## Техническое задание

Разработка стратегии обработки спектральных данных смесей растительных масел и примение методов машинного обучения для классификации смеси подсолнечных и рапсовых масел и количественного определения их компонентного состава.

## Исходные данные к работе

Пищевые масла подвержены фальсификации из-за хронического дефицита и колебаний цен. Это важная проблема, поскольку фальсификация может иметь пагубные последствия для здоровья потребителей. В последние годы для аутентификации пищевых продуктов используются передовые методы, такие как ядерный магнитный резонанс (ЯМР), инфракрасная спектроскопия с преобразованием Фурье (FTIR), рамановская спектроскопия и масс-спектрометрия (МС). Эти методы позволяют получить данные высокого разрешения в сочетании с хемометрическими методами, позволяющими различать настоящие и фальсифицированные пищевые масла.

Поскольку не существует методики ГОСТ для определения наличия рапсового масла в подсолнечном масле, актуальной оказалась разработка такой методики на основе ИК спектроскопии и хемометрической обработки данных. В качестве объектов использовались 5 разных подсолнечных масел и 5 рапсовых масел и их смеси в различном соотношении.

## Цель работы

Развитие метода обработки спектральных данных смесей пищевых масел и применения методов машинного обучения для их классификации и определения количественного состава.

## Виды работ и требования к их выполнению

1. Сбор и предобработка спектральных данных различных подсолнечных и рапсовых масел и их попарных смесей.

2. Классификация ИК спектров смесей данных пищевых масел.

3. Поиск алгоритмов, подходящих для количественного определения компонентов в смесях.

4. Анализ и сравнение прогнозных способности предложенных моделей.

## Введение

Растительные пищевые масла являются необходимым элементов питания человека, что делает их неотъемлемой частью пищевой промышленности. В связи с хроническим дефицитом сырья для их производства в индустрии производства масел существует проблема фальсификации, заключающаяся в добавлении посторонних примесей и подмене дорогостоящих высококачественных масел более дешевыми, и даже техническими маслами. Этот факт не только нарушает установленные стандарты производства и реализации пищевой продукции, но и создает угрозу для здоровья потребителей.

В настоящее время для установления случаев подмены масел используются передовые лабораторные методы, такие как спеткроскопия ядерного магнитного резонанса (ЯМР), рамановская спектроскопия, масс-спектроскопия (МС) и инфракрасная спектроскопия с преобразованием Фурье (FTIR). Эти методы позволяют получить данные высокого разрешения в сочетании с хемометрическими методами, позволяющими различать настоящие и фальсифицированные пищевые масла.

Также, возрастает доля научных исследований, посвященных применению методов машинного обучения (МО) в естесственно-научных областях при решение прикладных задач. Наличие данных высокого разрешения, содержащих в себе информацию, позволяющую различать образцы фальсифицированного и настоящего масла, позволяет автоматизировать процесс анализа результатов спектроскопических исследований и выявлять новые закономерности в данных, что также может привносить интерес в исследование.

В связи с отсутствием на данный момент методики ГОСТ для определения наличия рапсового масла в подсолнечном, актуальной оказалась разработка такой методики на основе FTIR спектроскопии с применением алгоритмов машинного обучения в качестве хемометрических методов для классификации образцов и выявления тенденций при анализе спектральных данных, что является целью данной работы.

## ГЛАВА 1. ОБЗОР И АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

### Методики идентификации растительных масел

Контроль качества растительных масел может осуществляться различными методами. Одним из возможных путей идентификации фальсификаций являются физико-химические показатели масел, такие как йодное число, перекисное число, кислотное число и т.д. [2]. Однако в современной промышленности требуется наличие более точных методов, позволяющих оценить более широкий диапазон параметров растительных масел для определения их качества.

Одним из таких способов является разделение смесей масел при помощи хроматографических методов. Суть данных методов заключается в различном распределении компонентов смеси между фазами установки в зависимости от их физико-химических свойств. Растительные масла являются многокомпонентными смесями, однако для каждого вида масла характерно преобладание остатков определенного набора жирных кислот, и в случае фальсификации масел добавлением сторонних видов, в результате проведения хроматографии можно установить наличие посторонних примесей путем анализа их жирнокислотного состава.

Так, в работе [1] для определения качества растительных масел использовались методы газовой хроматографии с масс-спектроскопическим детектором. В исследовании [3] описывается проблема определения качества облепихового масла методов высокоэффективной жидкостной хроматографии (ВЭЖХ), заключающаяся в проверке гипотезы о том, что по результатам проведения эксперимента можно установить наличие триглицеридов, не характерных для масла облепихи, например, подсолнечного масла. Было показано, что триглицериды, образованные остатками разных жирных кислот элюируются с различной скоростью, что позволяет разделять их по хроматограммам, полученным в результате проведения ВЖЭХ [3].

Еще одним способом получения данных высокого качества при анализе качества растительных пищевых масел является семейство спектральных методов, таких как флюоресцентная спектроскопия, спектроскопия в ультрафиолетовом диапазоне (УФ спектроскопия), флуоресцентная спектроскопия [3], инфракрасная спектроскопия с преобразованием Фурье, ЯМР спектроспия.

Метод ЯМР является высокоэффективным способом анализа качества масел. В работе [2] было показано, что проведение анализа спектров ЯМР значительно упрощает процесс анализа состава растительных масел в связи с отличающимися химическими сдвигами углеродных атомов различных жирных кислот, что результирует в наличие отдельных сигналов в спектре, позволяя определять жирнокислотный состав смесей и по наличию посторонних сигналов в сравнении с эталонным спектром образца рассматриваемого масла, позволяющих указать на наличие постороннего вида масла в составе смеси.

Также находит применение в исследовании качества растительных масел метод спектроскопии в ультрафиолетовом диапазоне (УФ-спектроскопия). Данный способ позволяет получать информацию об изменениях в окислительном профиле масел, позволяя выявлять нарушения в качестве по окислительным коэффициентам, свидетельствующим об окислительном состоянии соединений [5]. При этом можно получить данные о качестве масел, изменениях в их качестве при хранении и в процессе промышленной переработки при производстве. Для каждого вида масла характерен свой жирнокислотный состав, влияющий на уровень насыщения связей и способность к окислению, поэтому информация об окислительном состоянии вещества, полученная методом УФ-спектроскопии, обладает ценностью при проведении контроля качества масел как продуктов питания. Изменения в качестве растительных жиров, результирующие в окислительные процессы, происходящие при нарушениях условий хранения или производства, снижают ценность масел, так как высококачественные дорогостоящие масла обладают полезностью в том числе и из-за того, что могут выступать в роли источника антиоксидантов в питании человека. В свою очередь, изменения в окислительном состоянии масел влияют на эту способность, снижая уровень полезности растительных жиров.

Еще одним спектральным методом, позволяющим оценивать качество растительных масел, является масс-спектроскопия (МС). В работе [5] исследовалась фальсификация оливкового масла холодного отжима соевым маслом. Было показано, что МС может быть использована для решения данной задачи, так как данный способ позволяет выявлять различия в триглицеридном составе. В зависимости от вида масла, для него характерно преобладание того или иного набора жирных кислот, обладающих различной массой и зарядом. Получаемые в МС спектры позволяют оценить представленность компонентов в веществе, и путем сравнения полученных спектров с эталонными может быть установлен случай фальсификации масел путем смешивания более дорогих растительных жиров с более дешевыми.

Исследование [4] описывает методику применения FTIR спектроскопии для анализа смесей ряда растительных масел с подсолнечным. Авторы указывают, что при визуальном анализе полного спектра, полученного при проведении измерений, сложно найти существенные отличия между образцами. Однако при детальном рассмотрении характеристических регионов спектров имеются различия в интенсивности пиков поглощения. Такими регионами являются области 3100-2800 см-1, 1800-1600 см-1 и 1600-1390 см-1 [4], интенсивность поглощения в которых говорит о доле насыщенных и ненасыщенных остатков жирных кислот, а также об общем уровне насыщенности связей в триглицеридах, что позволяет проводить различие между чистыми образцами масел определенного вида, а также смесями различных видов масел, что является признаком недоброкачественности производителей и свидетельствует о наличии акта фальсификации продукции. Было показано, что в зависимости от количественного содержания примеси растительного масла в образцах интенсивность пиков поглощения изменяется [4], что говорит о возможности использования метода FTIR спектроскопии для определения наличия посторонних примесей в образцах реализуемых в продаже растительных масел.

### Машинное обучение в естественных науках

В настоящее время МО находит широкое применение в естественнонаучных областях. Модели МО используют для предсказания физических свойств материалов [6], моделирования и прогнозирования структуры белков [11], дизайне и разработке лекарств [16, 17], оценке токсичности наноматериалов [15].

Успех в стремительное развитие применения методов МО в науке обусловлен наличием больших объемов данных, получаемых исследователями в доменных областях, который продолжает пополняться с каждым днем [14].

Применение алгоритмов МО позволяет

Однако, на данный момент существует проблема репрезентации молекул химических веществ для подачи данных в модели МО [12]. Данная проблема важна, так как производительность и эффективность работы алгоритмов тесно связана с представлением и форматом данных, которые подаются на вход модели. Одним из самых популярных вариантом представления химических соединений является запись формул в формате SMILES [13]. Этот формат удобен для чтения людьми, но вычислительные машины справляются с ним не так хорошо, и несмотря на развитие моделей обработки естественного языка (NLP – natural language processing), эффективность алгоритмов МО при обработке данных в формате SMILES остается невысокой. В связи с этим при решении большого количества задач исследователи прибегают к использованию базовых физико-химических параметров молекул, а также используют методы векторизации строк формата SMILES при помощи полносвязных нейронных сетей. Описанная проблема вызывает научный интерес, так как разработка унифицированного метода представления молекул может совершить прорыв в области машинного обучения в химии, открывая новые возможности для исследования и позволяя улучшить производительность уже существующих алгоритмов.

## ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

### Машинное обучение

Машинное обучение - совокупность методов искусственного интеллекта (ИИ), в основе которых лежат математические методы статистики и теории вероятностей.

Одной из основных ценностей МО является возможность работать с большим данных, которые стремительно накапливаются в современном мире, проводить их анализ, выявлять новые закономерности, которые можно увидеть лишь в масштабах выборки размером в тысячи, десятки и сотни тысяч образцов. Такие объемы данных неподвластны восприятию человеком, однако могут быть обработаны и интерпретированы при помощи современных инструментов машинного обучения.

Основной задачей МО выступает выявление определенного правила или закономерности на основании больших данных. Такое правило описывается функцией, параметры которой алгоритм подбирает в процессе обучения на выбранном исследователем наборе данных. Методы МО в общем случае можно классифицировать на три вида: обучение с учителем, обучение без учителя, и обучение с подкреплением [6].

Обучение с учителем

Суть обучения с учителем заключается в том, что алгоритму дается набор входных параметров и набор соответствующих им ожидаемых результатов. При этом задачей является подбор оптимальной функции, позволяющей получать ожидаемые результаты с наименьшей долей ошибки. Ошибкой в методах обучения с учителем является метрика, позволяющая оценить степень различия между предсказанными результатами и ожидаемыми значениями целевой переменной из исходного набора данных.

При решении задач классификации оценить качество работы алгоритма  можно с помощью следующих методов. Первым способом, подходящим для задачи бинарной классификации, то есть классификации, где выбор состоит из двух возможных классов, является построение матрицы ошибок, содержащей информацию о доле ложно-положительных (FP - false positive), ложно-отрицательных (FN - false negative), истинно-отрицательных (TN - true negative) и истинно-положительных (TP - true positive) предсказаний. В группу ложных предсказаний относятся результаты, когда конкретный образец ложно определен к  тому или иному классу. В группе истинных предсказаний относятся случаи верного определения класса, к которому принадлежит образец. Данный метод оценки работы модели МО особенно важен для некоторых задач, например, для решения задач в медицине. Например, если алгоритм определения возможных заболеваний по результатам медицинских анализов ложно определит больного человека в группу здоровых, то это более критично, чем определение здорового человека в больные и рекомендация пройти более комплексное обследование и  дополнительную консультацию у специалиста. Из вышеописанной метрики оценки истинных и ложных предсказаний вытекает еще один метод оценки точности бинарной классификации - кривая ROC (receiver operating characteristic curve), которая отражает соотношение между специфичность и чувствительностью метода определения класса, что является критичным при проведении анализа, например, в области медицины [9]. Представим ситуацию, при которой перед исследователем стоит задача классификации больных и здоровых людей по результатам медицинских анализов. В таком случае, чувствительность, или true positive rate (TPR), отражает условную вероятность правильного определения заболевшего человека [9], и рассчитывается по формуле (1)

TPR = TPTP + FN (1),

где TP - количество истинно-положительных предсказаний, FN - количество ложно отрицательных результатов. Специфичность, или true negative rate (TNR), отражает условную вероятность верного определения здорового человека, и рассчитывается по формуле (2):

TNR = TNTN + FP (2),

где TN - количество истинно-отрицательных предсказаний, FP - количество ложно положительных предсказаний. В формулах (1), (2) знаменатель дроби представляет общее количество образцов, соответствующих либо классу больных, либо классу здоровых. Пример получаемого графика метрики показан на рисунке 1.

A graph of a number of black lines

Description automatically generated with medium confidence

Рис.1 Три эмпирические ROC-кривые [9].

Площадь под графиком (AUC - area under curve) ROC кривой позволяет оценить соотношение между специфичность и чувствительностью метода. показывает, что метод обладает идеальной способностью разделять больных и здоровых людей [9]. Приведенная аналогия определения здоровых и больных людей может быть представлена как пример задачи бинарной классификации. В связи с этим, метод оценки площади под графиком ROC кривой находит применение в оценке работы предсказательных алгоритмов машинного обучения.

Еще одной метрикой является общая точность классификации, выражаемая в долях или процентах и показывающая соотношение образцов, для которых был предсказан класс, к которому они относятся в исходном наборе данных. Этот способ оценки является актуальным как при решении задач бинарной классификации, так и в случае необходимости классификации образцов, которые могут относиться к множеству классов, то есть в таком случае, если набор данных описывает выборку с количеством классов больше двух.

В случае решения задач регрессии для оценки точности алгоритмов МО также используется несколько метрик:

1. Средняя абсолютная ошибка MAE – mean absolute error
2. Среднеквадратичная ошибка MSE – mean squared error
3. Коэффициент детерминации R2

где

В приведенных формулах *yi* - обозреваемое значение целевой переменной, *fi* - предсказанное значение, *n* - размер выборки. Формула (6) представляет метод расчета среднего значения целевой переменной во всей представленной выборке данных.

MAE и MSE позволяют оценить величину численного отклонения прогнозируемой величины переменной от обозреваемой, и поэтому по ним не всегда можно полностью оценить, насколько точно выбранный алгоритм решает регрессионную задачу. Данные величины могут быть использованы в качестве параметров оценки доверительных интервалов, в пределах которых в предсказании присутствует разброс. Чем меньше такой интервал, тем выше точность. В то же время, коэффициент детерминации позволяет оценить, насколько хорошо модель МО аппроксимирует исходные данные. показывает, что модель осуществляет предсказание с идеальной точностью.

Понимание методов, по которым происходит оценка эффективности и качества работы моделей обучения с учителем, дает возможность рассмотреть некоторые из наиболее распространенных алгоритмов классического МО, существующих на данный момент.  Первый тип алгоритмов - это линейные алгоритмы, основанные на методе наименьших квадратов (МНК), в которых осуществляется подбор линейного уравнения, позволяющего по набору входных параметров и коэффициентов к ним, подбираемых в процессе обучения, получать результаты, наиболее близкие к значению целевой переменной на обучающем наборе данных. Линейные модели регрессии работают методом непосредственного предсказания целевой переменной. В случае классификации линейным алгоритмом является логистическая регрессия, целью которой является подобрать уравнение такой прямой, у которой сумма квадратов расстояний от каждой обозреваемой точки до нее минимальна, то есть построить прямую, которая стремится наиболее точно разделить два класса. Примерами линейных алгоритмов классификации и регрессии являются  LinearRegression [11] - простая линейная регрессия по методу МНК, Ridge [12] - регрессия накладывающая ограничения на коэффициенты при построении уравнения, Lasso [13] - метод, основной тенденцией которого является сокращение количества ненулевых коэффициентов, что позволяет в некоторых случаях эффективно производить отбор параметров, влияющих на значение прогнозируемой величины , LogisticRegression [14] - алгоритм логистической регрессии.

К следующей группе относятся алгоритмы, основанные на построении деревьев принятия решения. Дерево принятия решения - это граф, в узлах которого находятся условия, в зависимости от которых так называем процесс принятия решения переходит в тот или другой узел. Условия в узлах формируются на основании обучающего набора  данных. Существуют алгоритмы как на базе одного дерева [15, 16], так и анасамблевые модели, которые сочетают в себе множество деревьев принятия решения, образуя случайный лес [17, 18]. Стратегия обучения алгоритмов случайного леса заключается в обучении множества деревьев принятия решения на различных подвыборках из общей обучающей выборки, и усреднении их предсказаний для повышения общей точности.

Еще одним распространенным методом является метод опорных векторов (SVM - support vector machines) [18]. Этот тип алгоритмов работает методом построения гиперплоскости - плоскости в многоразмерном пространстве параметров обучающей выборки, которая позволяет разделить образцы и осуществить эффективную классификацию.

Обучение без учителя

Алгоритмы обучения без учителя как правило используются с целью анализа имеющихся данных для выявления тенденций и шаблонов, характерных для образцов в исследуемой выборке. С этой целью используются алгоритмы кластеризации, например, метод K-средних [20], который осуществляет группировку набора данных в количество кластеров K, которое задается исследователем. В каждом кластере имеется центроида -  значение, являющееся средним для всех образцов внутри кластера. Алгоритм подбирает центроиды таким образом, чтобы уменьшить сумму квадратов Евклидовых расстояний (7) внутри кластера [20].

S2 = i=1n(xi-j)2 (7)

где *x* - координата *i-*ой точки, - координата центроиды *j-*го кластера.

Однако, оценка Евклидовых расстояний может терять свою эффективность при работе с данными высокой размерности. В данном случае, могут применяться методы снижения размерности, такие как анализ главных компонент (PCA - principal components analysis) [21], multidimensional scaling (MDS) [22], t--Distributed Stochastic Neighbor Embedding (t-SNE) [23].

Список использованных сокращений и условных обозначений

FTIR - инфракрасная спектроскопия с преобразованием Фурье

ИК - инфракрасная спектроскопия

МО - машинное обучение

ЯМР - спектроскопия ядерного магнитного резонанса

УФ спектроскопия - спектроскопия в ультрафиолетовом диапазоне

SMILES – simplified molecular input line entry system

ROC - receiver operating characteristic curve

MAE - средняя абсолютная ошибка

MSE - среднеквадратичная ошибка

МНК - метод наименьших квадратов

PCA - анализ главных компонент

MDS - multidimensional scaling

t-SNE - t--Distributed Stochastic Neighbor Embedding

## Список использованных источников

1. Зверева, И. С., Денисов, И. С., Зубчонок, Н. В., & Ермолаева, Т. Н. (2019). ИЗУЧЕНИЕ КАЧЕСТВА РАСТИТЕЛЬНЫХ МАСЕЛ МЕТОДАМИ ИК-СПЕКТРОМЕТРИИ И ГАЗОВОЙ ХРОМАТОГРАФИИ С МАСС-СПЕКТРОМЕТРИЧЕСКИМ ДЕТЕКТОРОМ. In *XXIX Российская молодежная научная конференция «Проблемы теоретической и экспериментальной химии».—Екатеринбург, 2019* (pp. 94-94). Издательство Уральского университета.
2. Ильина, Г. Г., Ламоткин, С. А., Колногоров, К. П., & Скаковский, Е. Д. (2014). Идентификация состава растительных масел хроматографическими и спектральными методами. *Труды БГТУ. Серия 2: Химические технологии, биотехнология, геоэкология*, (4 (168)), 207-210.
3. Дейнека, В. И., Дейнека, Л. А., & Сорокопудов, В. Н. (2009). Обращенно-фазовая ВЭЖХ в анализе растительных масел. Метод контроля подлинности и установления фальсификации облепихового масла. *Химико-фармацевтический журнал*, *43*(1), 33-36.
4. Alexa, E., Dragomirescu, A., Pop, G., Jianu, C., & Dragos, D. (2009). The use of FT-IR spectroscopy in the identification of vegetable oils adulteration. *J. Food Agric. Environ*, *7*(2), 20-24.
5. da Silveira, R., Vágula, J. M., de Lima Figueiredo, I., Claus, T., Galuch, M. B., Junior, O. O. S., & Visentainer, J. V. (2017). Rapid methodology via mass spectrometry to quantify addition of soybean oil in extra virgin olive oil: A comparison with traditional methods adopted by food industry to identify fraud. *Food Research International*, *102*, 43-50.
6. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547-555.
7. Kireeva, N. et al. Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison. *Mol. Inform*. 31, 301–312 (2012).
8. Fourches, D., Muratov, E. & Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model*. 50, 1189–1204 (2010).
9. Kumar, R., & Indrayan, A. (2011). Receiver operating characteristic (ROC) curve for medical researchers. *Indian pediatrics*, *48*, 277-287.
10. Pedregosa *et al.*, Scikit-learn: Machine Learning in Python JMLR 12, pp. 2825-2830, 2011.
11. Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). <https://doi.org/10.1038/s41586-021-03819-2>
12. Townsend, J., Micucci, C.P., Hymel, J.H. *et al.* Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nat Commun* **11**, 3230 (2020). <https://doi.org/10.1038/s41467-020-17035-5>
13. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36.
14. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., ... & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, *18*(6), 463-477.
15. Shirokii, N., Din, Y., Petrov, I., Seregin, Y., Sirotenko, S., Razlivina, J., ... & Vinogradov, V. (2023). Quantitative prediction of inorganic nanomaterial cellular toxicity via machine learning. *Small*, *19*(19), 2207106.
16. Staszak, M., Staszak, K., Wieszczycka, K., Bajek, A., Roszkowski, K., & Tylkowski, B. (2022). Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *12*(2), e1568.
17. Priya, S., Tripathi, G., Singh, D. B., Jain, P., & Kumar, A. (2022). Machine learning approaches and their applications in drug discovery and design. *Chemical Biology & Drug Design*, *100*(1), 136-153.
18. Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J. L., Sidhu, S. S., ... & Kim, P. M. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome medicine*, *6*, 1-18.