable levels of specificity and sensitivity of the method yet.

Ion mobility spectrometry is used for the analysis of EB mainly as a variant with a polycapillary column (MCC-IMS) [51-53]. As a result, two-dimensional exhalation "images" are obtained. Using a similar approach the EB of 19 patients with confirmed non-small cell lung carcinoma with different histology, using flexible bronchoscopy with video chips, was investigated in [51]. A total of 72 peaks were detected 5 of which were significantly different

for lung with LC and healthy lung. For adenocarcinoma a peak was identified that corresponded probably to an n-decane dimer, and for squamous cell cancer to butan-2-ol, or 2-methylfuran, or nonanal. The sensitivity, specificity, and predictive value of positive and negative results were 100%, 75%, 80%, and 100% for adenocarcinoma and 78%, 78%, 80%, 75% (butane-2-ol) and 78%, 78%, 80%, 88% (nonal) for squamous cell cancer. Note that the methodology proposed in [51] can hardly be called noninvasive since it is necessary to introduce sondes into a diseased and a healthy lung to determine the difference in intensities of different components present in exhalation.

When using MCC-IMS method to diagnose malignant pleural mesothelioma using EB analysis the values [52] were close to those of the previous work: sensitivity, specificity, predictive values of positive and negative results were 96%, 67%, 76%, 93% respectively. Note that the achieved level of positive predictive value is not enough to use MCC-IMS as the only method for screening examination because additional examination of a significant number of patients who do not have cancer would be required.

The use of IMS without MCC for diagnosis of LC as well as other diseases by exhalation is ineffective [54] which is associated with low selectivity of the method. Note that an important feature of the MCC-IMS system is the possibility of direct analysis without the use of sampling bags and TFE [52].

1.3.2. Methods of exhaled breath analysis based on multisensory systems operating on the pattern recognition principle

Along with the methods of direct detection of VOCs, one of the promising approaches for the realization of the diagnostics of LC by EB in the early stages is the use of multisensory system (MS) like "electronic nose" (EN). This term is understood as a compact and relatively low-cost gas analyzer consisting of an array of non-selective chemical sensors and image recognition system [55]. The principle of EN operation is to form a multidimensional response from an array of sensors with different cross-sensitivities, and then processing the response using chemometric methods to obtain the so-called image of a concrete gas mixture, in our case, exhaled air. Such an image can be called a "breath print" by analogy with a fingerprint. On the basis of the training data set which includes exhalation images of the group of patients with any disease and the group of patients with the confirmed absence of the disease (control group) a mathematical model-classifier is trained which allows making a prediction about the belonging

of the subject by his "breathing print».

The key role in the development of an EN-based diagnostic tool is played by the type of sensors. Thus, to investigate the diagnostic possibilities of detecting LC the following are used: sensors on conducting polymers, piezoelectric quartz resonators, sensors on surface acoustic waves (SAW), optical sensors, semiconductor metal-oxide (MO), etc. Advantages and limits of these types of sensors are closely connected with different character of forming of an analytical signal. Table 2 presents the main advantages and disadvantages of the above types of sensors [2,56].

Table 2. Advantages and disadvantages of sensors for the EN system

Types of sensors	**Principle**	**Advantages**	**Disadvantages**
Piezoelectric quartz resonators / surfactant sensors	Change of resonance frequency	High sensitivity, fast response	Complicated manufacturing process, sensitivity to humidity and temperature, low stability at high temperatures
Optical sensors	Changes in optical density, fluorescence intensity, luminescence	High sensitivity, service life	Difficult to miniaturize, high cost
Semiconductor metal oxide sensors	Change in sensor resistance or conductivity	Low cost, response time, durability, self-cleaning	Low selectivity, relatively high power consumption
Conductive polymers	Change of resistance, mass, optical properties	Low manufacturing cost, low power consumption	Response and relaxation time, low stability, low sensitivity, signal drift
Sensors based on field-effect transistors	Change in electric current	High adsorption capacity	Response time, low VOC sensitivity

1.3.2.1. Conductive polymers

The principle of operation of gas sensors on conductive polymers is to change the sensor resistance due to adsorption of gases by the sensor surface [57]. These sensors operate at ambient temperatures and can be coated with various materials to increase the sensitivity of the sensors to certain VOCs [57]. In the work of McWilliams et al. [57] investigated the possibility of early diagnosis of LC using Cyranose 320 EN system with an array of 32 sensors based on conductive polymers. An EB of 25 patients with LC (stage I and II) and a high-risk group of 166 active and former smokers without LC was analyzed. The results showed a significant effect of the smoking parameter and the gender of the subjects: the discrimination efficiency was higher for ex-smokers than for active smokers, at least in the case of adenocarcinoma (ROC AUC is the area under the curve of the mutual dependence of the probabilities of false positive and true positive results. ROC AUC for former male smokers was 0.846, for former female smokers 0.816, for active male smokers 0.745, for active female smokers 0.725). The authors suggest that changes in the VOC profile caused by active smoking mask to some degree the VOCs associated with tumorigenesis. Moreover, such changes in VOCs due to smoking are more strongly expressed in men than in women. The sensitivity and specificity of the developed method were 88.0% and 81.3% respectively.

1.3.2.2. Sensors on surface acoustic waves

In the work of Chen et al. [58] used a pair of surfactant sensors. The first sensor was coated with a polyisobutylene film and the second was used as a comparison. A preconcentration step was used to EB samples using TFME, followed by injection into a gas chromatographic capillary column. The eluted VOCs were fed to the surfactant sensors and the change in frequency was detected. The data obtained were analyzed using an artificial neural network. As a result, 80% of sensitivity and 80% of specificity were achieved for a total sample of 10 patients.

1.3.2.3. Piezoelectric quartz resonators

Quartz resonators consist of quartz crystals coated with specific metalloporphyrins. VOC is sorbed on the surface of the metalloporphyrins changing the mass of the crystal and the frequency of its vibrations. Such changes are detected and used to train classifiers.

Gasparri et al. [59] used an array of sensors based on quartz microbalances coated with different metalloporphyrins to discriminate 70 subjects with LC and 76 patients from the control group. The sensitivity and specificity achieved were 81% and 91%, respectively. A greater sensitivity to LC in stage I compared with stages II-IV was achieved (92% and 58% respectively).

1.3.2.4. Optical sensors

The principle of the optical sensors is based on the change in the optical characteristics in contact with the VOCs. In a simple variant the analyzed sample is blown through an array of such sensors as a result the color of the sensors changes and after a fixed time the obtained image of the sensors is analyzed. In Mazzone et al. [60] used a system consisting of 24 single-use optical sensors to identify patients with LC and determine the histological type. In this study 229 people were involved: 92 patients with LC (41 - stage I-II non-small cell cancer, 42 – stage III-IV) and 137 – control group with increased risk of the disease. All subjects with LC were grouped according to the histological type of cancer (adenocarcinoma, squamous cell, small cell carcinoma) and samples from each group were separately compared with samples from the control group. The possibility of discriminating groups of patients with early (I-II) and late (III-IV) stages of LC and the possibility of predicting survival was also assessed. It was shown that models built for each cancer type separately are more accurate than a generalized model. The sensitivity and specificity achieved ranged from 70-91% and 73-95% respectively depending on the histologic type. Differences between early and late stages were determined with a sensitivity of 81%, specificity of 93%, and survival (less than 12 months or more than 12 months) was assessed with a sensitivity of 70% and specificity of 86%.

1.3.2.5. Sensors based on field-effect transistors

A set of field-effect transistors based on silicon nanotubes were applied in the work of Shehada et al. [61] to detect and classify LC, gastric cancer, bronchial asthma, and chronic obstructive pulmonary disease (COPD). The total number of subjects was 374. The sample size of subjects with LC was 149, with gastric cancer was 40, with asthma or COPD was 56, and the control group consisted of 129 subjects. At the same time subjects with LC and gastric cancer were further discriminated into two groups according to the stage of the disease: early (I and II) and late (III and IV). As a result, the sensitivity and specificity for the constructed binary classifiers were: 87% and 82% (LC versus control group), 92% and 80% (LC versus asthma),

97% and 90% (LC versus gastric cancer) respectively. At the same time the authors noted that the ability to discriminate the group of patients with asthma from the control group was rather low (sensitivity – 48%, specificity – 91%). The authors associate this fact with the fact that asthma is characterized by only one marker - pentane - rather than by a set of markers as in cancer. In determining the stage of the disease in patients with LC a sensitivity of 34% and specificity of 95% was achieved.

1.3.2.6. Semiconductor metal-oxide sensors

Conductometric gas-sensitive metal oxide (MO) sensors are most commonly used in EN systems because of their low cost, stability, and sensitivity to a wide range of compounds [62]. Nanocrystalline oxides of SnO_2, ZnO, WO_3, etc., doped with Pd, Pt or other catalysts are most commonly used as sensor materials. These oxides are wide-zone semiconductors with n-type conductivity. The MO sensor surface has high adsorption properties and reactivity due to the presence of free electrons in the conduction area of the semiconductor, surface and bulk oxygen vacancies, and active chemisorbed oxygen. Sensors are stable in air when heated up to 500-600 °C and can be obtained in highly dispersed state with crystallite size of 3-50 nm and specific surface area up to 100-150 m^2/g [63].

When the sensor comes in contact with a gaseous environment atoms and molecules of volatile substances adsorb on its surface. In this case both physical adsorption due to weak attraction forces with a binding energy of 0.01–0.1 eV, and chemical adsorption with the appearance of a chemical compound due to exchange type forces with a binding energy of about 1 eV are possible [64].

In practice chemical adsorption is always activated, i.e., the gas particle must spend energy to overcome the potential barrier which is then returned as a result of the act of adsorption. Activated adsorption proceeds at a slower rate which increases with rising temperature [65]. In the great majority of cases gas sensors operate in an air environment where adsorption of oxygen molecules and atoms and water molecules has the main influence on their electrophysical and gas-sensitive properties.

Reducing gases react with chemisorbed oxygen which leads to a decrease in the negative charge density on the surface and an increase in conductivity. A significant change in the conductivity value of the sensor can be registered in the presence of analytes at concentrations of 0.1-10 ppm [63].

Structural changes such as changes in the size and geometry of MO grains result in changes in their conductivity and catalytic properties. Destruction of the MO film after a considerable time in service and phase separation between the metal oxide and the additives are additional factors affecting the stability of the sensor. Exposure to compounds that can bind irreversibly to the metal oxide results in inhibition of catalytic activity and contamination [62,66]. Nitrogen-, phosphorus-, and sulfur-containing compounds can act as such inhibitors [67].

As in practice work with sensors takes place not in vacuum but in air environment it is necessary to take into account that the surface of semiconductor sensor contains a significant amount of chemisorbed oxygen. At different temperature modes you can observe different forms of chemisorbed oxygen: 80-150ºC – oxygen is reduced to molecular anion O^{2-}, 150–260ºC – further reduction to atomic anion O^{-}, 260–460ºC – anion O^{2-}. Therefore, interaction with chemisorbed oxygen is more likely for the reducing molecules than independent adsorption on the surface of the sensing layer [68]. The operating temperature range of such sensors usually is from 200°C to 600°C

In the standard version the EB analysis procedure can in principle be separated into 3 steps. First a reference gas (e.g., room air where the test person is located) is passed through the MO sensor cell which forms a baseline. Next a sample of EB is fed for a certain period of time using the tap. Then the tap is switched back to the reference gas. At all stages the time dependence of conductivity of each sensor is recorded. An example of this dependence is illustrated in Figure 1.

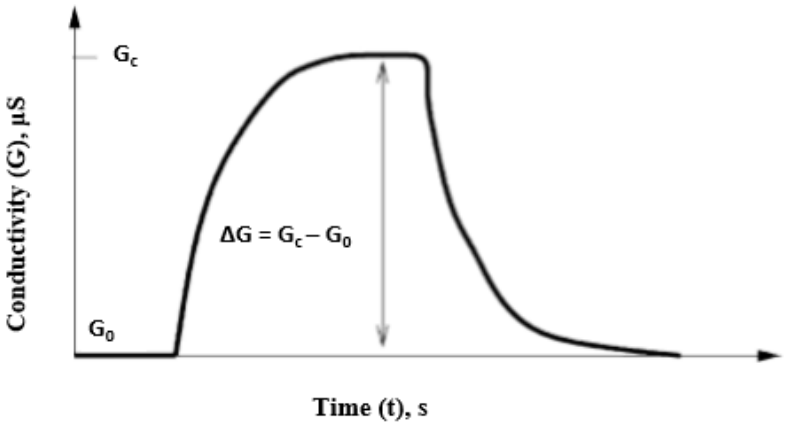

Figure 1. Example of the sensor conductivity (G) over the time (t) when the analyzed gas is supplied

Various features can be extracted from the obtained dependencies of the analyzed samples. The most common is the use of $\Delta G/G_0$. Also, G_c/G_0, G_{max}, integrals of different zones, 1st and 2nd order derivatives, conductivity value at a certain time relative to sample feeding, time of reaching a certain share of conductivity change are used as informative signs.

Table 3 contains information on the works in which the possibility of discriminating patients into LC groups and a control group using MO sensor-based ENs was investigated [56,69].

Table 3. Comparison of informativity criteria of developed tests for LC diagnosis in pilot studies with the use of EN systems based on MO sensors. The main criteria for discriminating the healthy and LC groups are sensitivity (Se), specificity (Sp) and accuracy (Acc).

Characteristics of the sample	**Se**	**Sp**	**Acc**	**Study**
N=101 (43 LC, 58 control group)	95.3%	90.5%	92.6%	[70]
N=89 (16 LC, 73 control group)	-	-	*	[71]
N=18 (9 LC, 9 control group)	100%	88.9%	94.4%	[72]
N=89 (47 LC, 42 control group)	93.6%	83.3%	-	[73]
N=76 (31 LC, 45 control group)	-	-	88%	[74]
N=37 (12 LC, 25 control group)	83%	88%	-	[75]
N=84 (32 LC, 52 control group)	85%	84%	-	[76]
N=290 (144 LC, 146 control group)	94.4%	32.9%	-	[77]
N=145 (52 LC, 93 control group)	83%	84%	-	[78]
N=16 (6 LC, 10 control group)	85.7%	100%	93.8%	[79]

*- sensitivity, specificity, and accuracy are not specified in the paper. The following significance levels were achieved in the discrimination: 0.045, 0.025, 0.001 for each channel of the EN system.

We should mention separately the works where commercially available VOC analysis systems for LC diagnosis have been investigated [74,78]. For example, van de Goor et al. [78] tested five Aeonose EN systems using an artificial neural network to classify patients into a group with LC and a group of healthy people (60 and 107 people, respectively). The results showed a diagnostic accuracy of 83% with a sensitivity of 83%, specificity of 84%, and ROC AUC of 0.84. Comparable results were shown with a sensitivity of 88%, specificity of 86%, and

diagnostic accuracy of 86%. In another study the group of de Vries et al. [74] used SpiroNose in combination with pulmonary function testing equipment to classify patients into LC, COPD, asthma and healthy patient groups (45 LC, 31 controls). Results showed that patients with LC and healthy controls were reasonably well distinguished ($p < 0.001$) and the accuracy on cross-validation was 88% with an ROC-AUC of 0.95 ± 0.11.

For EB analysis sampling procedures have high priority. According to the review presented by Krilaviciute et al. [20] out of 73 studies associated with the diagnosis of lung cancer via EB analysis only six of them realized the mode of direct online measurement while the remaining works used preliminary sampling. In other words, EB in most cases is collected in special containers for storage and transportation to analytical rooms. In addition, most studies use additional sorption procedures to concentrate VOCs [20]. Obviously, such procedures can cause loss of the relevant compounds and sample contamination associated with the sorbent material or storage container. Thus, the offline approach can lead to uncontrollable systematic uncertainty, increasing the analysis time [11]. The time factor becomes especially important for screening surveys because screening tools must be readily available to the general population. Online analysis has the potential to provide more reliable results because of the absence of sample pre-processing procedures and may be a suitable basis for establishing an effective method of screening for LC.

1.4. Methods of multivariate data processing

When using EN systems consisting of non-selective or particularly selective sensors and working on the principle of image recognition for solving classification problems the main part of the work on extraction of information lies on the stage of data processing. In the vast majority of cases the resulting data set has a high dimensionality so for the extraction of useful information multidimensional data processing methods are used.

1.4.1. Data preparation

The obtained analytical signal from an array of sensors can be represented in the form of a matrix **X** of dimension I of rows and J of columns. The rows of such a matrix are called samples they are numbered by the index i, varying from 1 to I. The columns are called variables or

attributes (for example, sensor response) and they are numbered with the index j, varying from 1 to J. Depending on the problem to be solved some dependent variable is known for the measured samples, for example, a mark of belonging to a certain group or its concentration of a component in the sample. And there can be several such variables. This information can be represented as a vector or matrix **Y** of size I rows (number of samples) and N columns (number of dependent variables).

If the sensor response is represented as a single value we have bimodal data as a 2D matrix X but if the sensor response is a set of values (for example, the time dependence of the response or the response scanned by varying one of the parameters) the data set is three-modal and is a 3D matrix (Figure 2) [80].

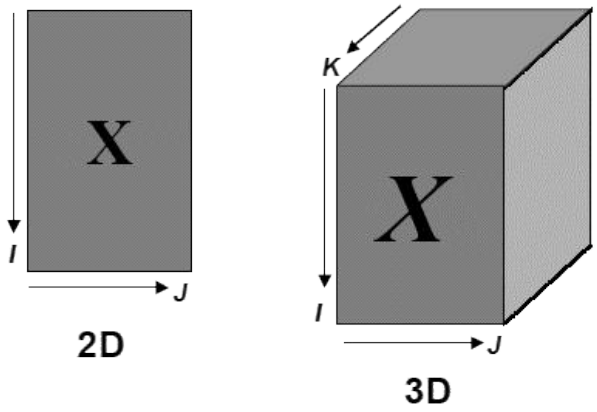

Figure 2. Representation of two- and three-modal data

It is much more convenient to work with bimodal arrays so a sweep procedure is applied to data having 3D matrix structures [80]. Thus, a 3D matrix X of dimension I×J×K is converted into a 2D matrix X of dimension I×JK that can be used for multivariate data analysis.

To avoid difficulties with a large number of features in the data set in advance various feature extraction techniques are applied so that the purpose is to obtain the most informative features using mathematical transformations of the original response matrix (while preserving the information related to the target value) [81]. On the example of the conductivity measurement of MO sensors, it is possible to use a set of responses in time with the following application of the dimensionality reduction method or it is possible to extract the features already at this stage. For MO sensors stationary responses are most often used: R, R/R_0, $(R-R_0)/R_0$. In addition to these, the signal integration or derivatives of 1st or 2nd order can also be used. Also

in some studies as extracted features researchers used: the time at which the signal reaches a certain ratio, the signal at a certain time and others.

Next the matrix X can be subjected to centering and normalization procedures. When the matrix X is centered, the matrix M, whose elements m_{ij} are equal to the mean value of the column mj, is subtracted from it. This operation is necessary for some projection methods such as principal component analysis (PCA).

Normalization in contrast to centering does not change the structure of the data, but simply changes the weight of different parts of the data during processing. When normalizing by columns matrix X is multiplied from the right by a diagonal matrix W of dimension J×J whose diagonal elements w_{jj} are equal to the inverse values of the standard deviation of column x_j. Data normalization is often used to equalize the contributions to the model from different variables [80].

The measurement results on multisensory systems often have a large number of variables (sensor responses and their derived quantities) so visualizing the data in a simple form is complicated if you want to look at the complete picture at once. For these purposes multivariate data analysis using various dimensionality reduction methods such as PCA or linear discriminant analysis (LDA) is used.

Subspace methods have a strong mathematical basis and are popular with many researchers. Despite the fact that PCA and LDA are the most popular methods they also have their disadvantages. PCA is a unsupervised learning method. PCA aims to cover the maximum variance in several dimensions, ignoring discriminatory information. LDA, on the other hand, is a supervised method but assumes unimodal normally distributed classes with different means and equal covariances between classes. In addition, it is well known that LDA is susceptible to overfitting showing too optimistic results when splitting classes on the training set for samples with low sample-to-trait ratios [81].

The principal components method uses new formal variables t_a (a=1,...A) which are a linear combination of the original variables x_j (j=1,...J). Using these new variables the matrix X is decomposed into the product of two matrices T and P:

$$X = TP^T + E = \sum_{a=1}^{A} t_a p_a^t + E \quad (2)$$

The matrix T is called the matrix of scores. Its dimension is (I×A). The matrix P is called the matrix of loadings. Its dimension is (J×A). E is a residual matrix of dimension (I×J). The new variables ta are called principal components. The number of columns t_a in matrix T and p_a

in matrix P is equal to A which is called the number of principal components. This value is obviously less than the number of variables J and the number of samples I. An important property of PCA is the orthogonality (independence) of the principal components [80]. The algorithm NIPALS (nonlinear iterative partial least square) or singular value decomposition is usually used to construct PCA.

1.4.2. Methods used to solve classification tasks

The principle of solving classification problems is based on the construction of models, i.e., a set of rules by which a new sample can be assigned to a certain class. Model construction or training is carried out on the basis of a training set of samples with available a priori information about class membership (for example, class of sick and healthy people). The most commonly used methods [82] in works using EN systems are: **kNN** (k nearest neighbors) method [83], logistic regression (**LR**) [84], support vector machine (**SVM**) method [85], fandom forest (**RF**) method, consisting of an ensemble of decision trees [86].

kNN. The simplest metric method in the classification problem is the k nearest neighbors kNN method. The idea is that the object belongs to the class to which most of its k nearest neighbors belong. The measure of proximity is given by a distance function. The classical kNN uses a Euclidean metric**.** For two points $x_1 = (x_{11}, x_{12}, \dots, x_{1j})$ and $x_2 = (x_{21}, x_{22}, \dots, x_{2j})$ the Euclidean distance is defined as follows:

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{n} (x_{1j} - x_{2j})^2} \quad (3)$$

Also, in an attempt to increase the accuracy of the classification a weighted version of kNN is sometimes used which takes into account not only the number of certain classes that fall into the region but also their distance from the new sample.

RF. RF is a machine learning algorithm [87] that uses a committee (ensemble) of decision trees. To construct a random forest of N decision trees, it is necessary to:

3) generate N random subsamples with repeats X_n, n = 1,…,N.

4) use each resulting subsample X_n as a training sample to construct the corresponding decision tree $b_n(x)$. Moreover:

- The tree is built until there are no more than n_{min} objects in each leaf. Very often trees are built to the end (n_{min} = 1) to get complex and overfitted decision trees with low bias.

- The process of tree building is randomized: at the stage of choosing the optimal feature to split, it is searched not among the whole set of features (J), but among a random subset of size q < J. And the subset of size q is chosen again each time when another vertex needs to be split. The selection of the best of these q features can be done with the help of informativity criterion. Generally, Gini informativity criterion or entropic informativity criterion are used.
- Classification of objects is done by voting: each committee tree assigns the object being classified to one of the classes and the object is assigned to the class for which the largest number of trees voted:

$$a(x) = sign\frac{1}{N}\sum_{n=1}^{N} b_n(x) \quad (4)$$

LR. Logistic regression is a method for constructing a linear classifier that allows us to estimate the a posteriori probability of objects belonging to classes. Provided that class labels take values $Y = \{-1, +1\}$ LR method constructs a linear classification algorithm a: $X \to Y$:

$$a(x, w) = sign\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right) = sign\langle x, w\rangle \quad (5)$$

where w_j – weight of the feature j, w_0 – decision threshold, $w = (w_0, \dots, w_n)$ – weight vector, $\langle x, w\rangle$ – the scalar product of the feature description of objects by the vector of weights. It is assumed that the null feature is artificially introduced: $f_0(x) = -1$. Thus, the task of training a linear classifier is to adjust the vector of weights w using the sample X^m. For this purpose, the LR method solves the problem of minimizing the empirical risk with a loss function of a special form:

$$\mathrm{Q}(w) = \sum_{i=1}^{m} \ln\left(1 + \exp(-y_i\langle x_i, w\rangle)\right) \to \min_{w} \quad (6)$$

SVM. The support vector method is one of the most popular training methods for solving classification problems and is based on the construction of a hyperplane separating sample objects in an optimal way. Let there be a set of objects in X space $\mathbb{R}^n$ with corresponding class labels $Y = \{-1, +1\}$. It is required to build a classification algorithm $a(x) = X \to Y$. Suppose we have a linearly separable set of samples and there is some hyperplane separating the classes -1 and +1. In this case, we will use the linear threshold classifier as a classification algorithm:

$$a(x) = sign(\langle w, x\rangle - b) = sign\left(\sum_{i=1}^{l} w_i x_i - b\right) \quad (7)$$

where $x = (x_1, \dots, x_n)$ – a vector of feature values of the object, $w = (w_1, \dots, w_n)\ \epsilon\ \mathbb{R}^n$ and $b\ \epsilon\ \mathbb{R}^n$ – hyperplane parameters. The SVM method builds the hyperplane that maximizes the

margin between classes for uniqueness. For a linear classifier the margin is defined by the equation:

$$M_i(w,b) = y_i(\langle w, x_i\rangle - b) \tag{8}$$

and characterizes how close the object is to its class. The smaller M_i, the closer the object x_i to the separating hyperplane and the higher the error probability. Accordingly, a negative margin M_i indicates that the algorithm $a(x)$ makes an error on the object x_i.

Then, for convenience, the normalization for the hyperplane equation $\langle cw, x\rangle - cb = 0$ is introduced so that $min\, M_i(w,b) = 1$. This limits the separating band between classes $\{x: -1 < \langle w, x_i\rangle - b < 1\}$ within which no object of the training sample can lie.

For the separating hyperplane to be as far away from the sampling points as possible the width of the band should be maximum. Let x_- and $x_+ -$ two random points of the classes -1 and $+1$ lying on the border of the strip, i.e., their margin is equal to one. Then the width of the separating band can be expressed as the projection of the vector $x_+ - x_-$ on the normal to the hyperplane w.

$$\frac{\langle x_+ - x_-, w\rangle}{\|w\|} = \frac{\langle x_+, w\rangle - \langle x_-, w\rangle - b + b}{\|w\|} = \frac{M_+(w,b) - M_-(w,b)}{\|w\|} = \frac{2}{\|w\|} \tag{9}$$

And for the separating hyperplane to be at the greatest distance from the sampling points the width of the band must be maximum:

$$\frac{2}{\|w\|} \to \max \quad \Rightarrow \|w\| \to \min \tag{10}$$

This leads us to the formulation of the optimization problem in terms of quadratic programming:

$$\begin{cases} \|w\|^2 \to \min\limits_{w,b} \\ M_i(w,b) \geq 1, \;\; i = 1, \dots, l \end{cases} \tag{11}$$

To generalize SVM to the case of linearly inseparable set of samples let the algorithm allow errors on the training objects but so that their number is minimal. For each object subtraction of some positive value ξ_i from margin is applied but it is required that the corrections introduced should be minimal. These changes will lead to the following formulation of the problem called SVM with soft margin:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \to \min\limits_{w,b,\xi} \\ M_i(w,b) \geq 1 - \xi_i, \;\; i = 1, \dots, l \\ \\ \xi_i \geq 0, \;\; i = 1, \dots, l \end{cases} \tag{12}$$

Since we have no information about which of the functionals $\frac{1}{2}\|w\|^2$ and $C\sum_{i=1}^{l}\xi_i$ is more important C factor is introduced which is optimized using cross-validation. As a result, this is a task that always has a single solution.

When the number of classes is more than two in practice, such a problem is usually split into several binary classification problems of One-vs-Rest or One-vs-One type. However, the multiclass support vector method (MSVM, multiclass SVM) proposed by Crammer and Singer [88] makes it possible to reduce the multiclass classification problem to a single optimization problem without the need to split it into several binary classification problems.

1.4.3. Methods for evaluating the results of classification and regression models

To assess the quality of a diagnostic test being tested information about the presence or absence of disease from a reference diagnostic test or so-called "gold standard" is needed. This is a test or combination of tests that can reliably determine whether or not a patient has a disease.

The diagnostics test can give a positive (the patient has the disease) or negative (the patient is healthy) result for the patient under examination. The result of applying a binary diagnostic test to a group of patients taking into account the gold standard test can be presented as a table consisting of 4 groups of outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Such a table is also called a contingency table or confusion matrix (Table 4).

Table 4. Confusion matrix of the results of the diagnostic test

		The result of the gold standard	
		1	0
Prediction result	1	TP	FP
	0	FN	TN

The diagnostic efficiency of a test or **accuracy (Acc)** is defined as the proportion of true results among all test results:

$$Acc = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{13}$$

Sensitivity (Se) is defined as the probability of obtaining a positive outcome for a subject with a disease:

$$Se = \frac{TP}{(TP+FN)} \quad (14)$$

Specificity (Sp) is defined as the probability of obtaining a negative outcome for a subject without disease:

$$Sp = \frac{TN}{(TN+FP)} \quad (15)$$

An assessment of sensitivity and specificity is important when selecting a test for a particular clinical application. The sensitivity of a test reflects the probability of a positive result in the presence of pathology. A high sensitivity of the test allows it to identify patients in the general population. The specificity of the test reflects the probability of a negative result in the absence of pathology so that under high specificity allows you to screen out healthy individuals from the population with suspected pathology. The combination of clinical sensitivity and clinical specificity characterizes the clinical efficacy of the test.

When interpreting laboratory test results, the probability of the actual presence of pathology with a positive result or the reliability of excluding pathology with a negative result is evaluated by determining the predictive value of positive or negative test results.

A positive predictive value (PPV) is defined as the probability of a subject having a disease with a positive outcome:

$$PPV = \frac{TP}{(TP+FP)} \quad (16)$$

Negative predictive value (NPV) is defined as the probability of a subject not having a disease with a positive outcome:

$$NPV = \frac{TN}{(TN+FN)} \quad (17)$$

If we consider not the class label but rather the probability of class 1 as the output value of the classifier, we can obtain a set of contingency matrices with different sensitivity and specificity values by varying the threshold by which the patient belongs to a healthy or sick group. The curve of receiver operating characteristic (ROC-curve) i.e., the curve of mutual dependence of probabilities of true positive results equal to sensitivity and false positive results equal to one minus specificity at all possible values of classification threshold is used for establishing the optimal threshold and for comparative analysis of classification algorithms efficiency. The ROC-curve is a graphical representation of the full spectrum of sensitivity and specificity, since all possible "sensitivity-specificity" pairs for a particular test can be displayed on it (Figure. 3).

Depending on the threshold value and on the distribution of probabilities predicted by the classification algorithm for the patient sample under study the ROC curve has a different shape and position. A desirable ratio between the sensitivity and specificity of the test is achieved by selecting the point of separation. The clearest distinction between sick and healthy subjects is achieved by using tests that have a characteristic results curve shifted toward the upper left corner of the graph.

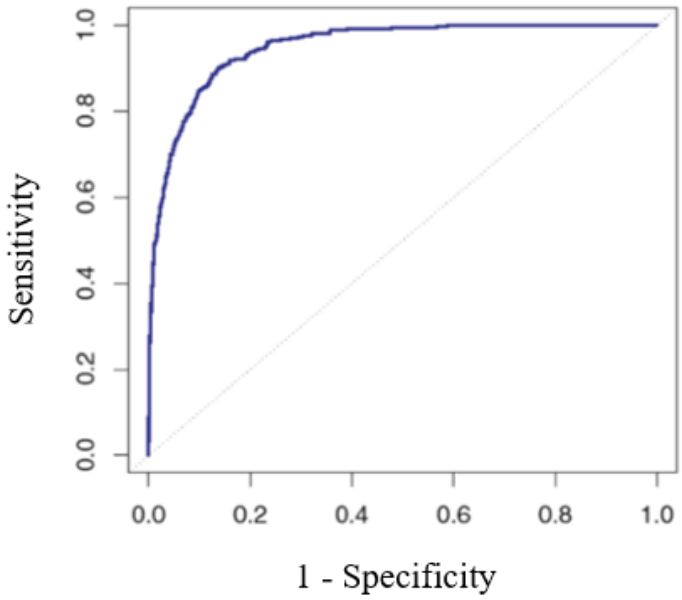

Figure 3. Example of the ROC curve

For an ideal test the curve passes through the upper left corner where the proportion of true positives is 100% or 1 (ideal sensitivity) and the proportion of false positives is 0 (ideal specificity). Therefore, the closer the curve is to the upper left corner, the higher the diagnostic efficiency (accuracy) of the test, and opposite, the smaller the curve bend and the closer it is to the straight line passing through the 45° angle, the less effective the diagnostic test. Points on such a diagonal correspond to the absence of diagnostic efficiency.

A method for evaluating ROC curves is to estimate the area under curves (ROC AUC, area under curve ROC). Theoretically the area varies from 0 to 1.0 but since diagnostically useful tests are characterized by a curve above the positive diagonal (Figure 3) it is commonly considered the variations from 0.5 (no diagnostic test performance) to 1.0 (maximum test performance). This estimate can be obtained directly by calculating the area under the polyhedron bounded on the right and bottom by coordinate axes and on the top left by experimentally obtained points. When ROC curves are visually evaluated their location relative

to each other indicates their comparative efficiency. The curve located above and to the left indicates greater diagnostic efficiency of the corresponding test.

The main criterion used to estimate a regression model is the **root mean square error (RMSE)**:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (18)$$

here $\hat{y}$ – predicted d value of the parameter using regression model, y_i – a priori known value of the parameter, n - number of samples. Note that the units of RMSE and the estimated parameter y correspond to each other.

The estimation of the predictive power of mathematical methods using the mentioned above quality metrics is optimally performed by **validation on the test set**. In this variant the initial data set is divided into training and test sets. The first one is used to adjust the parameters of the model and the second one is used to evaluate the quality of the resulting model. However, in cases where the data set is small the **cross-validation** method is used. The initial data set is separated into k non-overlapping parts of the same size. Then k iterations are performed and at each of them the model is trained on k-1 parts of the initial set (training subset) and the model is tested on one part of the initial set (test subset) which has not yet participated in training. The result is the average value of quality metric over all partitions in the control samples. When k equals the number of samples n we can get a cross-validation variant on individual objects **(leave one out cross-validation)**.

Let us note the importance of the way to evaluate the predictive power of mathematical models whether it is a classification model or a regression model. All of the above-mentioned quality metrics must be evaluated on a test set that has not been used to train the model.

1.4.4. Calibration transfer methods standardization of a pair of sensor systems

In addition to achieving high metrics of clinical informativeness of the test there is another obstacle to implementing a network of such systems which is the limited lifespan of multivariate calibration models. Rebuilding a full-fledged calibration is a costly and time-consuming procedure because of the large number of standard samples. The reasons why a graduation may become unusable may be different: time drift of the signal or change of the sensor characteristics during operation, matrix effects, change of external environmental parameters (temperature, humidity). There are also situations in which it is necessary to replace

the sensor or to transfer the calibration model from one sensor array to another. As a solution to the described problems methods of transferring calibration dependences are used [89]. In this paper the main attention is paid to the calibration transfer between two arrays of MO sensors since the environmental parameters (temperature and humidity) can be controlled in principle by additional tools (temperature control of the gas cell and humidifier installation) and the influence of time drift for MO sensors is at the level of reproducibility of extracted signals.

Calibration transfer methods aim at correcting a new measured data set by eliminating the new data variance related to the different characteristics of the respective sensors. For this purpose, a relationship between two experimental conditions is established and the data measured under the new conditions are corrected according to the established relationship and used for concentration prediction or in classification problems. A small set of standard samples measured on both arrays is called a calibration transfer set and is used to establish a relationship between the two multisensory systems. There are several approaches to implement calibration transfer: through correcting the output signal (concentrations, class labels) predicted on the new sensor array, through correcting the measurements obtained on the new sensor array (signal standardization), or by correcting the predictive model of the new array. [67]. A full review of the calibration transfer methods is done in the following papers [90,91].

Signal standardization is the most common among the methods of calibration transfer. In this case since the final task of this work is classification, the most preferable would be the use of correction of the data set obtained on another array. It should be noted that the given approach will allow to unite the results of measurements of the samples received on several devices in a uniform database and on its basis to build a classification model for diagnostics [67].

Standardization is performed using the ratio between the sensor responses obtained for the standardized samples on the main and on the second sensor array (which is to be standardized) to subsequently correct measurements for unknown samples. Data standardization methods can be separated by the way the relationship between the two sets of sensor signal data is established. Univariate Direct Standardization (UDS) uses the relationship between each channel separately. Piecewise Direct Standardization (PDS) uses a relation between a group of signals and Direct Standardization (DS) uses a relation between all signals. UDS and PDS are linear methods and take into account the linear shift of the sensor signal. PDS was proposed as an improvement of the UDS version with the proposition that spectral signals for adjacent wavelengths have a high correlation. However, sensor responses can be independent or partially

correlated depending on the fill and location of the sensors in the array and the analytes which raises doubts about the effectiveness of PDS for MO sensors. Given this fact, for MO sensors it is most common to use a direct standardization method that takes into account the responses of all sensors. The relationship between the two data sets can be established using various multivariate approaches: multiple linear regression, Projection to Latent Structures Regression-2 (PLS2), etc. [67].

In [92] a robust regression to the presence of outliers was applied to transfer the calibration relationship between two identical arrays consisting of 6 MO sensors each. An artificial neural network with back propagation of error on the resulting dataset was used as a calibration model when measuring gas samples collected from pulp and paper mills. A standardization dataset consisting of 27 gas mixtures of hydrogen sulfide, dimethyl sulfide, dimethyl disulfide, and methylmercaptan was used for transfer.

In [93] 5 identical arrays with 8 MO sensors were used to calibration transfer from the main array to the others. First, the signals for 4 compounds (ethanol, ethylene, carbon monoxide (II), and methane) were measured at 10 concentration points. Twenty samples each were used to construct the graduation and verification. The standardization set consisted of 8 samples: 2 concentration points each for 4 compounds. DS and PDS methods based on PLS2, orthogonal signal correction and weighted least squares method were used to transfer the calibration dependences.

Researchers note that the efficiency of data standardization depends largely on the data themselves and on the set of standard samples [67,94]. When standardizing data, it is desirable to use a small number of standard samples for transfer due to the labor-intensive procedure of their preparation and measurement. On the other hand, the number of standard samples should be sufficient to describe the variance between the two data sets for effective transfer of the calibration. In some cases, sample selection is based on existing knowledge of the analyte and the task at hand. In addition to manual selection, the Kennard-Stone algorithm [95] is also used. This algorithm is mainly used for uniformly distributed samples in the feature space and consists in sequentially selecting the sample that is most distant from the previously selected ones. As the initial state, 2 samples most distant from each other are selected. The Euclidean distance is mainly used as a measure. Also, for research purposes to evaluate the performance of the response standardization methods on average multiple random sampling for standardization can be used.

In current study the following methods were used to evaluate the transfer efficiency of grading: the univariate direct standardization (UDS), the univariate direct standardization without using the intercept of regression (UDSwoi), the direct standardization method based on regression on latent structures (DS-PLS2) and the direct standardization method using L1 regularization (DS-L1R).

Chapter 2. Experimental details

2.1. Description of sensor characteristics

Three multisensory systems (MS 1, MS 2.1, MS 2.2) were used in this work. The MS 1 system was used for the task of optimizing the sensor set, the procedure of analyzing the EB on model gas mixtures and further in the medical research. The MS 2.1 and MS.2.2 systems were used to investigate the possibility of combining responses, i.e., combining the database and using a single instrument classification model.

According to a review of scientific papers in which medical research on patients was conducted it was found that an average of no more than 10 sensors are used to develop models of disease diagnosis. Presumably, increasing the number of sensors in the system at least does not increase, and even decreases the informativity due to the high collinearity in the sensor responses. Therefore, in order to maximize the informativity of the multisensory system it is necessary to use sensors with as large a difference in cross-sensitivity as possible. Each multisensory system consisted of 6 MO gas sensitive semiconductor sensors made by sol-gel method in the laboratory of applied chemical physics of Kurchatov Institute. Each sensor is a multilayer system consisting of a semiconductor gas sensitive layer (sensor layer), an Al_2O_3 dielectric base, and a heater layer (heater). The sensor layer and the heater are applied on opposite sides of the substrate by screen printing. The semiconductor gas sensitive layer contained SnO_2 nanoparticles with various Pt, Pd, or La dopants [96]

Table 5. Composition of sensors for MS 1, MS 2.1 and MS 2.2

Sensor label	**Composition of the sensitive layer**
S1, S2	SnO_2 with dopant Pt (3%) and Pd (1%), Cd
S3, S4	SnO_2 with dopant Pt (3%) and Pd (1%)
S5, S6	SnO_2 with dopant Pt (3%) and Pd (1%), La

* - S1 and S2, S3 and S4, S5 and S6 have the same quality composition but have different heater resistances so that at the same applied voltage the heaters have different temperatures

Depending on the heater temperature and, consequently, the temperature of the sensor layer, the relative sensitivity of the sensors to different groups of compounds is different [97]. Thus, the informativity of the data on the analyzed samples can be improved by using several temperature regimes. It is known [98] that the sensor response to a certain gas has a maximum at a certain temperature. For example, a SnO_2-based sensor with Pd dopant has maximum sensitivity to H_2 at about 200°C, to propane at 350°C, to methane at 450°C, etc. These maxima are not clearly defined and sensitivity to a particular gas is observed over a relatively large temperature range. In the experiments carried out in this work, three temperature regimes chosen empirically and distributed over the operating temperature range were used. Also, the selected temperatures are located in the characteristic areas of the maximum sensitivity to substances that are easily oxidized (e.g., alcohols, ketones), relatively easily oxidized (long-chain alkanes), and to gases that oxidize with relative difficulty (propane). The sensor surface is not contaminated by the decomposition products of some compounds because the interaction of these products with oxygen on the catalytic surface at high temperature leads to complete oxidation of the adsorbed compound and decomposition products. In fact, the decomposition of the target compound on the catalytic surface is part of the sensor response (i.e., the process of catalytic oxidation by chemisorbed oxygen). For example, long-chain hydrocarbons, which are oxidized on the surface with the carbon chain broken, give the sensor sensitivity at a much lower temperature (below 300°C) than methane (around 450°C).

To achieve the required sensitivity to the analytes under study (biomarkers of pulmonary oncopathology), 3 temperature regimes of T1, T2, and T3 sensors were selected. To heat the sensor layer the microchip included a microheater layer formed from a Pt-containing suspension. The heater was simultaneously used as a thermistor due to the strictly linear dependence of its resistance RH on temperature. Temperature coefficient of heaters $\alpha = 0.0027$ ºC^{-1} was a constant over the entire operating temperature range from 150 to 600 ºC. To heat each sensor i, a constant voltage Ut_0 through a limiting resistor R_{0i}.was applied to its heater. By measuring the voltage across the heater UH_i, its resistance RH_i can be calculated using the equation:

$$RH_i = \frac{R_{0i} \times UH_i}{Ut_0 - UH_i} \tag{19}$$

If the heater resistance of the sensor i at 20 ºC is denoted by RH_i^{20} then the temperature coefficient of the heater $\alpha = (RH_i - RH_i^{20}) / (RH_i^{20} \times (T - 20)$, where RH_i is the heater resistance at temperature T. Then the heater temperature is calculated by the following formula:

$$T_i\ (°C) = \frac{RH_i - RH_i^{20}}{RH_i^{20} \times \alpha} + 20 \quad (20)$$

Tables 6, 7 and 8 show the calculated values of heater temperatures for 6 sensors at temperature modes T1, T2 and T3 corresponding to three different heating voltages Ut_0 4.48, 4.98 and 5.48 V for MS 1 and 2.67, 3.01 and 3.35 V for MS 2.1 and MS 2.2 respectively. Resistance and voltage were measured using a DT-832 model multimeter (Dadits, China).

Table 6. Sensor heater temperatures for three temperature modes T1, T2 and T3 for MS 1 (maximum relative error of heater temperature determination was 13.8%, 13.7% and 12.6% for T1, T2 and T3 respectively)

Sensor	R_{0i}, Ohm	RH_i^{20}, Ohm	Ut_0, mV		
			4480	4980	5480
			T, ^{0}C		
S1	33.0	7.9	360	409	464
S2	32.6	8.3	325	377	428
S3	32.6	10.2	360	411	462
S4	32.9	14.1	473	534	602
S5	32.8	11.4	444	502	560
S6	32.8	12.6	392	448	502
			T1	**T2**	**T3**

Table 7. Sensor heater temperatures for three temperature modes T1, T2 and T3 for MS 1 (maximum relative error of heater temperature determination was 14.3%, 12.1% and 11.4% for T1, T2 and T3 respectively)

Sensor	R_{0i}, Ohm	RH_i^{20}, Ohm	Ut_0, mV		
			2670	3010	3350
			T, ^{0}C		
S1	10.0	10.4	334	385	436
S2	10.0	10.6	335	387	438
S3	10.0	11.5	362	417	473
S4	10.0	10.8	337	390	442
S5	10.0	10.5	344	399	452
S6	10.0	11.2	345	401	455
			T1	**T2**	**T3**

Table 8. Sensor heater temperatures for three temperature modes T1, T2 and T3 for MS 1 (maximum relative error of heater temperature determination was 14.2%, 12.0% and 11.3% for T1, T2 and T3 respectively)

Sensor	R_{0i}, Ohm	RH_i^{20}, Ohm	Ut_0, mV		
			2670	3010	3350
			T, ^{0}C		
S1	10.0	11.1	349	402	455
S2	10.0	10.6	337	390	442
S3	10.0	11.5	343	397	450
S4	10.0	11.1	352	408	463
S5	10.0	11	350	407	461
S6	10.0	10.7	347	403	457
			T1	**T2**	**T3**

As can be seen from Tables 6, 7 and 8, the sensors of the same composition (S1 and S2, S3 and S4, S5 and S6) due to the differences in the heaters had different temperatures at the same order of temperature conditions so that sensors had different gas sensitive properties. Thus, when measuring the response readings of 6 sensors S1, ..., S6 at three temperature modes T1, T2, T3, 18 informative signs were obtained: S1_T1, S2_T1, ..., S5_T3, S6_T3.

Figure 4 shows the electrical circuit by which the conductivity G (S) of each sensor was measured $G_i = 1/RS_i$, where RS_i is the resistance of the sensor i, Ohm. Stabilized voltage US_0 = 5V (4950 mV) was applied to each sensor and the output voltage of current-voltage converter US_i proportional to the sensor current was measured: $US_i = US_0 \times RB_i/RS_i$, where RB_i is resistance of the bias resistor. The conductivity was calculated using the following formula:

$$G_i = \frac{US_i}{US_0 \cdot RB_i} \quad (21)$$

Figure 5 shows an example of the program interface showing the dependence of the sensor output voltages on time when a sample is sequentially passed through the sensor cell.

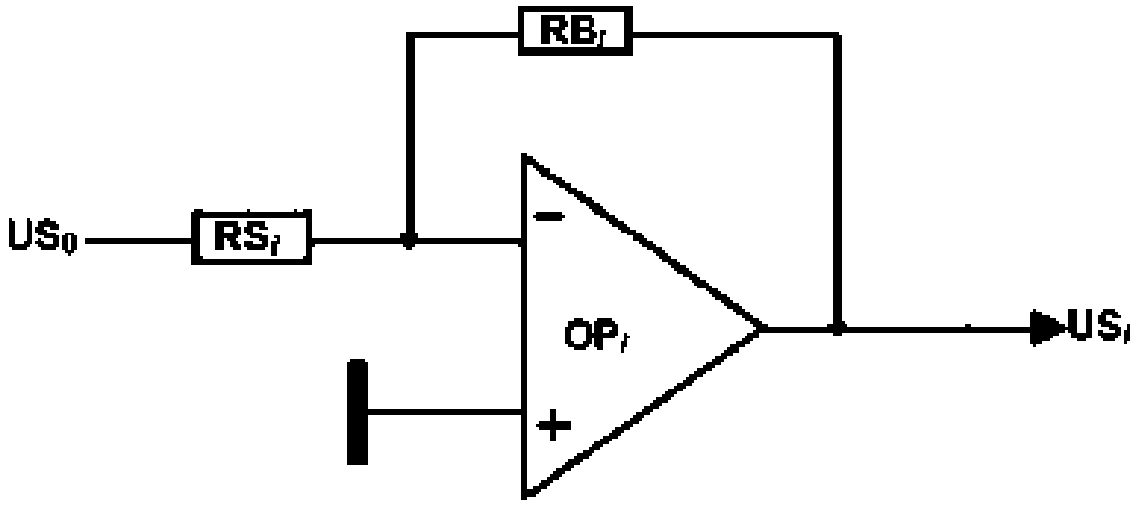

Figure 4. Electrical circuit for conductivity measurements for each sensor

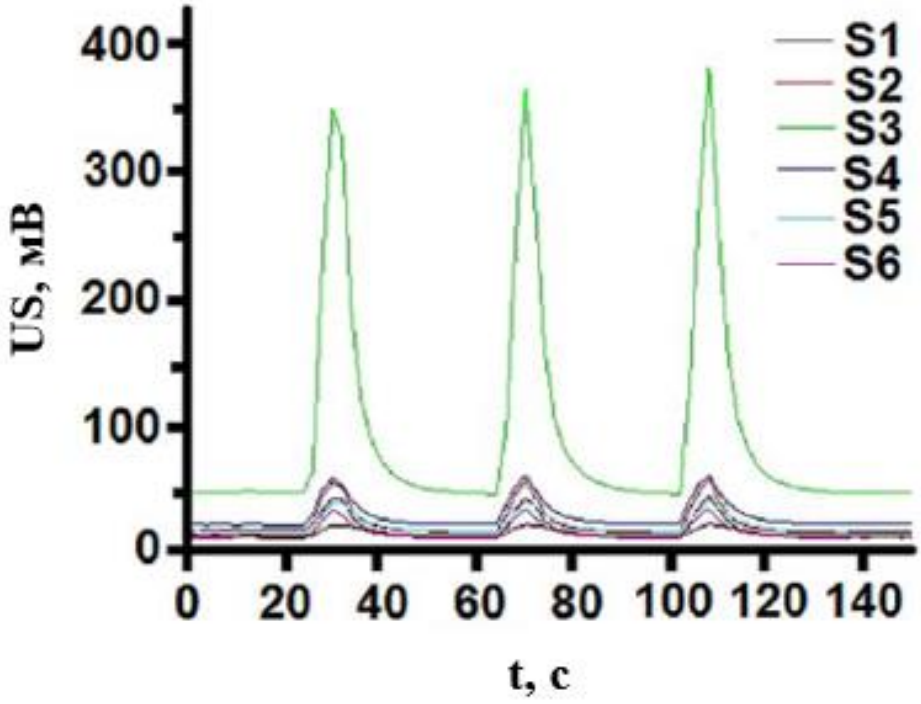

Figure 5. Output voltage dependence US_i for 6 sensors with 3 samples in series

Voltages US_i and UH_i for all 6 sensors were converted using a multichannel 24-bit sigma-delta ADC and recorded on a computer with a frequency of 0.25 Hz for MS 1 and 1.0 Hz for MS 2.1 and MS 2.2 in the form of tables and graphs. The time for establishing stable readings when switching temperature mode was from 3 to 7 minutes.

2.2. Technique for the preparation of model gas mixtures

In this work one-component graded air gas mixtures: air – n-heptane, air – propan-1-ol, air – ethylbenzene, air – o-xylene and three-component mixtures (air – n-heptane, propan-1-ol, o-xylene) that were prepared in accordance with the approaches described in GOST R ISO 6144-2008 were used for experiments. Gas mixtures were prepared using a syringe by injecting a known volume of liquid target components into a special bag with a known volume filled with background gas (room air).

Bags

To prepare the gas mixture, 1, 3 and 5 l Restek® (USA) Tedlar bags were used as a mixing chamber and simultaneously as a storage vessel. The pouches are equipped with a special membrane allowing the liquid of the target component to be injected inside. The pouch valve is made of inert polypropylene and has both a convenient inlet for connection to the sampler tube and the ability to inject the sample with a syringe through the membrane.

Syringe

For sampling liquid target components, we used Hamilton (USA) calibrated measuring syringes, 1 and 10 μl, with gas-tight gaskets to ensure tightness to avoid significant leakage of gas or liquid.

Bag preparation

Before preparing the gas mixture for each bag a bag cleaning procedure was performed by triple-filling and releasing with room air. Additional tests showed that the analytical signal of the room air sample after this cleaning procedure is at the level of the noise signal of the baseline sensors. After the cleaning procedure the bag is vacuumed with a pump.

Filling bags with ambient air

The bag is filled with background gas (room air) using a 12 V pump (Alldoo Micropump Co., China) with a known and constant pumping rate to the required volume of the bag. The rate of pumping of room air by the pump is controlled by voltage from the DC power supply B5-47 (Izmeritel Plant, Armenia) and is registered by a VA-10414 rotameter (Dwyer Instruments Inc., USA).

Component injection

The required volume of the target component is calculated based on the desired composition of the final gas mixture and the volume of the bag. After the liquid volume in the

syringe reaches the set value, the syringe needle is immediately injected into the package through the membrane by slowly pushing on the syringe piston while simultaneously removing the needle from the membrane. After the substance was injected, the gas mixture was incubated for a predetermined time to homogenize and achieve temperature equilibrium between the mixture and the environment. Additional experiments showed that for all investigated VOCs, 20 minutes was sufficient to achieve equilibrium and obtain a reproducible sensor signal.

VOCs under analysis

The following substances were used in this study: n-heptane (≥ 99%, for HPLC), ethylbenzene (≥ 99.5%, analytical standard), propan-1-ol (≥ 99.9%, for HPLC), o-xylene (≥ 99.0%, pure for analysis) by Sigma Aldrich (Merck KGaA, Darmstadt, Germany).

The concentration of analyte A in the final gas mixture was calculated according to the following formula (the volume of the introduced substance on the order of a few µl compared to the used volumes of the package on the order of 1-5 L can be neglected):

$$c_A\ (ppm) = \frac{V_A * \rho_A * R * T}{M_A * V_b * p} * 1000 \quad (22)$$

where

V_A – is the volume of substance A as a liquid, µl;

ρ_A – density of substance A, g/ml;

M_A – is the molar mass of substance A, g/mol;

V_b – volume of the bag before the substance is injected, l;

R – constant, equal to 8.314 J/(mol*K)

T – temperature, K;

p – pressure, kPa.

2.3. Analysis of model gas mixtures and exhaled breath samples in medical research using MS1

2.3.1. Scheme of the experimental setup for the analysis of model gas mixtures and exhaled breath samples

In contrast to fairly common approaches with pre-concentration of VOCs or any additional storage procedures this work used direct online analysis for the analysis of exhaled air samples. The use of additional intermediate steps in the sample preparation process can

adversely affect results: for example, loss of analyte or appearance of artifacts, degradation of adsorbed analytes during storage [38], thermal decomposition or isomerization of some compounds during thermal desorption [39,40], and degradation of sorbent material [41,42]. In order to eliminate the "memory effect" associated with the sorption of VOCs on the walls of the transport tubes, a pump was installed in the experimental scheme of online analysis of EB (Figure 6) to continuously purge the circuit and cell lines (3.5 l×min^{-1}). This provided a stable signal and consequently a stable baseline without drift. At the same time, it was found that a further increase in the speed leads to a significant decrease in the stability of the sensor signals.

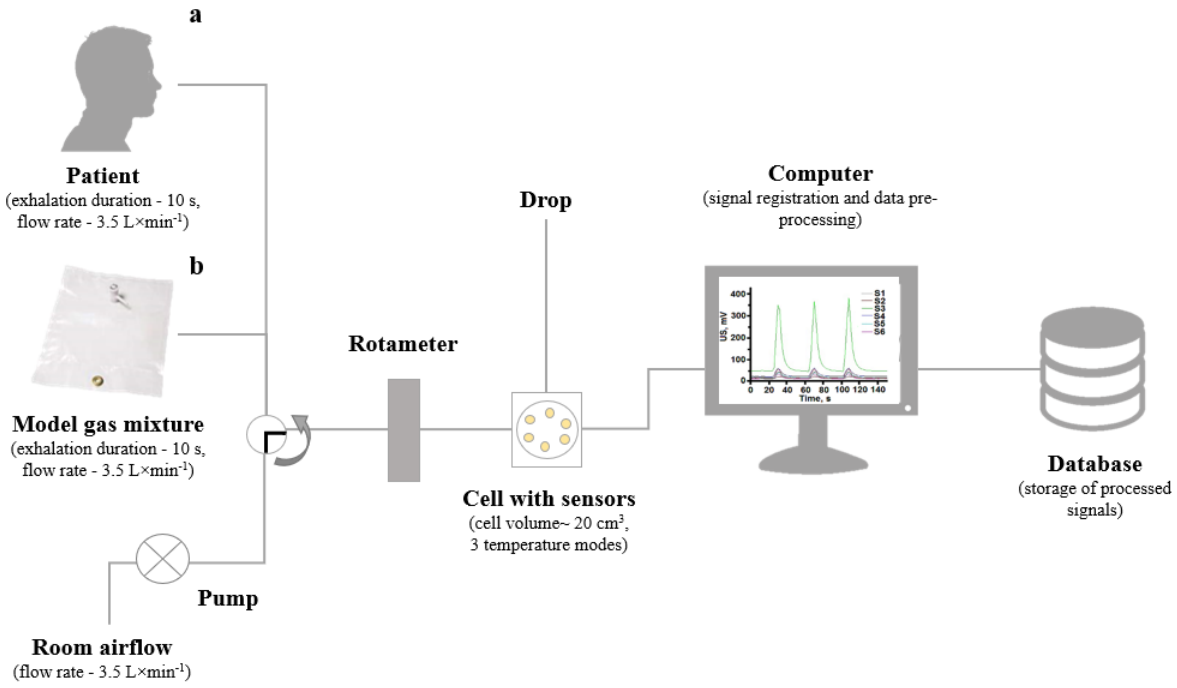

Figure 6. Experimental scheme of analysis for MS 1 (a - scheme of online analysis of EB in medical research, b - scheme of online analysis of model gas mixtures for assessment of cross sensitivities)

To account for the presence of exogenous VOCs in the ambient room air, the sensor cell was purged with room air; thus, there was no need to "purge" the subject's lungs with clean air for 3-5 minutes to reduce the influence of exogenous VOCs on the analysis result. Note that pre-cleaning the patient's breath with medical air, which has been used in similar studies, can lead to both a significant change in the patient's VOC profile and a loss of VOCs. The approach proposed in this paper simplifies the analysis by reducing the influence of exogenous VOCs.

Each subject waited for 10 min before the first EB measurement before starting the analysis. Moreover, an additional study using a charcoal filter showed no change in sensor responses. It is worth explaining that the primary criterion for selecting rooms suitable for EB analysis was the evaluation of the change in sensor conductivity between the basic experimental analysis scheme and the same scheme with the carbon filter connected at the input. The criterion was defined as follows: if the maximum relative change of the sensor is more than 5%, the room was defined as unsuitable for the analysis of EB of patients. Figure 7 shows an example of such a test for one room.

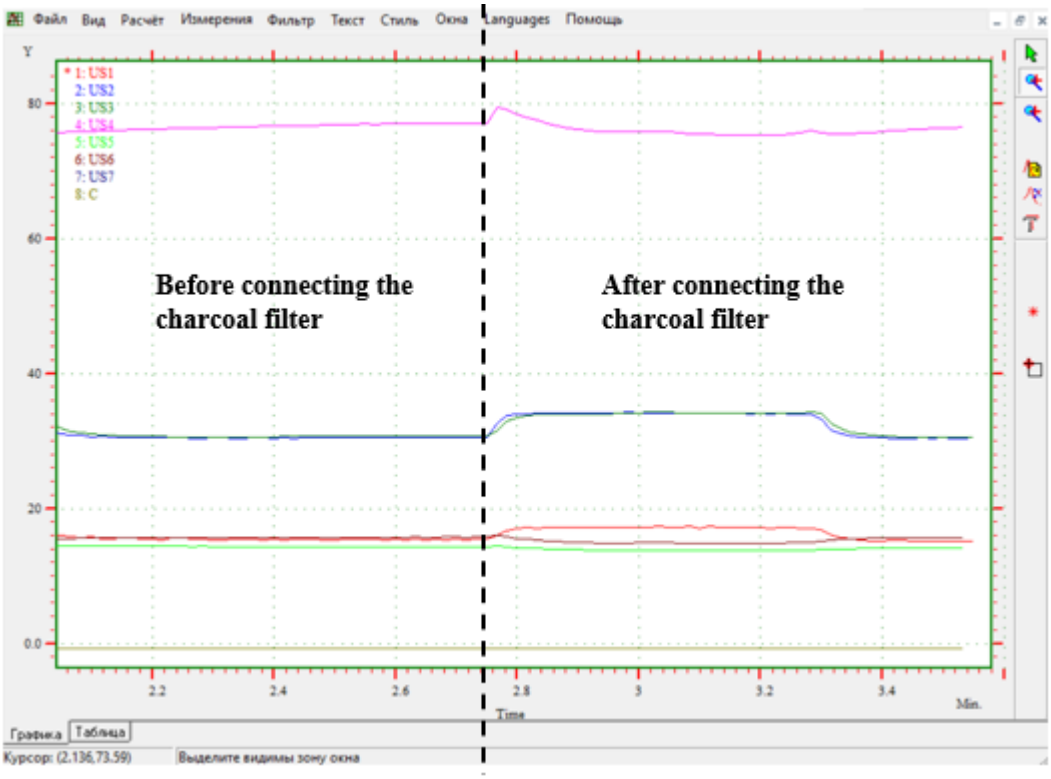

Figure 7. Change of conductivities of sensors when connected to the input of the experimental circuit of the carbon filter analysis. The figure shows an example of the application interface for displaying the conductivities of the sensors in relative units (ordinate axis) from time in minutes (abscissa axis)

The procedure for analyzing both patient EB and model gas mixtures can be roughly divided into three consecutive steps. In the first stage, the pump valve is open and ambient air is pumped through the sensor cell at a rate of 3.5 l×min^{-1}, while the valve for the sample to be analyzed is closed (as shown in Figure 6). This velocity was chosen as optimal, allowing a sufficient volume of patient EB (~600 ml) to flow through the sensor cell, for which an

acceptable exhalation duration was 10 seconds, and obtain an informative online analysis response. In the second step, the pump valve is closed and the sample delivery valve is opened. The sample is then fed for 10 seconds through a sensor cell at 3.5 L×min^{-1}, or in the case of a medical study, the patient exhales for 10 seconds through a sterile disposable mouthpiece through a cell at 3.5 L×min^{-1}, which corresponds to a volume of about 600 ml of EB. In the third step, the sample valve is closed, the pump valve is opened, and ambient air is again pumped through the sensor cell at the same rate of 3.5 L×min^{-1}. The flow rate is monitored using a rotameter. Exhalation velocity was also monitored by the patient with the rotameter readings monitored. In this study, a 12 V DC microvacuum pump for medical purposes (Alldoo Micropump Co., Ltd., Yueqing Zhejiang, China) and a VA-0414 rotameter (Dwyer Instruments Inc., Indiana, USA) were used. Similar actions were performed on all temperature modes after the sensors reached a stable signal.

The conductivity integral over time minus the area formed by the baseline was extracted from the conductivity curve as an analytical signal for each sensor at each temperature mode (Figure 8).

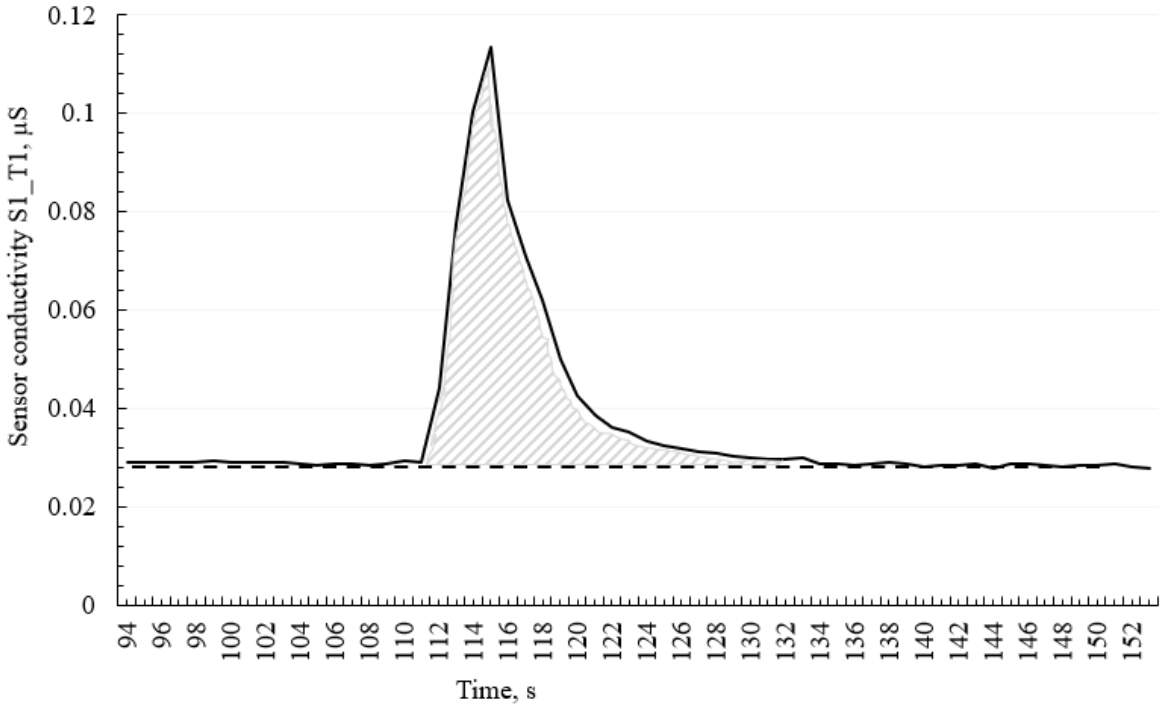

Figure 8. Principle of analytical signal extraction (the value of the area under the conductivity curve of the sensor minus the baseline acts as the analytical signal)

2.3.2. Determination of the relative sensitivities of volatile organic compounds for the sensors used

Single-component model gas mixtures of the following compounds: n-heptane, ethylbenzene, and propan-1-ol were used to optimize measurement parameters and assess the cross-sensitivity properties of the multisensory system. All these substances are potential biomarkers of LC, which were often used in similar studies [17,18]. All reagents were used to evaluate the sensitivity and linearity of the calibration dependences.

To assess the character of sensitivity, the calibration dependences for propan-1-ol, ethylbenzene and n-heptane were constructed. The range of investigated concentrations for calibration dependences was 0.5-500 ppm. Six parallel measurements (n=6) were used for each concentration level. In the chosen concentration range the linear approximation for building the calibration works well. Parameters of calibration dependences are presented in the table 9. Initial data for building the calibration dependences are located in the repository [99]. It should be noted that in the range of specified concentrations the dependences for heptane and propan-1-ol were linear, while for ethylbenzene nonlinearity of the analytical signal appears at very high concentrations (starting from 250 ppm).

Relative sensitivity factors (RSF) were calculated relative to propan-1-ol using the equation:

$$RSF_{A(i)} = \frac{k_{A(i)}}{k_{Prop(i)}} \quad (23)$$

where $k_{A(i)}$– is the slope coefficient of the calibration dependence for the sensor i for compound A, and $k_{Prop(i)}$- slope coefficient of the calibration dependence for the sensor i for propan-1-ol. Coefficients were calculated for three temperature regimes (Table 10).

Table 9. Parameters of calibration dependences constructed for n-heptane, propan-1-ol and ethylbenzene for the first temperature regime

	n-heptane				
Sensor	slope	intercept	R^2	RMSEC	RMSEP
S1	99.3	-4.93	0.9955	12.5	13.3
S2	24.07	-3.13	0.9970	10.1	10.6
S3	0.06	-12.1	0.9961	11.6	12.5
S4	0.05	-9.95	0.9912	17.4	17.1
S5	0.07	-15.58	0.9938	14.6	15.5
S6	0.02	-18.79	0.9913	17.3	18.1
	propan-1-ol				
Sensor	slope	intercept	R^2	RMSEC	RMSEP
S1	82.4	0.94	0.9981	7.2	7.3
S2	43.42	-2.05	0.9951	11.4	12.5
S3	0.62	-7.82	0.9951	11.4	11.4
S4	2.24	1.87	0.9932	13.6	14.3
S5	5.23	-4.24	0.9960	10.4	11.2
S6	1.94	-0.87	0.9933	13.4	14.3
	ethylbenzene				
Sensor	slope	intercept	R^2	RMSEC	RMSEP
S1	29.21	-13.53	0.9787	19.2	19.8
S2	10.88	-12.35	0.9683	23.4	23.4
S3	0.15	-17.09	0.9805	18.3	18.6
S4	0.56	-19.51	0.9692	23	23
S5	0.56	-10.78	0.9936	10.5	10.9
S6	0.17	-22.42	0.9519	28.8	29.2

Table 10. Representation of relative sensitivity coefficients regarding propan-1-ol for three temperature regimes (maximum relative error in RSF determination was 11.6%)

Compound	**S1_T1**	**S2_T1**	**S3_T1**	**S4_T1**	**S5_T1**	**S6_T1**
n-heptane	1.205	0.554	0.097	0.022	0.013	0.01
ethylbenzene	0.354	0.251	0.242	0.25	0.107	0.088
propan-1-ol	1.0	1.0	1.0	1.0	1.0	1.0

Compound	**S1_T2**	**S2_T2**	**S3_T2**	**S4_T2**	**S5_T2**	**S6_T2**
n-heptane	1.406	0.604	0.054	0.008	0.007	0.004
ethylbenzene	0.148	0.059	0.08	0.091	0.108	0.066
propan-1-ol	1.0	1.0	1.0	1.0	1.0	1.0

Compound	**S1_T3**	**S2_T3**	**S3_T3**	**S4_T3**	**S5_T3**	**S6_T3**
n-heptane	0.904	0.478	0.027	0.028	0.005	0.004
ethylbenzene	0.103	0.078	0.046	0.072	0.04	0.036
propan-1-ol	1.0	1.0	1.0	1.0	1.0	1.0

As can be seen from Table 8, the relative sensor sensitivities for the studied substances can differ by several orders of magnitude which indicates the good cross-sensitivity of the sensors chosen for the study.

2.3.3. Data processing and classifier training

Python 3.6 software (Python Software Foundation, USA) and libraries pandas, scipy, matplotlib, numpy, and scikit-learn were used for data analysis, visualization, and processing, estimation of distribution using statistical criteria, application of PCA, and training of mathematical model classifiers (kNN, RF, SVM, LR).

Figure 9 shows the scheme of processing the array of responses from MS 1 as part of the conducted medical research.

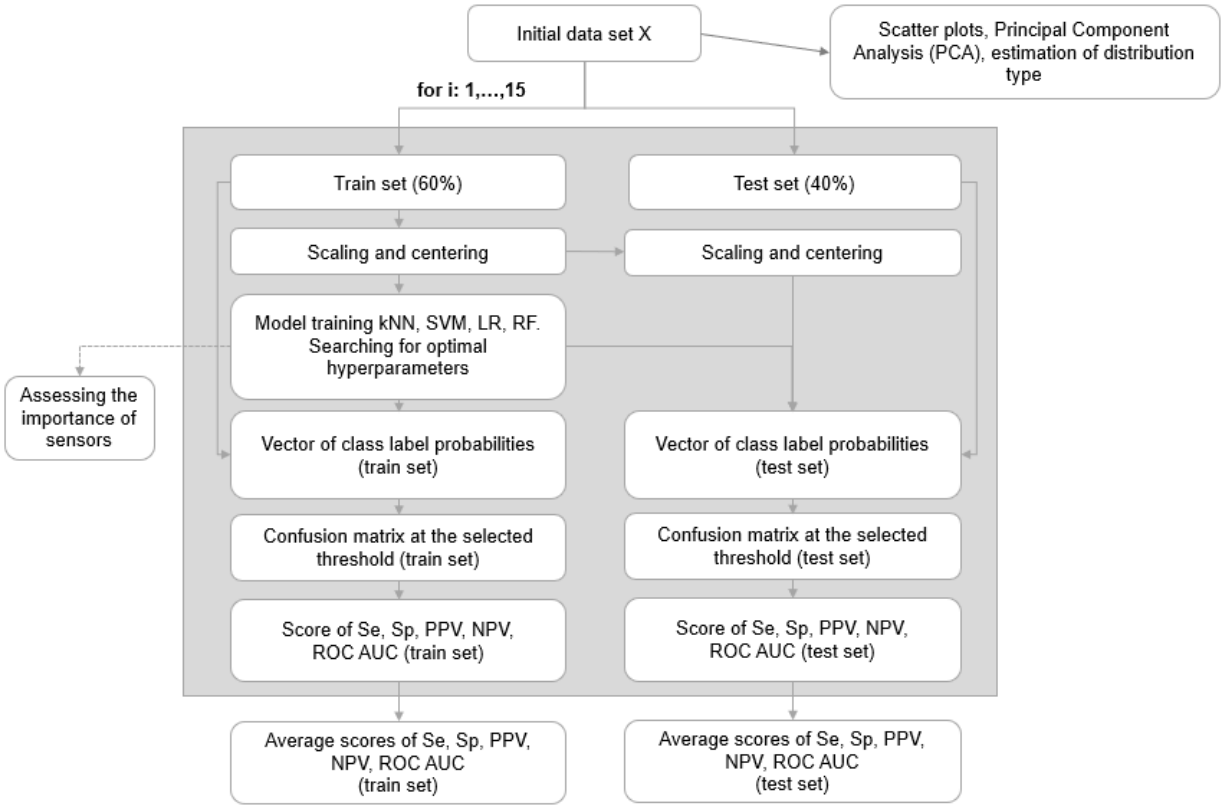

Figure 9. Scheme of data processing using MS 1 obtained in the course of medical research and presentation of the final results of the developed diagnostic test

A detailed processing script and input data: an array of responses with corresponding patient attributes (group, age, gender, etc.) are contained in the repository [99]. At each iteration for each algorithm an internal cross-validation is performed on the training set with 3 blocks. This check is performed for all possible combinations of hyperparameters of the algorithm initially set. Table 11 presents a description of the set of hyperparameter values for each classifier. The combination of hyperparameters that maximizes the quality metric:

$$balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (24)$$

averaged over the three blocks is chosen as optimal.

Table 11. Description of the hyperparameter grid to find the optimal combination on the training dataset (for each iteration of the initial data partitioning)

Classifier	Hyperparameter	Investigated values
kNN	Number of neighbors (n_neighbors)	[1, 2, …, 5]
kNN	Type of weighing (weights)	[uniform weighing, weighing by the Euclidean metric]
LR	Inverse regularization factor (C)	[0.01, 0.02, …, 1.00]
LR	Type of regularization (penalty)	[L1, L2]
RF	Number of base trees (n_estimators)	[10, 20, …, 50]
RF	Maximum tree depth (max_depth)	[1, 3, 5, …, 13]
SVM	Regularization parameter (C)	[0.0001, 0.001, 0.01, …, 10]
SVM	Kernel (kernel)	[linear, kernel with a Gaussian radial basis function]

2.4. Analysis of model gas mixtures for calibration transfer using MS 2.1 and MS 2.2

2.4.1. Scheme of the experimental setup for the analysis of model gas mixtures

Scheme of the installation to assess the possibility of calibration transfer in principle is similar to the scheme shown in Figure 6. The difference of the current scheme (Figure 10) from the above is only in the value of several parameters, namely: the flow rate of the gas mixture or room air flow to form the baseline - 0.4 l/min, the duration of sample feeding - 90 s, the time period of signal integration - 300 s.

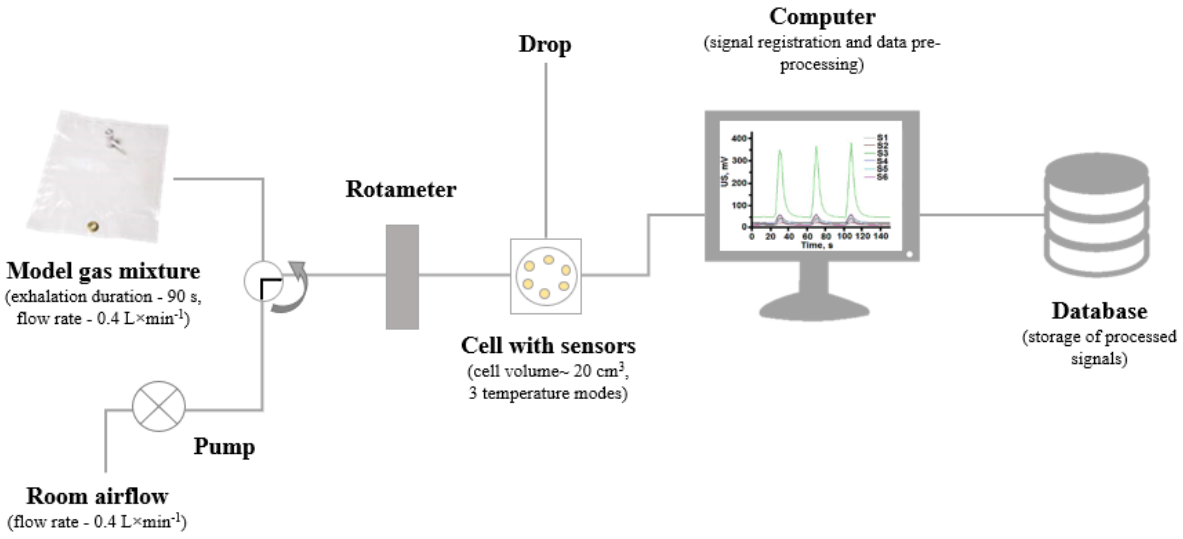

Figure 10. Experimental scheme of the analysis of model gas mixtures to assess the possibility of calibration transfer for MS 2.1 and MS 2.2

Table 12 shows the basic parameter values of the analysis for MS 1, MS 2.1 and MS 2.2.

Table 12. Parameters of sensor conductivity measurement for multisensory systems in the analysis of model gas mixtures

Parameter	**MS 1**	**MS 2.1 and MS 2.2**
Sampling duration	10 s	90 s
Flow rate of the gas mixtures and baseline room air	3.5 L/min	0.4 L/min
Time period of signal integration from the time of sample feeding	90 s	300 s

2.4.2. Data processing

Python 3.6 software (Python Software Foundation, USA) and libraries pandas, matplotlib, numpy and scikit-learn were used for data analysis, visualization and processing, PCA application, as well as for classification models (MSVM and SVM) and regression models (UDS, UDSwoi, DS-PLS2, DS-L1R).

For modeling and experiments to assess the possibility of conducting the calibration transfer using response standardization methods, data obtained from the analysis of calibration

samples of one-component gas mixtures of three VOCs and samples of two gas mixtures with identical qualitative composition by the set of entering VOCs, but different in total quantitative composition were collected. Table 13 and 14 show the concentrations of the above samples. In the case of samples of one-component mixtures, 2 samples were prepared for each concentration and each component and the total number of samples analyzed for each MS was 42. In the case of samples of gas mixtures consisting of three VOC components, 8 samples were prepared for each mixture. The total number of analyzed samples for each MS was 16.

Table 13. Composition of one-component calibration gas mixtures for modeling the classification task

	Concentration, ppm		
Sample number	**propan-1-ol**	**n-heptane**	**o-xylene**
1	1.2	1.2	1.5
2	2.4	2.4	3.0
3	6.0	6.1	7.4
4	11.9	12.1	14.9
5	23.9	24.3	29.7
6	59.6	60.7	74.3
7	119.3	121.5	148.5

Table 14. Composition of gas mixtures (GM 1 and GM 2) for modeling the classification task

# of gas mixture	**Component**	**Concentration, ppm**
1	propan-1-ol	33
	n-heptane	17
	o-xylene	20
2	propan-1-ol	26
	n-heptane	17
	o-xylene	24

The original sensor response matrices with information about the qualitative composition of the samples obtained in the analysis of one-component calibrated gas mixtures and three-component mixtures are located in the repository [99].

Scheme of the experiments to assess the possibility of conducting the calibration transfer using standardization methods is shown in figure 11. At the end of the experiment, we compared the average values of multiclass classification accuracy (in the case of single-component gas mixtures) and binary classification accuracy (in the case of multicomponent mixtures) for each combination of "model - test set for training" by 15 random partitions of the data set into training and test sets in the ratio 70% to 30% and random selection of standardization samples. During the training phase, an internal cross-validation on the training dataset with the number of blocks equal to three is performed for each algorithm at each iteration. Table 15 presents a description of the set of hyperparameter values for the MSVM and SVM classifiers used. The detailed experiment algorithm can be found in the script located in the repository [99].

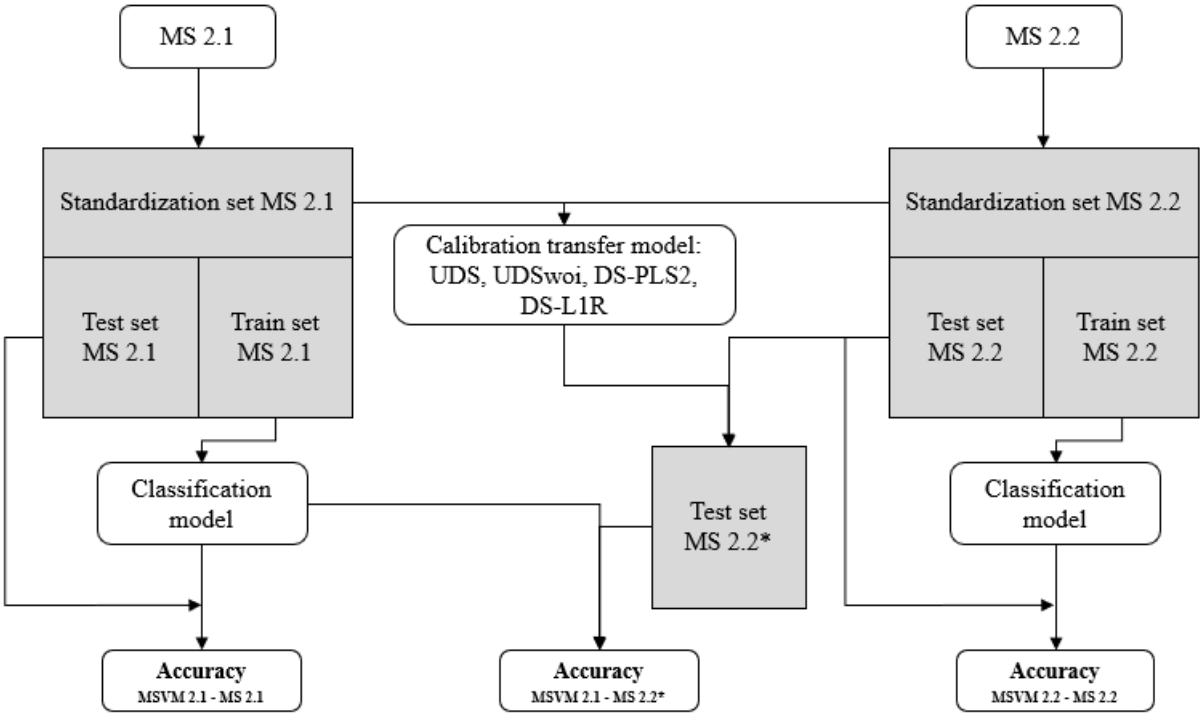

Figure 11. Scheme of data processing for two sensor arrays and correction of MS 2.2 responses using methods of calibration transfer

Table 15. Description of the hyperparameter grid for finding the optimal combination on the training dataset

Classifier	**Hyperparameter**	**Investigated values**
MSVM	Regularization parameter (C)	$[10^{-2}, 10^{-1}, \ldots, 10^{-2}]$
SVM	Regularization parameter (C)	$[10^{-8}, 10^{-7}, \ldots, 10^{-4}]$

Chapter 3. Development of an online analysis of exhaled breath method for the diagnosis of lung cancer using a multisensory system

3.4. Description of the medical study

All studies involving both patients and healthy volunteers were conducted in accordance with ethical standards and the Declaration of Helsinki of 1964. All experiments were performed with the permission of the N.N. Petrov National Medical Research Center of Oncology by Ethics Committee No. 15/83, dated March 15, 2017 [1,100]. The MS 1 multisensory system was installed in the research center. The following inclusion criteria were used: patients aged 20 years or older; suspicion of LC based on clinical symptoms or radiological examination (patients were given the opportunity to have all necessary additional diagnostic tests such as CT scan, fibrobronchoscopy and/or transthoracic trepan biopsy); active smoker or past smoker who had quit no more than 10 years before participation in the study. Exclusion criteria for the study: Patients with severe comorbidities (including decompensated cardiovascular, endocrine, such as diabetes, or pulmonary pathology; decompensated or subcompensated organ failure; uncorrectable coagulopathies; cerebrovascular disorders; unstable angina pectoris); history of cancer 5 years prior to study, excluding skin cancer or cervical cancer in situ; history of LC or lung surgery; expected survival 1 year or less.

In the study [100] EB samples from 118 patients (49 females (42%) and 69 males (58%)) were analyzed, including 65 patients with LC and 53 healthy patients who voluntarily participated in the study and signed consent to participate. In all patients who were in the RL group, the diagnosis was confirmed by morphological examination after EB analysis. The following forms of LC were diagnosed: non-small cell LC - 59, small cell LC - 6. Thirty patients had early-stage disease (I/II) and 35 had late-stage disease (III/IV). In addition, the control group included 54 healthy volunteers without signs of lung disease based on clinical symptoms and radiological examination. Data on sex and age of the control group are also presented in Table 16.

Table 16. Characteristics of the studied groups (LC group and healthy group)

Group	LC group	Healthy group
Number of participants	65	53
Age, average ± sd*	65 ± 9	56 ± 12
Male	42 (65%)	27 (51%)
Female	23 (35%)	26 (49%)

*sd – standard deviation

Patients were informed to abstain from eating and smoking for one hour before EB analysis with MS 1. Immediately before analysis patients rinsed their mouths with clean warm water and waited in the room where measurements were taken for at least 10 minutes before the first measurement.

3.4. Description of patient exhaled breath analysis procedure

The analysis procedure of patient EB consisted of three consecutive steps. In the **first step** the pump valve is opened and ambient air is blown through the gas cell at 3.5 $L \times min^{-1}$ while the patient valve is closed (as shown in Figure 12). In the **second step** the pump valve closes and the patient valve opens. The patient exhales through a sterile disposable mouthpiece for 10 seconds through the gas chamber at a rate of 3.5 $L \times min^{-1}$ which corresponds to a volume of about 600 ml of EB. The rate was controlled by the patient with monitoring of rotameter readings. In the **third step** the patient valve was closed, the pump valve was opened and ambient air was blown through the gas cell again at the same rate of 3.5 $L \times min^{-1}$. The flow rate is also monitored using the rotameter.

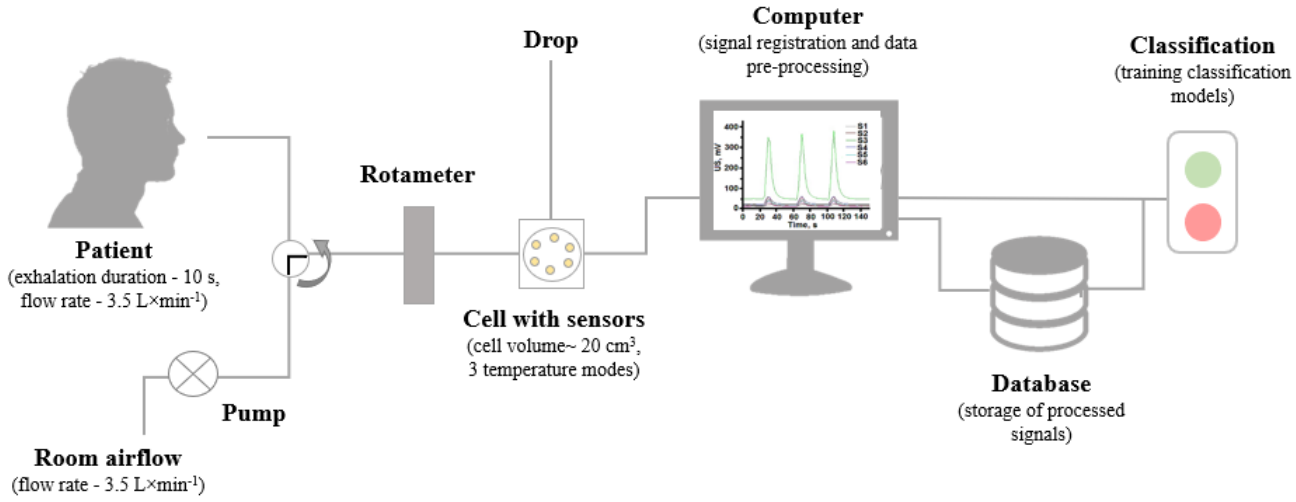

Figure 12. Scheme of the online analysis of EB

The conductivity integral over time minus the area formed by the baseline was extracted from the conductivity curve as an analytical signal for each sensor at each temperature. The time between successive measurements at the same sensor temperature was 1-2 minutes for one patient.

The reproducibility of the extracted analytical signal for all sensors was about 2-15%. Initially three consecutive EB measurements were performed for each of the three temperature regimes. During the course of the study, it was found possible to perform only one measurement on each temperature mode resulting in a total analysis time for one patient ranging from 15 to 25 minutes.

The higher the humidity level the less the MO sensors are affected by its fluctuations. Therefore, all measurements were performed in rooms with relatively high humidity levels (60% ± 5%). Due to the use of high operating temperatures the effect of temperature fluctuations in the air passing through the cell is insignificant. Therefore, no preheating was used for the cell.

3.4. Selection of the most effective data processing algorithm and classification model

The resulting feature matrix X has a dimension of 118×18, where 118 is the total number of patients, and 18 is the number of extracted features (integral of the conductivity peak from

time) from 6 sensors on 3 temperature modes. The y vector has a dimensionality of 118×1 with a value of 0 (the patient belongs to the healthy group) or 1 (the patient belongs to the LC group).

First of all, the classes are not balanced (53 vs. 65 for the healthy group and LC group respectively) so stratification by class was used in the cross-validation partitions so that the ratio of the LC group to the control group was the same between partitions.

To study and visualize the data in two-dimensional space 3 matrices were built for each temperature mode with point scatter diagrams. More intergroup separation is observed for the first temperature mode (Figure 13). As it can be seen visually the greatest separation between the LC group and the healthy group is observed for pairs with the presence of sensor #4.

The Shapiro-Wilk criterion [101] which is one of the most effective criteria for testing normality was used to test the distribution of the signs for normality. The hypothesis of distribution normality was rejected for all the features (α=0.05). Figure 9 suggests a lognormal distribution of the data in the features. Indeed, the hypothesis of normal distribution was already rejected for only 6 out of 18 features (α=0.05) upon log transformation of the feature matrix (Table 17). Nevertheless, most of the above-mentioned classifier models do not require the features to be normally distributed. Since the data are not normally distributed it is worth excluding the frequently used LDA-based classifier from consideration.

For the PCA analysis, the data were scaled and centered beforehand. Figure 14 shows the dependence of the explained variance on the number of principal components used in the model. Figure 15 shows the representations of the EB samples in the space of 3 principal components.

Figure 13. Matrix of scatter plots for 6 sensors at temperature mode T1. Cells (i,j) contain the scattering point diagram for the conductivity integral (μS×s) of i and j sensors. The diagonal cells (i,i) show the smoothed distribution for sensor i. The grouping of the EB sample is shown in color (blue - control group (0), orange - LC group (1))

Table 17. Check for normality of the original data and transformed by natural logarithms using the Shapiro-Wilk criterion (α=0.05)

	Internal data			**Data after log transformation (natural logarithm)**		
#	**W-value**	**p-value**	**H0 is rejected**	**W-value**	**p-value**	**H0 is rejected**
S1_T1	0.930	1×10^{-5}	yes	0.982	0.116	no
S2_T1	0.872	1×10^{-8}	yes	0.990	0.512	no
S3_T1	0.917	2×10^{-6}	yes	0.986	0.259	no
S4_T1	0.875	2×10^{-8}	yes	0.978	0.051	no
S5_T1	0.917	2×10^{-6}	yes	0.990	0.513	no
S6_T1	0.900	2×10^{-7}	yes	0.983	0.154	no
S1_T2	0.853	2×10^{-9}	yes	0.995	0.933	no
S2_T2	0.823	1×10^{-10}	yes	0.993	0.844	no
S3_T2	0.897	2×10^{-7}	yes	0.976	0.034	yes
S4_T2	0.847	1×10^{-9}	yes	0.977	0.037	yes
S5_T2	0.704	4×10^{-14}	yes	0.987	0.299	no
S6_T2	0.897	2×10^{-7}	yes	0.987	0.347	no
S1_T3	0.679	1×10^{-14}	yes	0.974	0.021	yes
S2_T3	0.692	2×10^{-14}	yes	0.970	0.009	yes
S3_T3	0.785	7×10^{-12}	yes	0.978	0.047	yes
S4_T3	0.733	2×10^{-13}	yes	0.969	0.009	yes
S5_T3	0.806	4×10^{-11}	yes	0.987	0.307	no
S6_T3	0.822	1×10^{-10}	yes	0.985	0.199	no

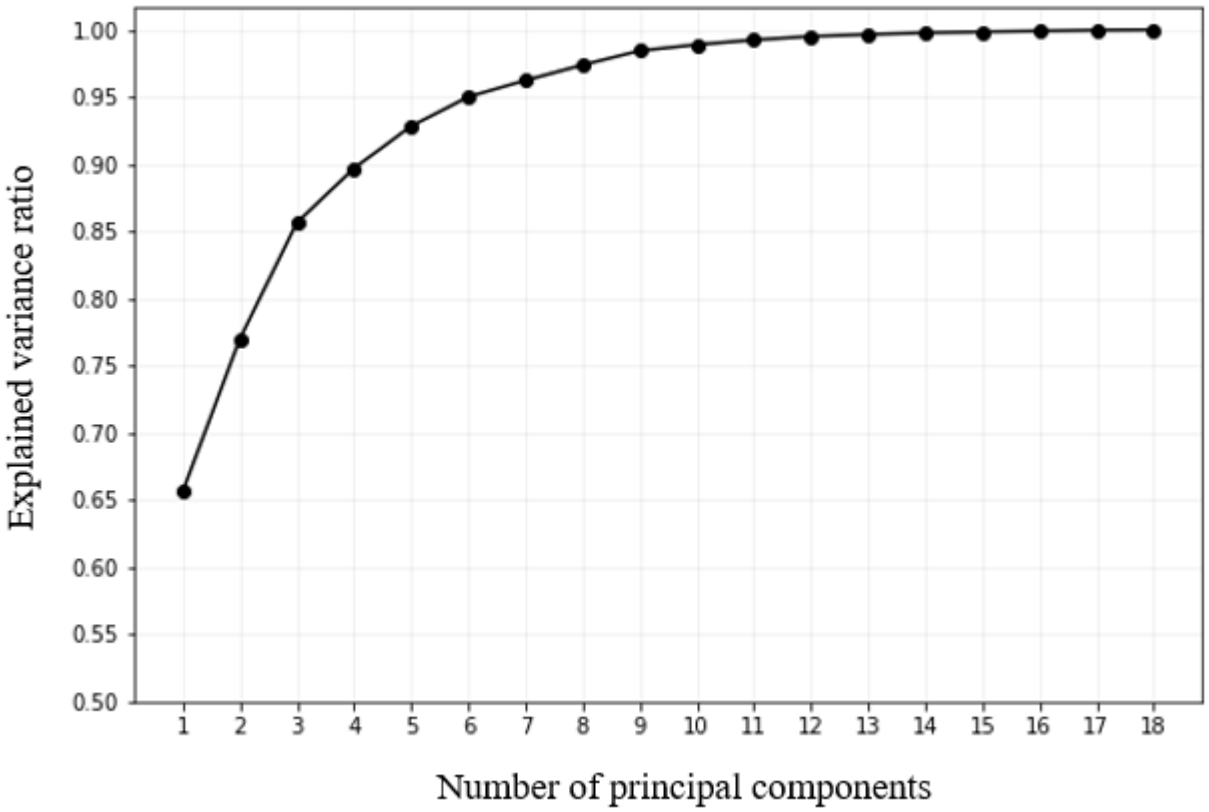

Figure 14. Dependence of the explained variance ratio on the number of principal components in the model

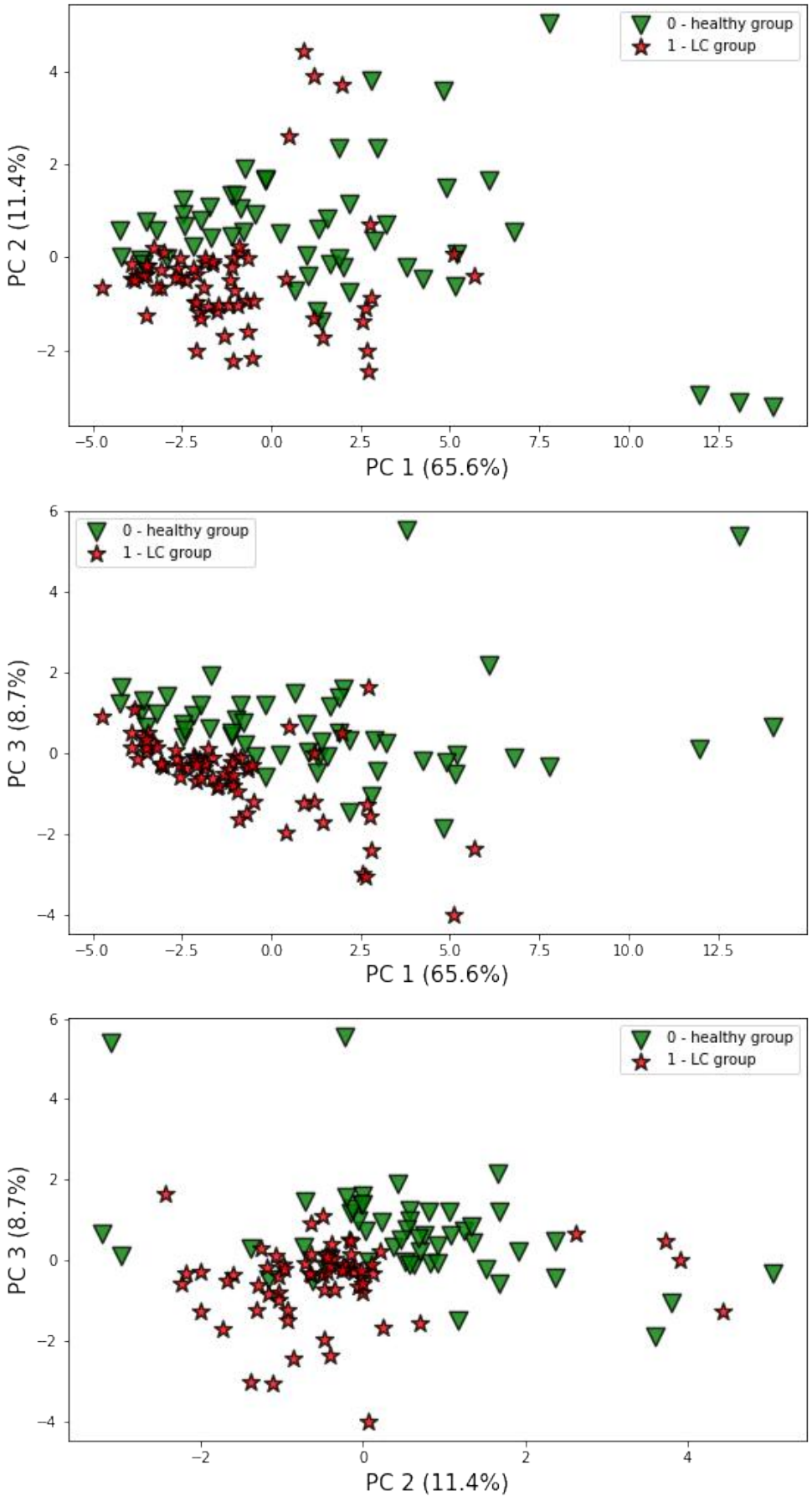

Figure 15. EB samples in the space of the pairs of three principal components (PC1-PC2, PC1-PC3, PC2-PC3)

According to the data processing scheme shown in Figure 9 for each iteration (15 in total) a random partitioning of the sensor response matrix and class label vector was performed on the training and test data sets in the ratio of 60 to 40. Then the considered classification models were trained on the training sets taking into account the internal cross-validation to select the optimal values of hyperparameters for each model. The trained models were then used to predict probabilities and class labels for both the training and test set samples.

Figure 16 shows the constructed ROC curves for the classification models for one of the iterations. After the ROC curves were constructed the ROC AUC metrics were calculated.

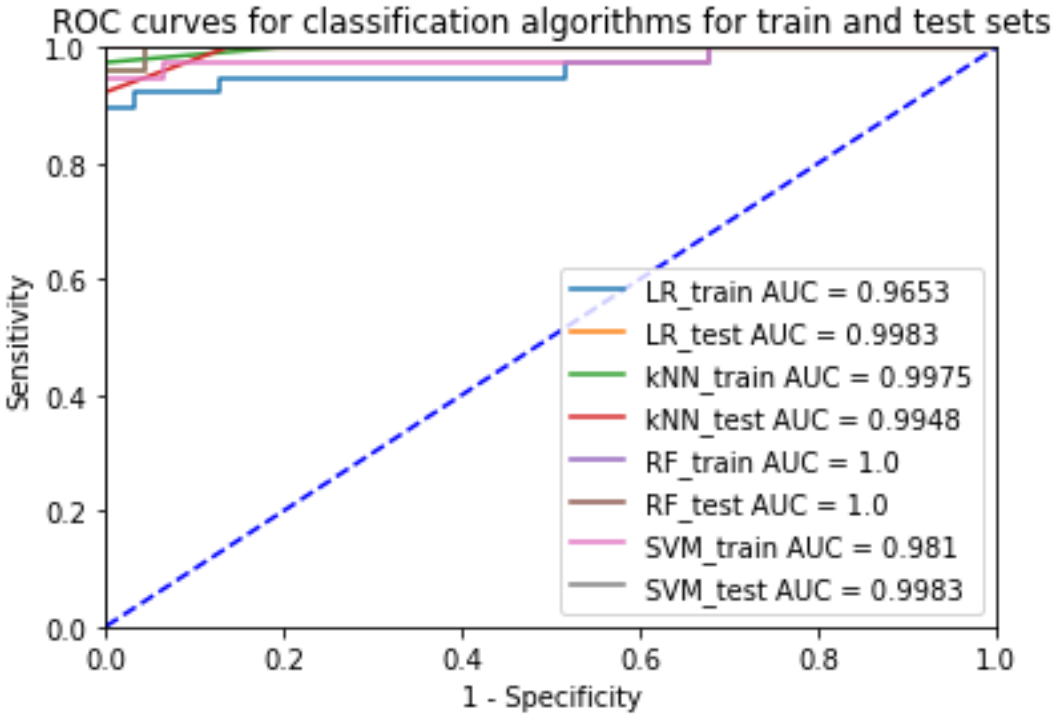

Figure 16. ROC curves of classification algorithms obtained when predicting class labels for one data partitioning into training and test sets

Using a standard classification threshold (p=0.5) each sample was classified as true positive (TP), false positive (FP), true negative (TN) or false negative (FN). Next confusion matrices were constructed. Tables 18 and 19 show examples of constructed error matrices for one of the iterations for the training and test sets respectively.

Table 18. Class prediction confusion matrices for a training dataset on a single random partition

kNN		Predicted class	
		1	0
Actual Class	1	36	0
	0	3	31

LR		Predicted class	
		1	0
Actual Class	1	36	3
	0	2	29

RF		Predicted class	
		1	0
Actual Class	1	39	0
	0	0	31

SVM		Predicted class	
		1	0
Actual Class	1	37	2
	0	0	31

Table 19 Class prediction confusion matrices for a test dataset on a single random partition

kNN		Predicted class	
		1	0
Actual Class	1	24	2
	0	0	22

LR		Predicted class	
		1	0
Actual Class	1	26	0
	0	2	20

RF		Predicted class	
		1	0
Actual Class	1	26	0
	0	0	22

SVM		Predicted class	
		1	0
Actual Class	1	25	1
	0	1	21

Then diagnostic test metrics were calculated for each classification model and type of data set used, and then averaged over all 15 iterations. Table 20 summarizes the averaged quality metrics (sensitivity, specificity, and ROC AUC) for both the classification of samples from the training set and for the test dataset.

Table 20. Mean values of metrics with confidence interval (at significance level α=0.05) when classifying algorithms on the training and test dataset at 15 times random partitioning of the dataset

Metric	Dataset type	Classification model			
		LR	kNN	SVM	RF
ROC AUC	Training	0.981±0.007	0.998±0.001	0.993±0.005	0.999±0.000
	Test	0.959±0.012	0.961±0.018	0.972±0.014	0.960±0.014
Acc	Training	0.956±0.015	0.970±0.012	0.990±0.007	0.988±0.007
	Test	0.931±0.018	0.940±0.016	0.936±0.016	0.898±0.023
Se	Training	0.957±0.011	0.947±0.022	0.986±0.011	0.981±0.012
	Test	0.938±0.021	0.905±0.026	0.920±0.025	0.853±0.041
Sp	Training	0.954±0.025	1.000±0.000	0.995±0.006	0.997±0.004
	Test	0.924±0.036	0.981±0.015	0.954±0.030	0.951±0.026
PPV	Training	0.964±0.018	1.000±0.000	0.996±0.004	0.998±0.003
	Test	0.939±0.027	0.983±0.013	0.962±0.024	0.956±0.023
NPV	Training	0.946±0.014	0.939±0.025	0.983±0.013	0.977±0.014
	Test	0.929±0.023	0.899±0.027	0.913±0.025	0.851±0.035

3.4. Analysis of the results

Table 12 shows that the classification results on the test datasets are somewhat worse than on the training datasets. It should be noted that this is a common situation which is related to the retraining of models. Still under real conditions the metrics values obtained on the test dataset will be closest to the achieved results of the diagnostic test. Thus, we will use the results on the test sets to select the optimal classifier (Table 21). In this study high values of the metrics were achieved ROC AUC. It is difficult to identify any particular algorithm here due to the fact that the differences between the mean values are not statistically significant (at a significance level of α=0.05). Note that in this paper this ROC AUC metric can be interpreted as the probability with which a randomly selected patient with LC will be assigned a higher weight than a random healthy patient. Turning to the main metrics assessed in the pilot works on the development of MS-based diagnostic tests, namely, Se and Sp, less frequently PPV and NPV, it is worth noting

that in real conditions of screening examination, the ratio of patients with the disease to the number of healthy patients will differ markedly from 1:1. Therefore, consideration of the diagnostic test first begins with Sp and NPV. The corresponding values should be at least 0.98-0.99, otherwise, the number of unnecessary biopsies with the risk of complications increases dramatically, the use of additional examination methods increases as well as the cost of the examination. The requirements for Se and PPV are less strict - for Se not less than 0.90, for negative predictive value – 0.85. According to the results, we can see that the model based on the kNN classifier meets the above requirements with Se – 0.905±0.026, Sp – 0.981±0.015, PPV – 0.983±0.013, NPV – 0.899±0.027.

Table 21. Classification results of model on the test sets

Metric	**Classification model**			
	LR	**kNN**	**SVM**	**RF**
ROC AUC	0.959±0.012	0.961±0.018	0.972±0.014	0.960±0.014
Acc	0.931±0.018	0.940±0.016	0.936±0.016	0.898±0.023
Se	0.938±0.021	0.905±0.026	0.920±0.025	0.853±0.041
Sp	0.924±0.036	0.981±0.015	0.954±0.030	0.951±0.026
PPV	0.939±0.027	0.983±0.013	0.962±0.024	0.956±0.023
NPV	0.929±0.023	0.899±0.027	0.913±0.025	0.851±0.035

Nevertheless, the set of classifiers used allows us to estimate the importance of a feature in the classification model. For example, the importance of the features in LR can be represented as absolute values of the model coefficients (Figure 17). Since this model uses L1 regularization for the least important features the model zeroes the coefficients during training.

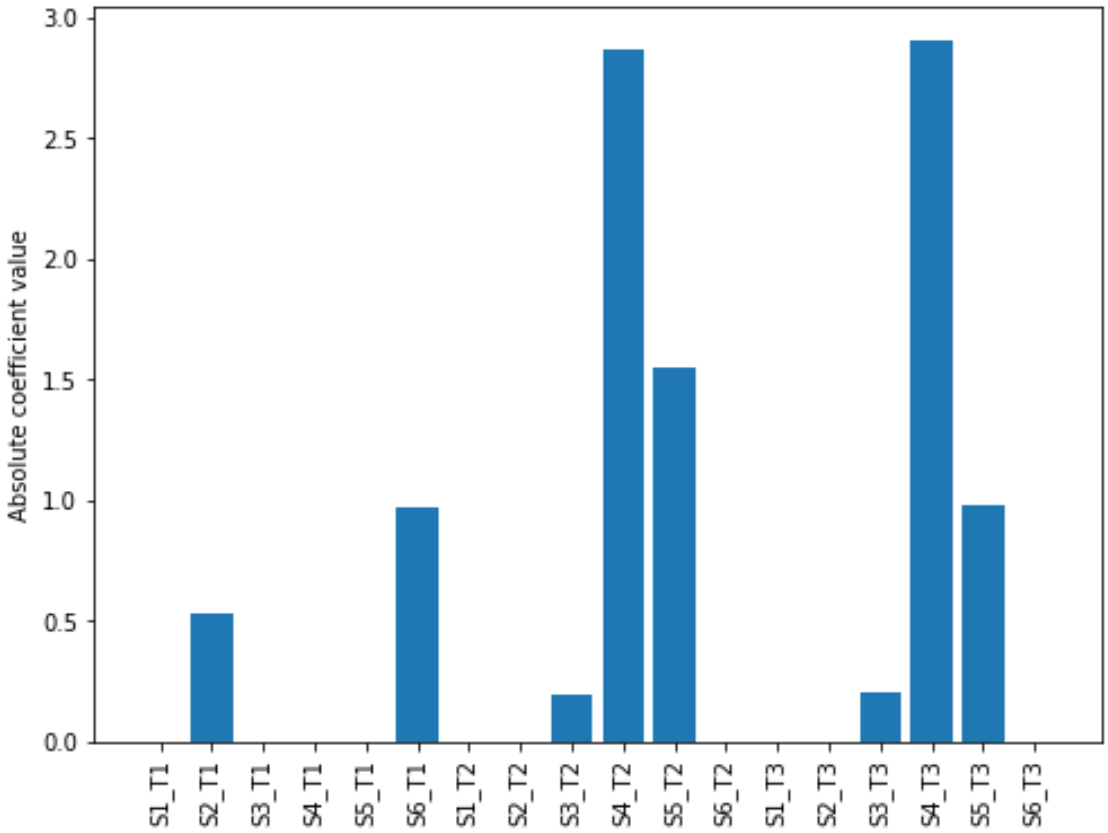

Figure 17. Feature importance in LR based on absolute values of model coefficients

Since each sensor is used at different temperatures, we expect some collinearity between the data of one sensor at these temperatures. Thus, the effective number of sensors is less than 18. This can also be observed when evaluating the importance of the features (Figure 14) in the LR model. Using similar approaches in investigating the contribution of the features it is possible to optimize the qualitative composition of the sensors in the working set. Also from Figure 14 we can distinguish sensor S4 as the most important in classification which was noted earlier in the analysis of the pairwise scatter plots.

The results of this pilot study demonstrated the applicability of online analysis of EB using an array of MO gas-sensitive sensors for diagnostic purposes. Separation of patients of LC group and healthy group was observed with acceptable levels of sensitivity and specificity. In Table 22 results from current work using the kNN-based model are added to the pilot studies on separation of the LC group and the healthy group for comparative evaluation.

Table 22. Comparison of obtained results with other pilot works using gas-sensitive MO sensors (Se - sensitivity, Sp - specificity, Acc - accuracy)

Reference	Group description	Se	Sp	Acc
This work (kNN)	N=118 (65 LC, 53 healthy group)	90.5 ± 2.6%	98.1 ± 1.5%	94.0 ± 1.6%
[70]	N=101 (43 LC, 58 healthy group)	95.3%	90.5%	92.6%
[72]	N=18 (9 LC, 9 healthy group)	100%	88.9%	94.4%
[73]	N=89 (47 LC, 42 healthy group)	93.6%	83.3%	-
[74]	N=76 (31 LC, 45 healthy group)	-	-	88%
[75]	N=37 (12 LC, 25 healthy group)	83%	88%	-
[76]	N=84 (32 LC, 52 healthy group)	85%	84%	-
[77]	N=290 (144 LC, 146 healthy group)	94.4%	32.9%	-
[78]	N=145 (52 LC, 93 healthy group)	83%	84%	-
[79]	N=16 (6 LC, 10 healthy group)	85.7%	100%	93.8%

Implementing the online mode avoids additional sampling and sample preparation procedures that can negatively affect the quality of results due to loss or contamination of VOCs. Since the use of sampling bags can lead to contamination or uncontrolled sorption of important VOCs, this greatly affects the results. And procedures such as pre-cleaning a patient's lungs with medical air can lead to uncontrolled changes in a patient's VOC profile and, as a result, incorrect patient separation.

Conclusions

Memory effects are also significantly reduced in online mode with ambient air pumping at increased speed. The developed approach provides adequate EB analysis results even with a limited number of sensors. On the other hand, by using three temperature modes – the number of sensors is tripled, since switching from one mode to another significantly changes the relative sensitivity of each sensor.

The results obtained in this work show the possibility of implementing the presented MS system in diagnostic practice. In particular, the values of the main informativeness criteria of the developed diagnostic test (sensitivity, specificity, PPV, NPV, ROC AUC), which are most important for medical personnel when evaluating diagnostic systems, are among the highest in comparison with other studies (Table 14) mentioned above [70-79].

Chapter 4. Development of a calibration transfer and response standardization method between two multisensory systems

4.1. Description of the study design

As it was mentioned earlier, the existing technologies of manufacturing of MO sensors do not allow obtaining such sensors, which would have identical characteristics and, consequently, identical character of the response to the analyte. This fact limits the use of the developed classification model for other MS consisting of a set of corresponding sensors. And the obtained informativity criteria for the developed model of classification on new samples can be expected only if the samples were measured on the same set of sensors.

The following experiment was performed to check the possibility of two MFs working without any adjustments. Responses of calibration samples for two MS (MS 2.1 and MS 2.2) were recorded. Information about the MS sensor set is presented in section 2.1, information about the composition and concentrations of the calibration samples is presented in section 2.4.2 in tables 13 and 14. Multiclass classification models based on the Crammer-Singer support vector machine (MSVM) MSVM 2.1 and MSVM 2.2 were trained on the training set of the MS 2.1 and MS 2.2 array calibration sample responses, respectively. Next, classification was performed on a test set of graded sample responses for the MS 2.1 array using the MSVM 2.1 model and for the MS 2.2 array using the MSVM 2.1 and MSVM 2.2 models. As a result, we obtained classification accuracy for the following "model – test set source" combinations: MSVM 2.1 - MS 2.1, MSVM 2.1 - MS 2.2, MSVM 2.2 - MS 2.2. Figure 18 shows the results obtained on the test response sets for 15 random partitions of the data into the training and test sets. The data partitioning was conducted in a 70%/30% ratio respectively.

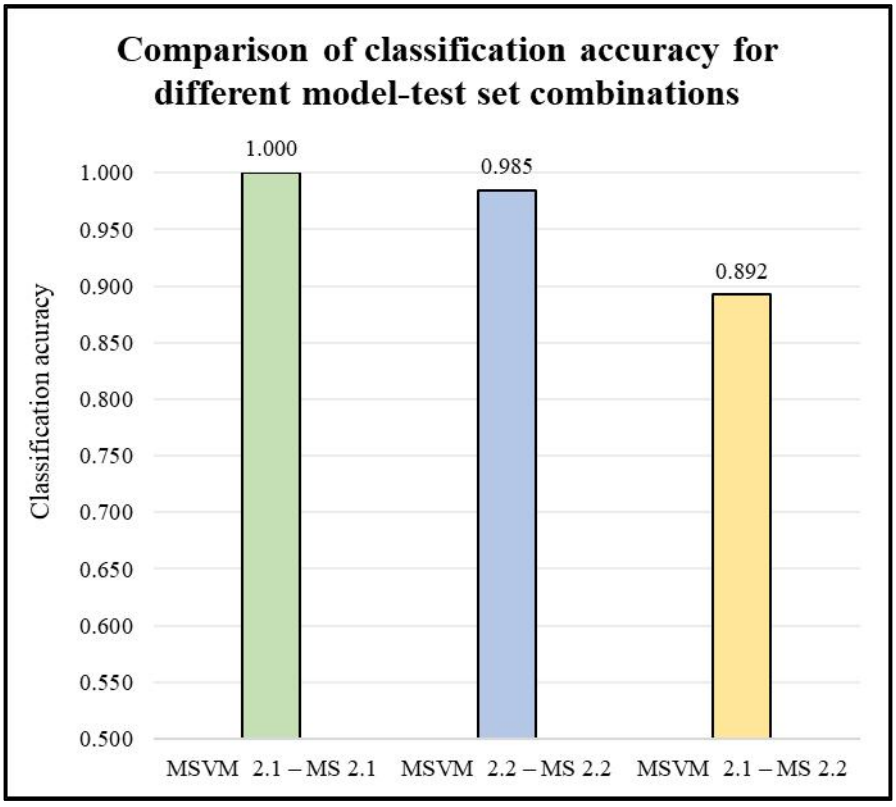

Figure 18. Evaluation of the change in classification accuracy when using the classification model trained on MS 2.1 samples for the MS 2.2 test set (MSVM 2.1 - MSVM 2.2). The graph shows the average accuracy values for 15 random partitions of the original response set into the training set and the test set in a 70/30 ratio. The standard deviation for each of the three series is zero.

According to the obtained computations it is clear that models with training and test sets of samples measured on the same MS classify test samples almost error-free (1.000 for MSVM 2.1 - MS 2.1 and 0.985 for MSVM 2.2 - MS 2.2); when using different data sources for training model and test the classification results are somewhat worse (0.892 for MSVM 2.1 - MS 2.2). The results once again confirm that for a model trained on data from one sensor array, one should not expect comparable classification accuracy results on data from the same samples from another sensor array with similar but physically different sensors. Therefore, to solve the problem, we considered the option of working with mathematical transformations to standardize the data between MS.

To study the possibility of joint operation of two arrays of similar sensors by correcting the responses of one of the sets using methods of transferring calibration dependences,

experiments were conducted on model gas mixtures (air – VOC) in accordance with the scheme shown in Figure 11.

The principle of the calibration transfer methods based on the correction of the sensor array responses is described in chapter 1.4.1 of this paper. In this work four methods were used to correct the responses:

Univariate direct standardization (**UDS**) *method*. For each pair of corresponding sensors from two systems (MS 2.1 and MS 2.2) using the method of least squares we determined linear regression coefficients of the form:

$$X_j^{MS\ 2.1} = k_j * X_j^{MS\ 2.2} + b_j \tag{25}$$

Next, the MS 2.2 test set was adjusted to the relationship found.

Univariate direct standardization method without intercept (**UDSwoi**). The principle is similar to the UDS method, except that the found value of the free term b_j is not used in the equation of relationship:

$$X_j^{MS\ 2.1} = k_j * X_j^{MS\ 2.2} \tag{26}$$

Direct standardization method with L1 regularization (**DS-L1R**). This type of regularization allows to adjust the regularization coefficient so as to obtain the required number of non-zero terms in the constructed regression model. This feature allows to use the response standardization procedure with the number of standardization samples smaller than the number of independent variables in the system. An additional regularization term is introduced into the optimized model functional, and the estimation of model parameters $\hat{\beta}$ is expressed by the following formula:

$$\hat{\beta} = arg\ min\left(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}\beta_j x_{ij})^2 + \lambda|\beta|\right) \tag{27}$$

Direct standardization method with regression on latent structures 2 (**DS-PLS2**). In brief, this approach simultaneously decomposes the response matrices of standardization samples for MS 2.1 and MS 2.2:

$$\mathrm{X}^{MS\ 2.2} = TP^T + E; \qquad X^{MS\ 2.1} = UQ^T + F; \qquad T = XW + G\ ; \tag{28}$$

The construction of projections is consistent with the maximization of the correlation between the corresponding vectors $\mathrm{X}^{MS\ 2.2}$-scores t_a and $\mathrm{X}^{MS\ 2.1}$-scores u_a taking into account the given number of principal components.

The efficiency of calibration transfer and the possibility of scaling the classification model from one sensor array to another in this experiment was evaluated by the accuracy of multiclass classification.

Special attention should be paid to the method of taking standardized samples. Recall that by this term we mean a number of identical samples, measured on all devices, for which a response correction is necessary. In analytical chemistry, standard samples are used primarily for this kind of problem, but when analyzing samples of complex composition for which there are no certified samples, the above term is used. In addition to randomized selection, mathematical algorithms are also used. Most often in scientific works to select the minimum in the number of samples, but close to the representativeness of the entire sample uses the algorithm Kennard-Stone algorithm [95]. In this work, the task was to evaluate the efficiency of standardization on average, i.e., without considering the influence of the qualitative composition of the standardization sample set and taking into account the partitioning of the data for training the model and test, so to obtain stable results, multiple calculations of classification accuracy were performed with randomized samples for standardization and randomized partitioning of the data set into the training and test sample, respectively.

In order not to compromise the results of the experiment it was ensured that there were no standardization samples in the test data sets.

4.2. Evaluation of standardization results in the classification of individual VOC samples

As a result of repeated experiments in accordance with the scheme shown in Figure 11 for each algorithm of standardization of the response, the average values of accuracy of multiclass classification of samples for each number of standardization samples in the range under study (from 2 to 11 inclusive) were obtained. These results are presented in Figure 19.

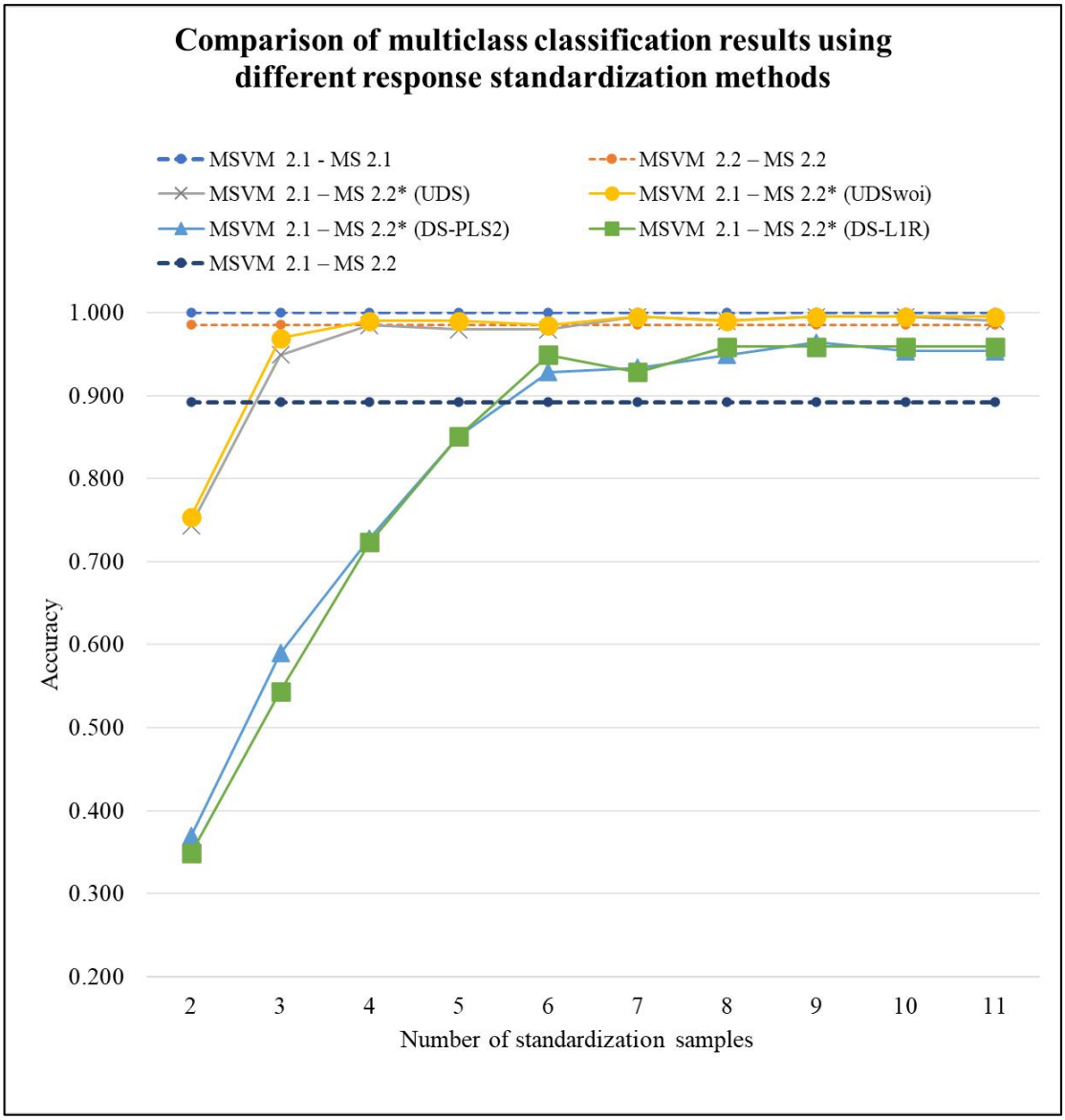

Figure 19. Dependence of the average classification accuracy on the test data set for different methods of response standardization on the number of standardization samples used. (Each point on the graph corresponds to the average value of classification accuracy over 15 repetitions of the experiment with randomized selection of standardization samples and splitting the data into the training and test samples in the ratio 70/30)

To examine the reproducibility of the results for each value of classification accuracy, confidence intervals were estimated (shown in Figure 20). It can be observed that for most models using standardization methods with a small number of samples, the confidence intervals are quite wide and decrease as the number of standardization samples increases. Presumably, as the number of standardization samples increases, the contribution of those standardization samples that poorly describe the overall difference between the sensors decreases. An exception

is the model DS-L1R with regularization, where the width of the confidence does not have any tendency to increase or decrease and is at the same level within the considered range of the number of standardization samples.

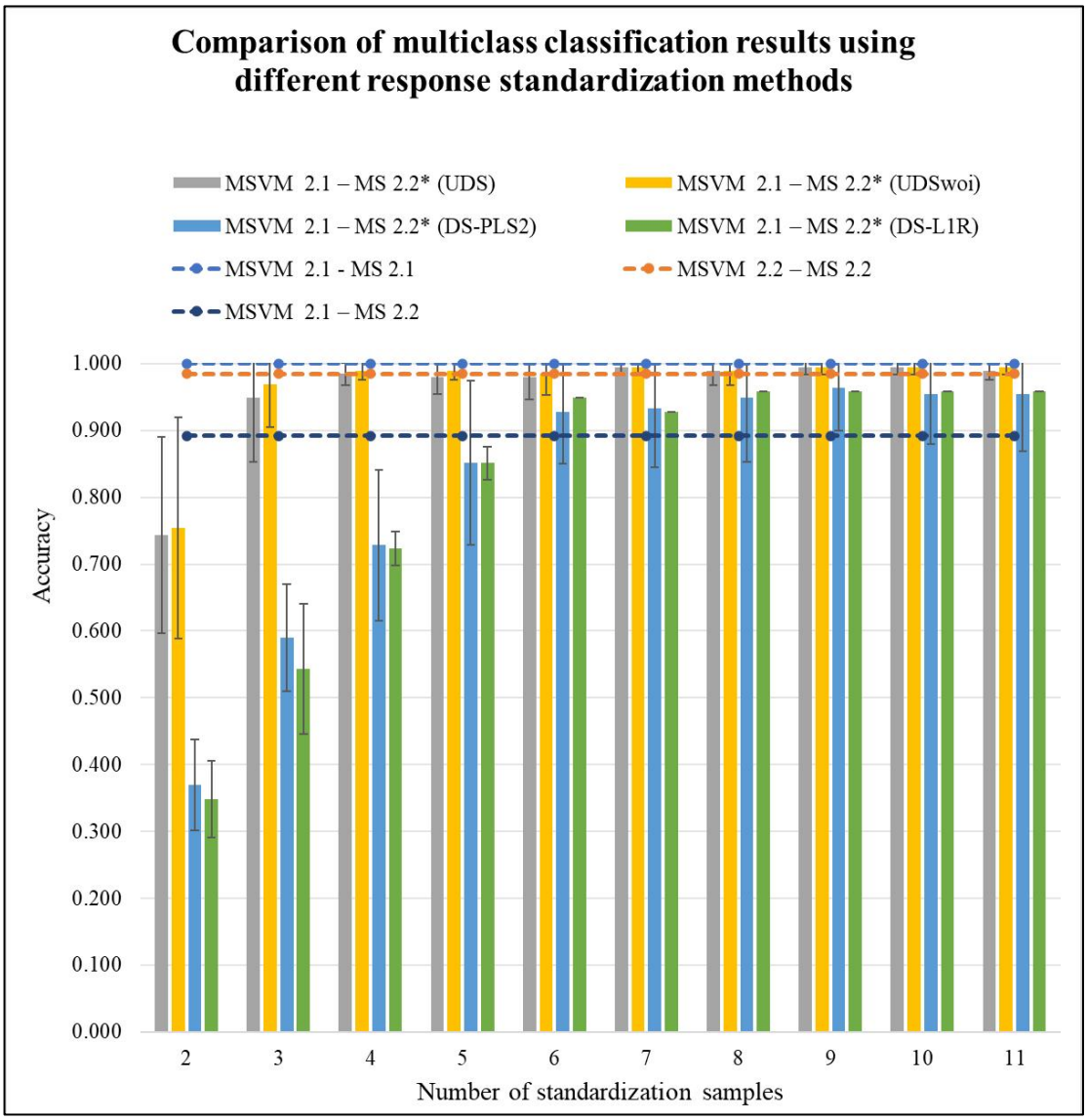

Figure 20. Dependence of the average classification accuracy on the test data set for different methods of response standardization on the number of standardization samples used. (Each point on the graph corresponds to the average classification accuracy over 15 repetitions of the experiment with randomized selection of standardization samples and splitting the data into the training and test samples in the ratio 70/30) The graph also shows the error bars corresponding to the confidence intervals.

On the basis of the obtained results, it is possible to mark out methods UDS and UDSwoi as the most acceptable for standardization of responses MS 2.2 as exactly for these methods it is

possible to reach level of accuracy of classification close to level of model MSVM 2.2 - MS 2.2 for the least quantity of samples for standardization, equal 4.

4.3. Evaluation of standardization results in the classification of VOC mixtures

In research studies related to quantitative multicomponent analysis of EB the authors often point out that the difference in the VOC profile between a group of healthy people and patients with pathology is not the absence or presence of certain VOCs but the range of concentrations. Therefore, this work also modeled the task of reproducing two groups of samples with identical qualitative composition but different quantitative compositions for some components of the mixture. Table 15 shows the compositions of mixtures with VOC concentrations.

Table 15. Composition of gas mixtures (GM 1 and GM 2) for modeling the classification task

№ of mixture	**Mixture component**	**Concentration, ppm**
1	propan-1-ol	33
	n-heptane	17
	o-xylene	20
2	propan-1-ol	26 (-20%)
	n-heptane	17 (+ 0%)
	o-xylene	24 (+20%)

Note that the concentrations of each of the analytes for mixtures 1 and 2 were chosen so that the extracted response for most sensors was close to the response obtained from the analysis of EB samples with an order of magnitude accuracy.

For each MS 2.1 and MS 2.2 system, 8 samples of each GM 1 and GM 2 mixture were prepared and measured. In accordance with the experiment scheme shown in Figure 16, for each response standardization algorithm, average values of binary classification accuracy were obtained for each number of standardization samples in the studied range (from 2 to 7 inclusively). The obtained results are presented in Figure 21.

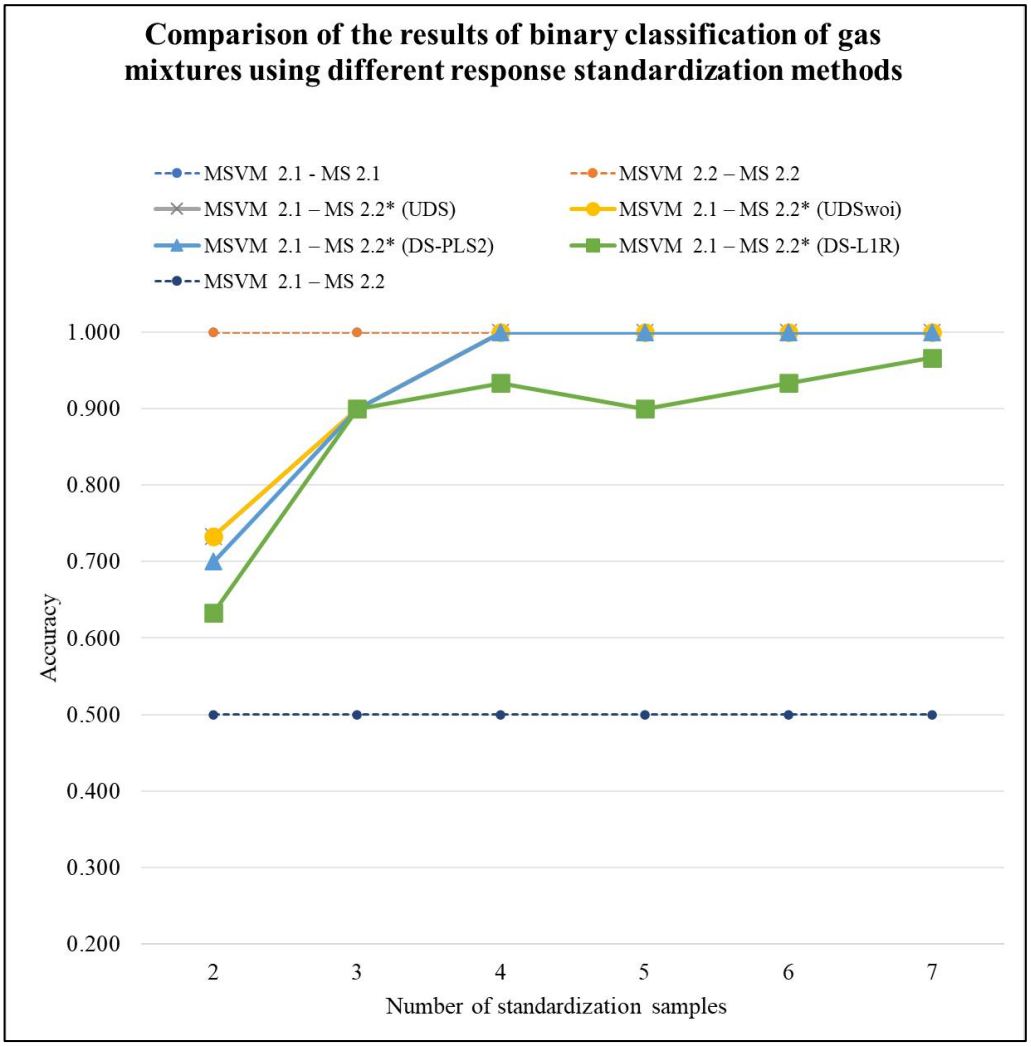

Figure 21. Dependence of the average classification accuracy on the test set of samples (gas mixtures 1 and 2) for different response standardization methods on the number of standardization samples used. (Each point on the graph corresponds to the average value of the classification accuracy over 15 repetitions of the experiment with randomized selection of standardization samples and splitting the data into the training and test samples in the ratio 70/30)

To investigate the reproducibility of the results for each value of classification accuracy, the confidence intervals were estimated (Figure 22). In this case, the maximum reproducibility of the results is achieved for all standardization methods, except for DS-L1R, for which the confidence intervals remain at the same level in the considered range of standardization samples. As in the previous problem, the minimum required number of samples for standardization of MS 2.2 responses was 4.

The PCA method was used to qualitatively evaluate the result of standardization of the responses. Specifically, a PCA model was constructed using MS 2.1 data. Relative explained

variance for the first component was 80.5%, for the second component - 8.1%. Then MS 2.2 data were projected into the principal component space of PCA model. MS 2.2* (UDS) data were also obtained: MS 2.1 data subjected to one-factor standardization over the three standard samples after projection onto the principal components. Figure 23 shows the responses of the graded propan-1-ol, n-heptane, and o-xylene test kit samples for MS 2.1, MS 2.2, and MS 2.2* in the first two principal component space.

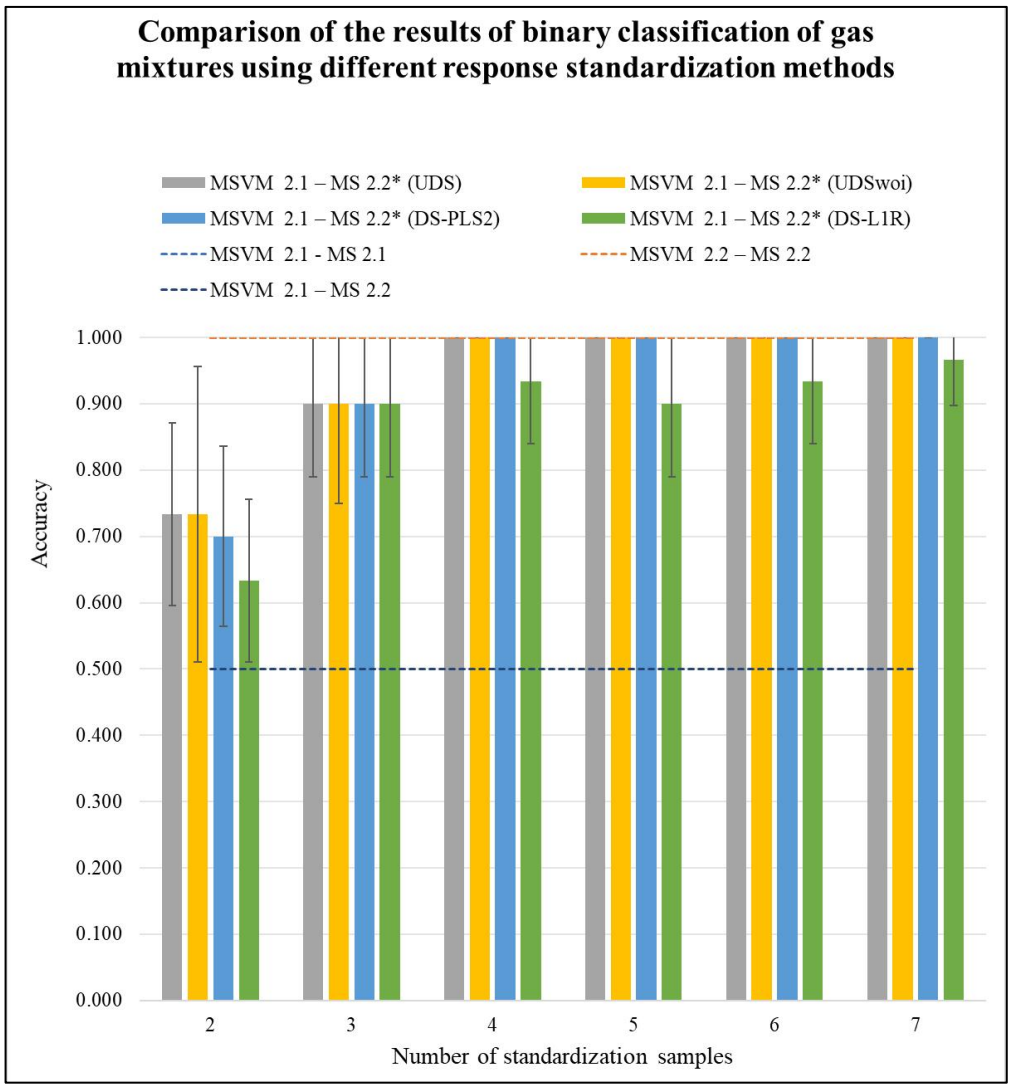

Figure 22. Dependence of the average classification accuracy on the test set of samples (gas mixtures 1 and 2) for different response standardization methods on the number of standardization samples used. (Each point on the graph corresponds to the average value of the classification accuracy over 15 repetitions of the experiment with randomized selection of standardization samples and splitting the data into the training and test samples in the ratio of 70/30) The graph also shows the error bars corresponding to the confidence intervals.

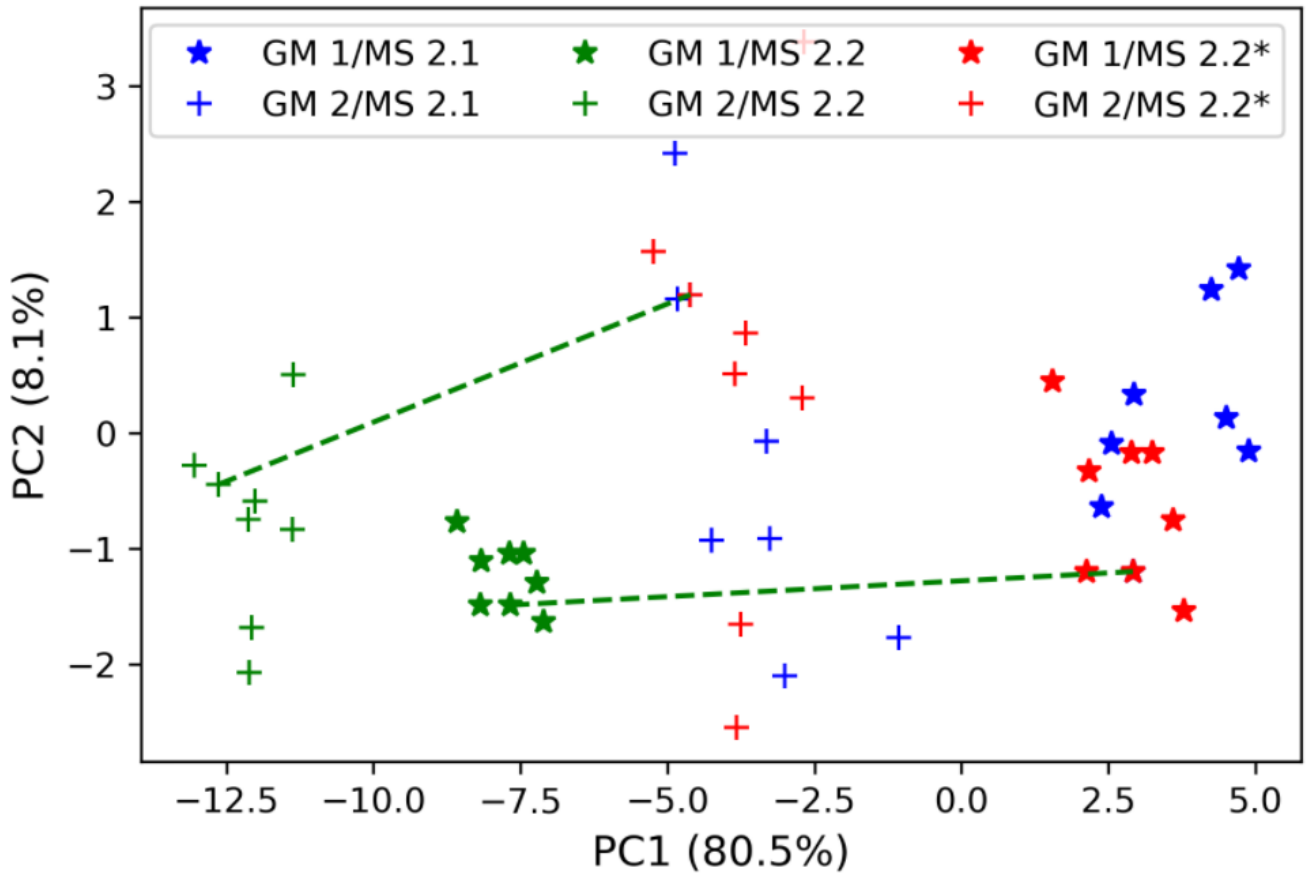

Figure 23. Visual interpretation of the response correction process and mutual arrangement of the samples in the principal component space (the fraction of explained variance for each principal component is shown in parentheses). Mixture samples measured at MS 2.1 are marked in blue, mixture samples at MS 2.2 are marked in green, and samples at MS 2.2 corrected by UDS using two standardization samples are marked in red (for these two samples, the mapping after response correction is shown using the green dotted line). PC1 is the axis of the first principal component, PC2 is the axis of the second principal component.

Conclusions

The results obtained in Chapter 4 demonstrate the feasibility of standardizing the response of one MS for the joint use of one classification model trained on the responses of another MS. The univariate direct standardization methods demonstrated the highest and most acceptable classification accuracy with a minimum number of standardization samples equal to four.

Conclusions

The scheme of online analysis of EB using a system of gas-sensitive metal-oxide sensors, which does not require additional sample preparation, was proposed, created and tested.

Using the developed system, a comparative examination of the group of healthy (53) and LC patients (65) was carried out. Based on the results obtained, an algorithm for the processing of experimental data was developed and validated, allowing the efficient isolation of RL patients with sensitivity (90.5 ± 2.6)%, specificity (98.1 ± 1.5)%, accuracy (94.0 ± 1.6)%, ROC AUC 0.961 ± 0. 018, the predictability of a positive result (98.3 ± 1.3)% and the predictability of a negative result (89.9 ± 2.7)% without regard to external patient condition factors (age, gender, smoking, etc.) and based solely on the multisensory system responses.

Experiments were carried out to investigate the possibility of calibration transfer between two multisensory systems which showed the necessity of using the calibration transfer for one of the systems by standardizing the response. On the model tasks of multiclass classification of mixtures (air – VOC) and binary classification of 2 mixtures with identical qualitative composition but differing in concentrations of individual components not more than 20% the efficiency of using the method of univariate direct standardization for which the maximum accuracy is achieved at the lowest number of samples for standardization equal to four.

List of abbreviation

Acc – accuracy of the test

COPD – chronic obstructive pulmonary disease

DS – direct standardization

DS-L1R – direct standardization with L1 regularization

DS-PLS2 – direct standardization with regression on latent structures 2

EB – exhaled breath

EN – electronic nose

FN – false negative result

FP – false positive result

GC-MS – gas chromatography mass spectrometry

kNN – k nearest neighbor classifier

LC – lung cancer

LDA – linear discriminant analysis

LR – logistic regression

MCC-IMS – multi-capillary column ion mobility spectrometry

MO – metal oxide

MS – multisensory system

MSVM – multiclass support vector machine

NPV – negative predictive value

PC – principal component

PCA – principal component analysis

PCA – principal component analysis

PDS – piecewise direct standardization

PPV – positive predictive value

PTR-MS – proton transfer reaction mass spectrometry

RF – random forest model

RMSE – root mean squared error

ROC – receiver operating characteristic

ROC AUC – area under ROC curve

SAW – surface acoustic waves

sd – standard deviation

Se – sensitivity of the test

SIFT-MS – Selected ion flow tube mass spectrometry

Spe – specificity of the test

SPME – solid-phase microextraction

SVM – support vector machine model

TN – true negative result

TP – true positive result

UDS – univariate direct standardization

UDSwoi – UDS without intercept

VOC – volatile organic compound

References

1. Arseniev A. et al. Combined diagnostics of lung cancer using exhaled breath analysis and sputum cytology // Probl. Oncol. 2020. Vol. 66, № 4. P. 381–384.
2. Ganeev A.A. et al. Analysis of exhaled air for early-stage diagnosis of lung cancer: opportunities and challenges // Russ. Chem. Rev. 2018. Vol. 87, № 9. P. 904–921.
3. Saalberg Y., Wolff M. VOC breath biomarkers in lung cancer // Clin. Chim. Acta. Elsevier B.V., 2016. Vol. 459. P. 5–9.
4. Rattray N.J.W. et al. Taking your breath away: Metabolomics breathes life in to personalized medicine // Trends Biotechnol. Elsevier Ltd, 2014. Vol. 32, № 10. P. 538–548.
5. Xu F., Zou L., Ong C.N. Multiorigination of chromatographic peaks in derivatized GC/MS metabolomics: A confounder that influences metabolic pathway interpretation // J. Proteome Res. 2009. Vol. 8, № 12. P. 5657–5665.
6. Zhou J. et al. Review of recent developments in determining volatile organic compounds in exhaled breath as biomarkers for lung cancer diagnosis // Anal. Chim. Acta. Elsevier Ltd, 2017. Vol. 996. P. 1–9.
7. Atkinson A.J. et al. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework // Clin. Pharmacol. Ther. 2001. Vol. 69, № 3. P. 89–95.
8. Hakim M. et al. Volatile organic compounds of lung cancer and possible biochemical pathways // Chem. Rev. 2012. Vol. 112, № 11. P. 5949–5966.
9. Rocco G. et al. Breathprinting and Early Diagnosis of Lung Cancer // Journal of Thoracic Oncology. International Association for the Study of Lung Cancer, 2018. Vol. 13, № 7. 883–894 p.
10. Amann A. et al. Methodological issues of sample collection and analysis of exhaled breath // Exhaled Biomarkers. 2010.
11. Turner C. Techniques and issues in breath and clinical sample headspace analysis for disease diagnosis // Bioanalysis. 2016. Vol. 8, № 7.
12. Pleil J.D., Lindstrom A.B. Collection of a single alveolar exhaled breath for volatile organic compounds analysis // Am. J. Ind. Med. 1995. Vol. 28, № 1. P. 109–121.
13. Miekisch W. et al. Impact of sampling procedures on the results of breath analysis // J. Breath Res. 2008. Vol. 2, № 2.
14. Miekisch W., Schubert J.K. From highly sophisticated analytical techniques to life-saving diagnostics: Technical developments in breath analysis // TrAC - Trends Anal. Chem. 2006. Vol. 25, № 7. P. 665–673.

15. McCafferty J.B. et al. Effects of breathing pattern and inspired air conditions on breath condensate volume, pH, nitrite, and protein concentrations // Thorax. 2004. Vol. 59, № 8. P. 694–698.
16. Соодаева С.К., Климанов И.А. Нарушения окислительного метаболизма при заболеваниях респираторного тракта и современные подходы к антиоксидантной терапии. 2009. P. 34–37.
17. Horváth I. et al. Exhaled breath condensate: Methodological recommendations and unresolved questions // Eur. Respir. J. 2005. Vol. 26, № 3. P. 523–548.
18. Kubáň P., Foret F. Exhaled breath condensate: Determination of non-volatile compounds and their potential for clinical diagnosis and monitoring. A review // Anal. Chim. Acta. 2013. Vol. 805. P. 1–18.
19. Buszewski B. et al. Human exhaled air analytics: Biomarkers of diseases // Biomed. Chromatogr. 2007. Vol. 21. P. 553–566.
20. Krilaviciute A. et al. Detection of cancer through exhaled breath: A systematic review // Oncotarget. 2015. Vol. 6, № 36. P. 38643–38657.
21. US EPA. Method TO-15: Compendium of methods for the determination of toxic organic compounds in ambient air // EPA Methods. 1999. № January. P. 1–32.
22. Beauchamp J. et al. On the use of Tedlar® bags for breath-gas sampling and analysis // J. Breath Res. 2008. Vol. 2, № 4.
23. Schmekel B., Winquist F., Vikström A. Analysis of breath samples for lung cancer survival // Anal. Chim. Acta. Elsevier B.V., 2014. Vol. 840. P. 82–86.
24. Trabue S.L., Anhalt J.C., Zahn J.A. Bias of Tedlar Bags in the Measurement of Agricultural Odorants // J. Environ. Qual. 2006. Vol. 35, № 5. P. 1668–1677.
25. Mieth M. et al. Multibed Needle Trap Devices for on Site Sampling and Preconcentration of Volatile Breath Biomarkers. 2009. Vol. 81, № 14. P. 5851–5857.
26. Hyšpler R. et al. Determination of isoprene in human expired breath using solid-phase microextraction and gas chromatography-mass spectrometry // J. Chromatogr. B Biomed. Sci. Appl. 2000. Vol. 739, № 1. P. 183–190.
27. Mutlu G.M. et al. Collection and analysis of exhaled breath condensate in humans // Am. J. Respir. Crit. Care Med. 2001. Vol. 164, № 5. P. 731–737.
28. Dyne D., Cocker J., Wilson H.K. A novel device for capturing alveolar breath samples for solvent analysis // J. Automat. Chem. 1997. Vol. 19, № 2. P. 59.
29. Poli D. et al. Exhaled volatile organic compounds in patients with non-small cell lung cancer: Cross sectional and nested short-term follow-up study // Respir. Res. 2005. Vol. 6. P. 1–10.
30. Kusano M., Mendez E., Furton K.G. Development of headspace SPME method for analysis of

volatile organic compounds present in human biological specimens // Anal. Bioanal. Chem. 2011. Vol. 400, № 7. P. 1817–1826.

31. Van Den Velde S. et al. Differences between alveolar air and mouth air // Anal. Chem. 2007. Vol. 79, № 9. P. 3425–3429.
32. Amann A. et al. Applications of breath gas analysis in medicine // Int. J. Mass Spectrom. 2004. Vol. 239, № 2–3. P. 227–233.
33. Alonso M., Castellanos M., Sanchez J.M. Evaluation of potential breath biomarkers for active smoking: Assessment of smoking habits // Anal. Bioanal. Chem. 2010. Vol. 396, № 8. P. 2987–2995.
34. Alonso M. et al. Capillary thermal desorption unit for near real-time analysis of VOCs at sub-trace levels. Application to the analysis of environmental air contamination and breath samples // J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 2009. Vol. 877, № 14–15. P. 1472–1478.
35. Scheepers P.T.J. et al. Determination of exposure to benzene, toluene and xylenes in Turkish primary school children by analysis of breath and by environmental passive sampling // Sci. Total Environ. Elsevier B.V., 2010. Vol. 408, № 20. P. 4863–4870.
36. Gordon S.M. et al. Volatile organic compounds as breath biomarkers for active and passive smoking // Environ. Health Perspect. 2002. Vol. 110, № 7. P. 689–698.
37. Woolfenden E. Sorbent-based sampling methods for volatile and semi-volatile organic compounds in air. Part 1: Sorbent-based air monitoring options // J. Chromatogr. A. Elsevier B.V., 2010. Vol. 1217, № 16. P. 2674–2684.
38. Helmig D. Artifact-free preparation, storage and analysis of solid adsorbent sampling cartridges used in the analysis of volatile organic compounds in air // J. Chromatogr. A. 1996. Vol. 732, № 2. P. 414–417.
39. Calogirou A. et al. Decomposition of terpenes by ozone during sampling on tenax // Anal. Chem. 1996. Vol. 68, № 9. P. 1499–1506.
40. Dewulf J., Van Langenhove H. Anthropogenic volatile organic compounds in ambient air and natural waters: a review on recent developments of analytical methodology, performance and interpretation of field measurements // J. Chromatogr. A. 1999. Vol. 843, № 1–2. P. 163–177.
41. Peng C.Y., Batterman S. Performance evaluation of a sorbent tube sampling method using short path thermal desorption for volatile organic compounds // J. Environ. Monit. 2000. Vol. 2, № 4. P. 313–324.
42. Cao X.L., Nicholas Hewitt C. Build-up of artifacts on adsorbents during storage and its effect on passive sampling and gas chromatography-flame ionization detection of low concentrations of volatile organic compounds in air // J. Chromatogr. A. 1994. Vol. 688, № 1–2. P. 368–374.
43. Materić D. et al. Methods in Plant Foliar Volatile Organic Compounds Research // Appl. Plant

Sci. 2015. Vol. 3, № 12. P. 1500044.

44. Wang C. et al. Noninvasive detection of colorectal cancer by analysis of exhaled breath // Anal. Bioanal. Chem. 2014. Vol. 406, № 19. P. 4757–4763.
45. Wang C., Sahay P. Breath analysis using laser spectroscopic techniques: Breath biomarkers, spectral fingerprints, and detection limits // Sensors. 2009. Vol. 9, № 10. P. 8230–8262.
46. Schwarz K. et al. Breath acetone - Aspects of normal physiology related to age and gender as determined in a PTR-MS study // J. Breath Res. 2009. Vol. 3, № 2.
47. Smith D. et al. Mass spectrometry for real-time quantitative breath analysis // J. Breath Res. 2014. Vol. 8, № 2.
48. Smith D. et al. Quantification of acetaldehyde released by lung cancer cells in vitro using selected ion flow tube mass spectrometry // Rapid Commun. Mass Spectrom. 2003. Vol. 17, № 8. P. 845–850.
49. Rutter A. V. et al. Quantification by SIFT-MS of acetaldehyde released by lung cells in a 3D model // Analyst. 2013. Vol. 138, № 1. P. 91–95.
50. Sulé-Suso J. et al. Quantification of acetaldehyde and carbon dioxide in the headspace of malignant and non-malignant lung cells in vitro by SIFT-MS // Analyst. 2009. Vol. 134, № 12. P. 2419–2425.
51. Baumbach J.I. et al. Significant different volatile biomarker during bronchoscopic ion mobility spectrometry investigation of patients suffering lung carcinoma // Int. J. Ion Mobil. Spectrom. 2011. Vol. 14, № 4. P. 159–166.
52. Lamote K. et al. Detection of malignant pleural mesothelioma in exhaled breath by multicapillary column/ion mobility spectrometry (MCC/IMS) // J. Breath Res. IOP Publishing, 2016. Vol. 10, № 4. P. 46001.
53. Bessa V. et al. Detection of volatile organic compounds (VOCs) in exhaled breath of patients with chronic obstructive pulmonary disease (COPD) by ion mobility spectrometry // Int. J. Ion Mobil. Spectrom. 2011. Vol. 14, № 1. P. 7–13.
54. Arasaradnam R.P. et al. Non-invasive exhaled volatile organic biomarker analysis to detect inflammatory bowel disease (IBD) // Dig. Liver Dis. Editrice Gastroenterologica Italiana, 2016. Vol. 48, № 2. P. 148–153.
55. Wilson A.D. Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath // Metabolites. 2015. Vol. 5, № 1. P. 140–163.
56. Behera B. et al. Electronic nose: A non-invasive technology for breath analysis of diabetes and lung cancer patients // Journal of Breath Research. 2019. Vol. 13, № 2.
57. McWilliams A. et al. Sex and smoking status effects on the early detection of early lung cancer in high-risk smokers using an electronic nose // IEEE Trans. Biomed. Eng. 2015. Vol. 62, № 8.

P. 2044–2054.

58. Chen X. et al. A study of an electronic nose for detection of lung cancer based on a virtual SAW gas sensors array and imaging recognition method // Meas. Sci. Technol. 2005. Vol. 16, № 8. P. 1535–1546.
59. Gasparri R. et al. Volatile signature for the early diagnosis of lung cancer // J. Breath Res. IOP Publishing, 2016. Vol. 10, № 1. P. 16007.
60. Mazzone P.J. et al. Exhaled breath analysis with a colorimetric sensor array for the identification and characterization of lung cancer // J. Thorac. Oncol. International Association for the Study of Lung Cancer, 2012. Vol. 7, № 1. P. 137–142.
61. Shehada N. et al. Silicon Nanowire Sensors Enable Diagnosis of Patients via Exhaled Breath // ACS Nano. 2016. Vol. 10, № 7. P. 7047–7057.
62. Meixner H., Lampe U. Metal oxide sensors // Sensors Actuators, B Chem. 1996. Vol. 33, № 1–3. P. 198–202.
63. Marikutsa A. V. et al. Active sites on the surface of nanocrystalline semiconductor oxides ZnO and SnO 2 and gas sensitivity // Russian Chemical Bulletin. 2017. Vol. 66, № 10.
64. Волькенштейн Ф.Ф. Электронные процессы на поверхности полупроводников при хемосорбции. Наука. Гл. ред. физ.-мат. лит., 1987.
65. Мясников И.А, Сухарев В.Я., Куприянов Л.Ю. З.С.А. Полупроводниковые сенсоры в физико-химических исследованиях. Москва: Наука, 1991. 327 p.
66. Pijolat C. et al. Gas detection for automotive pollution control // Sensors Actuators, B Chem. 1999. Vol. 59, № 2. P. 195–202.
67. Rudnitskaya A. Calibration update and drift correction for electronic noses and tongues // Front. Chem. 2018. Vol. 6, № September.
68. А.В. Ш. Селективное определение газов полупроводниковыми сенсорами. 2005.
69. Baldini C. et al. Electronic nose as a novel method for diagnosing cancer: A systematic review // Biosensors. 2020. Vol. 10, № 8. P. 1–21.
70. Blatt R. et al. Lung cancer identification by an electronic nose based on an array of MOS sensors // IEEE Int. Conf. Neural Networks - Conf. Proc. 2007. P. 1423–1428.
71. Tran V.H. et al. Breath analysis of lung cancer patients using an electronic nose detection system // IEEE Sens. J. 2010. Vol. 10, № 9. P. 1514–1518.
72. Yu K. et al. A portable electronic Nose intended for home healthcare based on a mixed sensor array and multiple desorption methods // Sensor Letters. 2011. Vol. 9, № 2.
73. Wang D. et al. A hybrid electronic noses' system based on MOS-SAW detection units intended for lung cancer diagnosis // J. Innov. Opt. Health Sci. 2012. Vol. 5, № 1. P. 1–7.
74. De Vries R. et al. Integration of electronic nose technology with spirometry: Validation of a

new approach for exhaled breath analysis // J. Breath Res. IOP Publishing, 2015. Vol. 9, № 4. P. 46001.

75. Tan J.L., Yong Z.X., Liam C.K. Using a chemiresistor-based alkane sensor to distinguish exhaled breaths of lung cancer patients from subjects with no lung cancer // J. Thorac. Dis. 2016. Vol. 8, № 10. P. 2772–2783.
76. van Hooren M.R.A. et al. Differentiating head and neck carcinoma from lung carcinoma with an electronic nose: a proof of concept study // Eur. Arch. Oto-Rhino-Laryngology. Springer Berlin Heidelberg, 2016. Vol. 273, № 11. P. 3897–3903.
77. Kort S. et al. Multi-centre prospective study on diagnosing subtypes of lung cancer by exhaled-breath analysis // Lung Cancer. Elsevier Ireland Ltd, 2018. Vol. 125. P. 223–229.
78. van de Goor R. et al. Training and Validating a Portable Electronic Nose for Lung Cancer Screening // J. Thorac. Oncol. International Association for the Study of Lung Cancer, 2018. Vol. 13, № 5. P. 676–681.
79. Marzorati D. et al. A Metal Oxide Gas Sensors Array for Lung Cancer Diagnosis Through Exhaled Breath Analysis // Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. 2019.
80. Родионова О.Е., Померанцев А.Л., Ран Н.Н.С. Хемометрика в аналитической химии [Electronic resource]. 2006.
81. Marco S., Gutierrez-Galvez A. Signal and data processing for machine olfaction and chemical sensing: A review // IEEE Sens. J. 2012. Vol. 12, № 11. P. 3189–3214.
82. Leopold J.H. et al. Comparison of classification methods in breath analysis by electronic nose // J. Breath Res. 2015. Vol. 9, № 4. P. 046002.
83. Wlodzimirow K.A. et al. Exhaled breath analysis with electronic nose technology for detection of acute liver failure in rats // Biosens. Bioelectron. Elsevier, 2014. Vol. 53. P. 129–134.
84. Benedek P. et al. Exhaled biomarker pattern is altered in children with obstructive sleep apnoea syndrome // Int. J. Pediatr. Otorhinolaryngol. 2013. Vol. 77, № 8. P. 1244–1247.
85. Hakim M. et al. Diagnosis of head-and-neck cancer from exhaled breath // Br. J. Cancer. Nature Publishing Group, 2011. Vol. 104, № 10. P. 1649–1655.
86. Mazzone P.J. et al. Diagnosis of lung cancer by the analysis of exhaled breath with a colorimetric sensor array // Thorax. 2007. Vol. 62, № 7. P. 565–568.
87. Breiman L. Random Forests // Mach. Learn. 2001. Vol. 45, № 1. P. 5–32.
88. Crammer K., Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines // J. Mach. Learn. Res. - JMLR. 2002. Vol. 2, № 2. P. 265–292.
89. Rudnitskaya A. et al. Measurements of the effects of wine maceration with oak chips using an electronic tongue // Food Chem. Elsevier Ltd, 2017. Vol. 229. P. 20–27.

90. Pillonel L., Bosset J.O., Tabacchi R. Data transferability between two MS-based electronic noses using processed cheeses and evaporated milk as reference materials // Eur. Food Res. Technol. 2002. Vol. 214, № 2. P. 160–162.
91. Pérez Pavón J.L. et al. Strategies for qualitative and quantitative analyses with mass spectrometry-based electronic noses // TrAC - Trends Anal. Chem. 2006. Vol. 25, № 3. P. 257–266.
92. Deshmukh S. et al. Calibration transfer between electronic nose systems for rapid In situ measurement of pulp and paper industry emissions // Anal. Chim. Acta. 2014. Vol. 841. P. 58–67.
93. Fonollosa J. et al. Evaluation of calibration transfer strategies between Metal Oxide gas sensor arrays // Procedia Eng. Elsevier B.V., 2015. Vol. 120. P. 261–264.
94. Panchuk V. et al. Extending electronic tongue calibration lifetime through mathematical drift correction: Case study of microcystin toxicity analysis in waters // Sensors Actuators, B Chem. Elsevier B.V., 2016. Vol. 237. P. 962–968.
95. Kennard R.W., Stone L.A. Computer Aided Design of Experiments // Technometrics. 1969. Vol. 11, № 1. P. 137–148.
96. Vasiliev A.A. et al. Reducing humidity response of gas sensors for medical applications: Use of spark discharge synthesis of metal oxide nanoparticles // Sensors (Switzerland). 2018. Vol. 18, № 8.
97. Hierlemann A., Gutierrez-Osuna R. Higher-order chemical sensing // Chem. Rev. 2008. Vol. 108. P. 563–613.
98. Malyshev V. V., Pislyakov A. V. Dynamic properties and sensitivity of semiconductor metal-oxide thick-film sensors to various gases in air gaseous medium // Sensors Actuators, B Chem. 2003. Vol. 96, № 1–2. P. 413–434.
99. Alexander Kononov. Breath Analysis [Electronic resource]. 2021. URL: https://github.com/camberbatch/Breath_analysis.
100. Kononov A. et al. Online breath analysis using metal oxide semiconductor sensors (electronic nose) for diagnosis of lung cancer // J. Breath Res. 2020. Vol. 14, № 1.
101. Shapiro A.S.S., Wilk M.B. An Analysis of Variance Test for Normality (Complete Samples) Published by : Biometrika Trust Stable URL : http://www.jstor.org/stable/2333709 // Biometrika. 1965. Vol. 52, № 3/4. P. 591–611.