



Научная статья

МЕТОД RANDOM FOREST В ЗАДАЧАХ ИНТЕРПРЕТАЦИИ РЕЗУЛЬТАТОВ БИОЛЮМИНЕСЦЕНТНОГО АНАЛИЗА СЛЮНЫ ПРИ ПЕРСОНАЛИЗИРОВАННОЙ ДИАГНОСТИКЕ

Г.В. Жукова¹, П.А. Мартышук¹, Е.Р. Афер¹, А.Н. Шуваев¹, Н.А. Розанова²,
Д.В. Сергеев², В.А. Кратасюк^{1,3}

¹Сибирский федеральный университет, Российская Федерация, 660041, г. Красноярск, Свободный пр., 79

²Научный центр неврологии, Российская Федерация, 125367, г. Москва, Волоколамское шоссе, 80

³Институт биофизики СО РАН, обособленное подразделение ФНЦ «Красноярский научный центр Сибирского отделения Российской академии наук», Российская Федерация, 660036, г. Красноярск, Академгородок, 50, стр. 50

Развитие направления персонифицированной медицины и биотехнологии напрямую связано с получением данных, в значительной степени зависящих от индивидуальных характеристик исследуемых. При этом используемые в традиционной медицине допустимые интервалы исследуемых параметров не всегда позволяют адекватно охарактеризовать состояние пациента. Возникает необходимость поиска таких методов обработки результатов, которые позволяют учитывать разные индивидуальные особенности организма испытуемых и образа его жизни.

Определены возможности применения анализа биомедицинских данных Random Forest для корректной интерпретации результатов персонифицированных диагностических тестов на примере биолюминесцентного тестирования, позволяющего оценить влияние разных характеристик испытуемых и условий их жизни. Метод позволяет минимизировать риски неверного диагноза и корректировать схему мониторинга исследуемых.

В данной работе используются результаты диагностики трудовой нагрузки работников железнодорожного транспорта, полученные биолюминесцентным методом. Оценку состояния организма проводят, изучая влияние слюны испытуемых на интенсивность свечения биферментной системы: НАДН: ФМН-оксидоредуктаза + люцифераза. Такой анализ является интегральным, реагирует на множество факторов, каждый из которых потенциально может повлиять на результат. Оценку эффективности разных методов обработки данных проводили на примере группы диспетчеров Красноярского филиала АО «РЖД». Для анализа были использованы как статистические методы, так и алгоритм машинного обучения Random Forest.

В результате было показано, что для оценки значимости некоторых биохимических показателей слюны для прогнозирования состояния организма работников железнодорожного транспорта целесообразно использовать метод Random Forest, который позволяет определить наиболее значимые факторы и построить графики частичного влияния факторов на целевую переменную. Данное исследование позволяет оптимизировать систему диагностики организма человека при использовании интегрального биолюминесцентного анализа. Метод Random Forest может быть составной частью персонифицированного биолюминесцентного биосенсора для оценки влияния стрессовых и трудовых нагрузок на организм человека.

Ключевые слова: персонифицированная диагностика, машинное обучение, анализ данных, многофакторный анализ, слюна, биолюминесценция, биосенсор, сигнальные системы.

© Жукова Г.В., Мартышук П.А., Афер Е.Р., Шуваев А.Н., Розанова Н.А., Сергеев Д.В., Кратасюк В.А., 2025

Жукова Галина Викторовна – аспирант Института фундаментальной биологии и биотехнологии, младший научный сотрудник лаборатории биолюминесцентных биотехнологий (e-mail: gvivanova@sfu-kras.ru; тел.: 8 (913) 047-88-37; ORCID: <https://orcid.org/0000-0002-8646-1224>).

Мартышук Полина Александровна – студент Института фундаментальной биологии и биотехнологии (e-mail: p.martishuk@yandex.ru; тел.: 8 (904) 896-97-34).

Афер Егор Романович – студент Института фундаментальной биологии и биотехнологии (e-mail: egorafer@yandex.ru; тел.: 8 (902) 957-64-89).

Шуваев Андрей Николаевич – кандидат физико-математических наук, заведующий базовой кафедрой медико-биологических систем и комплексов Института фундаментальной биологии и биотехнологии (e-mail: ashuvaev@sfu-kras.ru; тел.: 8 (950) 997-77-92; ORCID: <https://orcid.org/0000-0002-3887-1413>).

Розанова Наталья Александровна – лаборант-исследователь, аспирант лаборатории нейробиологии и тканевой инженерии Института мозга (e-mail: nataliarozanova@gmail.com; тел.: 8 (968) 030-47-28; ORCID: <https://orcid.org/0000-0001-9619-4679>).

Сергеев Дмитрий Владимирович – кандидат медицинских наук, старший научный сотрудник, врач-невролог (e-mail: sergeev@neurology.ru; тел.: 8 (926) 860-08-97; ORCID: <https://orcid.org/0000-0002-9130-1292>).

Кратасюк Валентина Александровна – доктор биологических наук, профессор, заведующий кафедрой биофизики Института фундаментальной биологии и биотехнологии; ведущий научный сотрудник (e-mail: valkrat@mail.ru; ORCID: <https://orcid.org/0000-0001-6764-5231>).

В настоящее время персонифицированный подход становится все более популярным в медицине, в связи с чем увеличивается объем полученных при диагностике индивидуальных данных, на которые влияет множество факторов: возраст, образ жизни, условия труда, стрессовые состояния. Допустимые интервалы исследуемых параметров часто не позволяют адекватно охарактеризовать состояние пациента и анализировать риски здоровью с учетом его индивидуальных особенностей. Перспективной биологической жидкостью, которая может использоваться для решения данной проблемы, является слюна. Ее применение в диагностических целях имеет ряд преимуществ: неинвазивный метод сбора, зависимость состава и количества слюны от патологических и физиологических состояний организма [1–3]. Эти проблемы не решены для нового биolumинесцентного неинвазивного биотеста, основанного на оценке влияния слюны на биферментную систему: NAD (P) H: FMN-оксидоредуктаза + люцифераза и зависимости состава слюны от патологических и физиологических состояний организма [4]. Множество факторов влияют на результаты анализа [5], что делает необходимым найти метод для оценки их вклада в биомедицинские данные с учетом разнообразия персонифицированных характеристик испытуемых, найти возможность усреднения данных, выделения критериев нормального состояния и отклонения от нормы.

Существует три основных группы анализа биомедицинских данных: статистические методы, кластеризация и машинное обучение¹ [6–11]. При этом статистическая компонента может быть достаточно сильно задействована в алгоритмах машинного обучения. Статистический анализ обычно оперирует небольшими по объему наборами данных и направлен на понимание взаимосвязей между переменными и проверку гипотез. Машинное обучение более активно применяется в биотехнологиях, так как позволяет автоматизировать различные процессы, обрабатывать более крупные и сложные наборы данных, фокусируется на точности прогнозирования и классификации в задачах без явной переменной-результата, что дает возможность обрабатывать самые разные типы и структуры данных без строгих предположений [6]. Статистический анализ начинается с гипотезы о главном факторе, характеризующем состояние организма, и использует результаты для ее проверки. При этом анализ опирается на за-

ранее определенные модели о распределении данных (например, нормальном распределении)².

Для анализа слюны корреляционный анализ принято считать базовым статистическим методом [7]. В частности, вычисление корреляции Спирмена позволяет оценить силу и направление связи между двумя ранжированными переменными и понять, насколько хорошо эта связь описывается с помощью монотонной функции. Для этого не нужны предположения о характере распределений признаков из-за непараметричности этого теста. Также часто при анализе слюны используется непараметрический тест Манна – Уитни для выявления достоверных статистических различий между двумя независимыми группами по уровню какого-либо признака [7, 10], даже для малых выборок³. Данные методы позволяют найти взаимосвязи между параметрами и провести первичный анализ набора данных. Однако опыт применения базовых статистических методов не всегда позволяет выявить ключевые факторы, влияющие на целевой показатель слюны.

Другим решением проблемы недостаточности арсенала статистического анализа в интерпретации данных по слюне могут быть алгоритмы машинного обучения, которые используются для обнаружения заболеваний различных систем организма [12–18]. Исходя из специфики задачи (поиск значимых факторов, влияющих на индикаторные показатели слюны), можно выбрать алгоритмы, находящие закономерности, по ранее размеченным наборам данных. Метки данных дают модели обратную связь, позволяя скорректировать результат и повысить точность прогноза. Для анализа данных слюны может быть использовано семейство алгоритмов, основанных на построении деревьев решений, таких как CART, Random Forest и бустинги [13, 15, 16, 19]. Метод Random Forest является одним из самых популярных методов для решения задач неглубокого машинного обучения³. Помимо определения класса объекта и нахождения точного значения целевой переменной, алгоритм Random Forest позволяет предсказывать значимость каждого параметра модели для принятия решения, учитывая его уровень относительно вершины дерева, частоту попадания этого параметра в решающий узел, количество объектов, правильно классифицированных с его помощью. Наиболее значимыми являются параметры, находящиеся ближе к вершине дерева [20]. Алгоритм случайного леса создает график частичной зависимости (PDB) для каждого параметра модели. Для рег-

¹ Макарова Н.В. Статистический анализ медико-биологических данных с использованием пакетов статистических программ Statistica, SPSS, NPSS, SYSTAT: методическое пособие / под ред. проф. С.С. Алексанина. – СПб.: Полиграфический центр Санкт-Петербургского ун-та ГПС МЧС России, 2012. – 179 с.; Шорохова И.С., Кисляк Н.В., Мариев О.С. Статистические методы анализа: учебное пособие. – 2-е изд. – М.: ФЛИНТА, 2017. – 300 с.

² Макарова Н.В. Указ. соч.

³ Жерон О. Прикладное машинное обучение с помощью Scikit-Learn и Tensorflow: концепции, инструменты и техники для создания интеллектуальных систем / пер. с англ. Ю.Н. Артеменко. – СПб.: Диалектика-Вильямс, 2020. – 1040 с.

рессионной задачи график показывает усредненные отношения между признаком и целевым значением. Для классификации строится усредненная зависимость между вероятностью верного определения класса объекта и признаком, используемым при построении модели⁴.

Цель исследования – понять, можно ли метод анализа данных Random Forest использовать для корректной интерпретации результатов персонифицированных данных в биолюминесцентном биосенсоре для оценки влияния стрессовых и трудовых нагрузок на организм человека.

Материалы и методы. Для оценки эффективности программы обработки были использованы результаты диагностики, полученные при исследовании группы диспетчеров Красноярского филиала ОАО «РЖД» ($n = 43$). Исследования проводились ежедневно в течение месяца согласно рабочему графику в 2022 и 2023 гг. Для расширения выборки был применен ресэмплинг, при этом контроль за переобучением осуществлялся путем разделения на обучающую и тестовую выборки. Исследуемые параметры были получены при анализе слюны, собранной два раза в сутки: до и после смены. Были оценены такие показатели слюны, как pH, концентрация лактата, остаточное свечение биферментной системы светящихся бактерий: NAD (P) H: FMN-оксидоредуктаза + люцифераза (ЛИ, %) [21], а также биохимические показатели крови и другие результаты профосмотра сотрудников, выполненного на базе ЧУЗ «Клиническая больница “РЖД-медицина”», данные самоанкетирования, включающие субъективную оценку стресса и работоспособности до и после смены, информацию о режиме дня, потреблении пищи, лекарственных средств, энергетических напитков и табака.

Все исследования проведены в соответствии с принципами биомедицинской этики и одобрены локальным этическим комитетом Сибирского федерального университета (г. Красноярск), протокол № 5 от 11.11.2019).

Каждый участник исследования предоставил добровольное письменное информированное согласие, подписанное им после разъяснения ему потенциальных рисков и преимуществ, а также характера предстоящего исследования.

Машинное обучение на задачах классификации и регрессии методом Random Forest выполнено в Python 3 при помощи библиотеки scikit-learning. Важность факторов в этой библиотеке вычисляется по улучшению чистоты коэффициента Джини после разделения по каждому из этих факторов при классификации либо по уменьшению дисперсии прогноза по данному фактору для регрессии. В работе использовалась возможность библиотеки scikit-learning визуализации частичного влияния каждого

фактора на целевую переменную. Статистический анализ выполнен в программе R с использованием базовых библиотек и библиотеки ggplot2. Статистически значимым принималось отличие при p -значениях, меньше либо равных 0,05.

Результаты и их обсуждение. Размер базы биомедицинских данных после предобработки составил 310 строк. Для увеличения базы данных был проведен ресэмплинг, позволивший увеличить количество строк в 4 раза путем формирования новых строк случайным выбором из имеющихся. Перед работой данные разделили на тренировочную выборку (80 % от исходной выборки) и тестовую выборку (20 % исходной выборки). Для дальнейшего анализа были использованы переменные, представленные в табл. 1.

Таблица 1

Факторы, использованные для проверки эффективности программы обработки

Название поля	Размерность	Диапазон
Концентрация лактата в слюне	ммоль/л	0,2–80,7
Уровень глюкозы в крови	ммоль/л	3,2–9,8
Уровень холестерина общий	моль/л	3,4–6,8
pH	pH	3–8
Время забора пробы	До смены (утром) – 1; после смены (вечер) – 0	0,1
Уровень стресса (отношение к работе)	Очень волнуюсь – 3; волнуюсь – 2; спокоен – 1; нет волнения – 0	0–3
Интервал между курением и забором пробы	Минуты	0–120
Пол	Женский – 1; мужской – 0	0,1
Курение	Курил – 1; не курил – 0	0,1
Время суток (утро / вечер)	Утро – 1; вечер – 0	0,1
ЛИ (остаточное свечение)	%	1,08–142,50

В качестве предобработки для выявления наиболее перспективных факторов для включения в машинное обучение методом статистической обработки были оценены коэффициенты корреляции между целевыми параметрами из табл. 1.

Корреляционный анализ показал значимые отрицательные связи целевой переменной ЛИ с отношением к работе испытуемых по шкале от 1 до 5 и временем, прошедшим после курения, для женской группы испытуемых. Выявлены значимые положительные связи целевой переменной с уровнем глюкозы в крови (в женской группе) и с уровнем холестерина (в мужской группе). Корреляционные коэффициенты, представленные в табл. 2, значимы на p -уровне ($< 0,05$).

⁴ Jain A., Fandango A., Kapoor A. Tensorflow Machine Learning Projects. – Packt Publishing, 2018. – 322 p.

Таблица 2

Корреляция факторов с целевой переменной (ЛИ)

Факторы	Коэффициент корреляции
Отношение к работе (женщины)	-0,618
Время между курением и забором пробы (женщины)	-0,699
Глюкоза, ммоль/л (женщины)	0,446
Холестерин общий, моль/л (мужчины)	0,671

В качестве перспективных для анализа алгоритмов был выбран ансамблевый алгоритм машинного обучения, основанный на построении деревьев решений, метод случайного леса. Решали два типа задач: классификационная (с помощью Random Forest Classifier библиотеки scikit-learn), в которой алгоритм прогнозировал принадлежность к группам высокого, низкого или среднего значения ЛИ, и регрессионная (с помощью Random Forest Regressor библиотеки scikit-learn) с прогнозом точного значения ЛИ по задаваемым факторам. Были выявлены наиболее значимые факторы и построены графики частичной зависимости. В обоих случаях определялись точность прогноза, значимость каждого из факторов по влиянию на ЛИ (рис. 1), а также анализировалось частичное влияние фактора на ЛИ в зависимости от уровня фактора. Важность факторов и стандартное отклонение вычислялись по результатам построения 2000 деревьев в каждой задаче и определялись по индексу Джини на основании снижения неопределенности при разбиении узлов: чем выше в дереве фактор, тем более значимым он считался.

Проверка гипотезы значимости отличия от нуля каждого из факторов тестом Стьюдента с последующей поправкой Бонферрони на множественное сравнение показала достоверное отличие каждого из признаков. При этом наиболее важными по влиянию на ЛИ оказались как при классификации, так и при регрессии три фактора: концентрация лактата, а также уровни глюкозы и рН. Дополнительно к важным факторам при классификации можно отнести уровень холестерина в крови.

Стоит отметить разную важность фактора курения. Если сам факт, курит ли исследуемый человек или нет, обладает практически нулевой важностью, то временной интервал между курением и забором пробы достоверно выше для обеих задач (p -значения – $1,18E-07$ для классификации и $1,26E-06$ для регрессии). Такое расхождение можно объяснить быстрым нивелированием эффекта курения у слюны.

Среднее значение метрик обучения составило: $MAE = 28,73$; $MS = 880,99$; $R^2 = 0,175$ в случае регрессии. При решении классификационной задачи точность прогноза составила 96 %. Высокая точность классификации при низком R^2 обусловлена различной природой задач регрессии и классификации. В классификации модель распознает классы по нелинейным комбинациям признаков, тогда как регрессионная модель страдает от высокой дисперсии отклика.

Частичное влияние также было оценено для двух типов задач (рис. 2, 3).

Граничным уровнем глюкозы, при котором она перестает оказывать воздействие на люциферазный индекс, является уровень выше 6 ммоль/л (см. рис. 2). Наибольшая вероятность получить значение ЛИ, близкое к верхней границе предполагаемой нормы, наблюдается при низком уровне глюкозы в крови. Для рН уровень меньше 5 практически не оказывает влияние на уровень целевой переменной. Однако при увеличении рН влияние на ЛИ характеризуется неоднозначной динамикой с большим разбросом значений. Концентрация лактата в слюне эффективно влияет на показатель люциферазного индекса при низких (до 8 ммоль/л) значениях.

Для классификационной задачи уровень глюкозы демонстрировал противоположную регрессионной зависимость – влияние этого фактора увеличивалось (рис. 3).

При этом сама динамика имела форму насыщения и значимость переходила в плато при уровне глюкозы порядка 5 ммоль/л. Лактат и рН характеризовались картиной, схожей с регрессионной. При 8 ммоль/л лактата появлялся четкий переход уровня

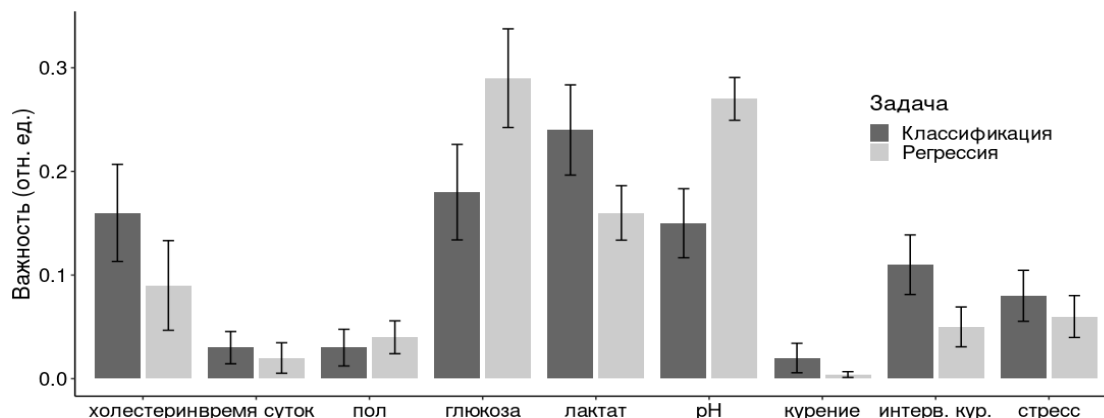


Рис. 1. Распределение факторов по важности влияния на ЛИ в регрессионной и классификационной задачах

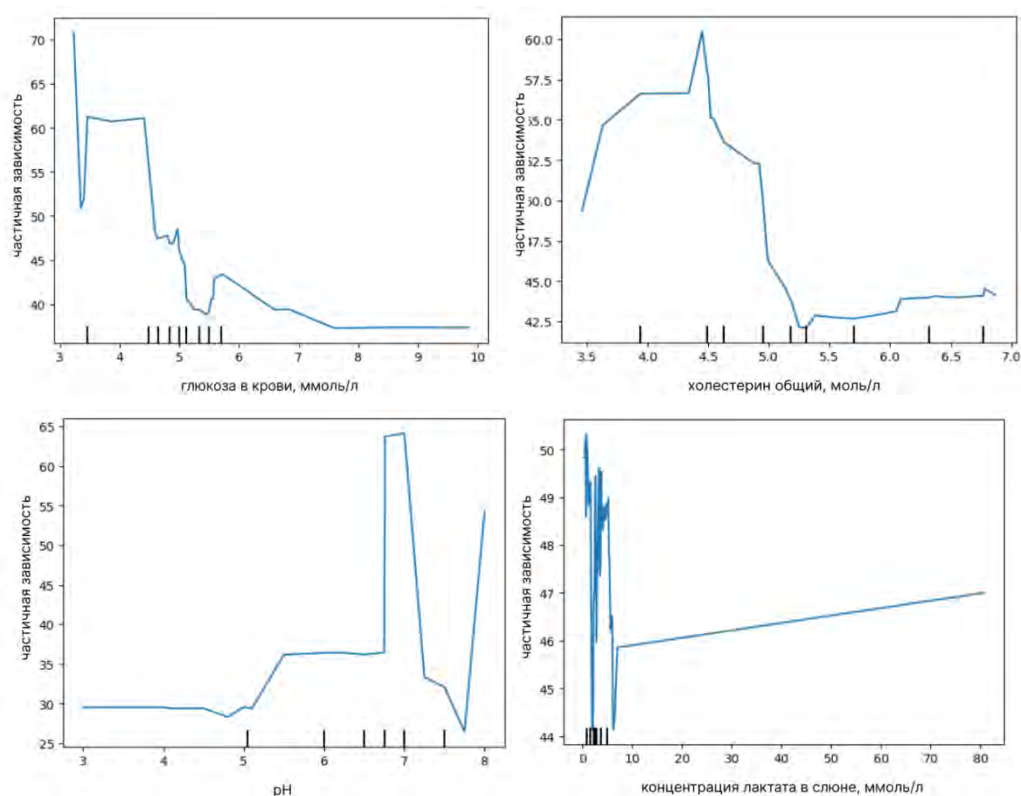


Рис. 2. Графики частичного влияния факторов на целевую переменную (ЛИ) для регрессионной задачи

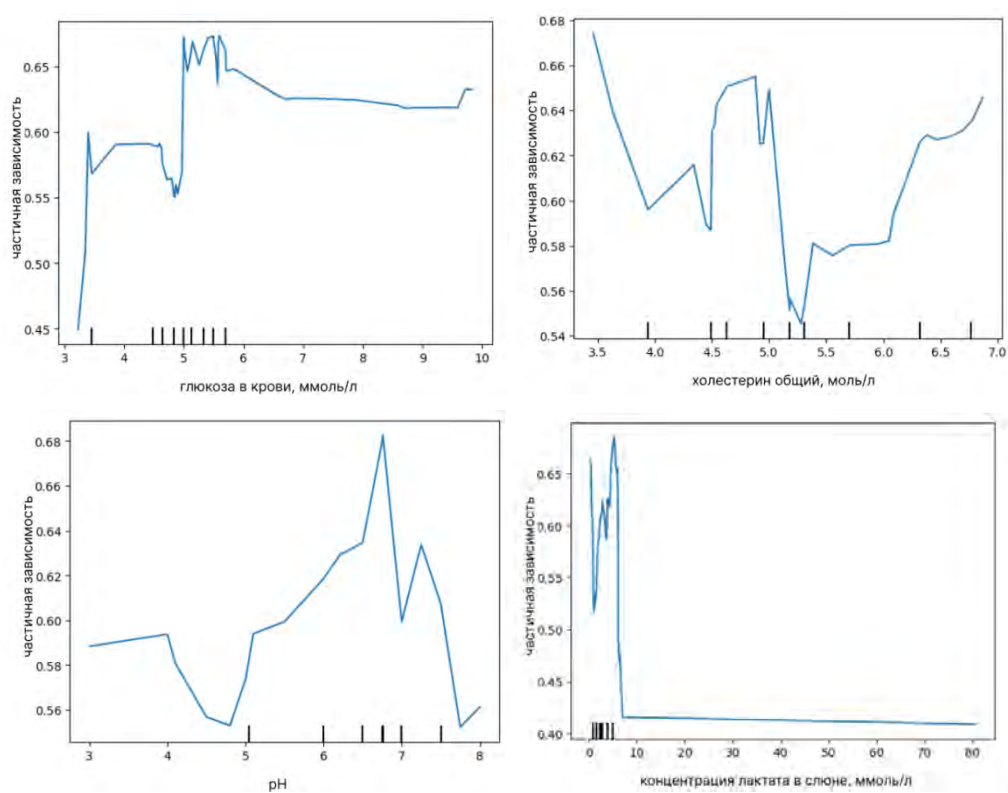


Рис. 3. Графики частичного влияния факторов на целевую переменную (ЛИ) для классификационной задачи

значимости, после которого отсутствовало влияние на люциферазный индекс. Зависимость от pH имела хаотическую динамику с общей тенденцией к росту при увеличении уровня фактора.

Выводы. Анализ факторов, проведенный с использованием алгоритма Random Forest, показал, что не все параметры в равной степени значимы для прогнозирования целевого показателя. Глюкоза демонстрирует противоположные эффекты в регрессионной и классификационной задачах: при регрессии влияние глюкозы уменьшается при превышении уровня в 6 ммоль/л, тогда как в классификации значимость фактора достигает плато при уровне около 5 ммоль/л. Это может указывать на сложные нелинейные взаимоотношения между уровнем глюкозы и состоянием исследуемого материала. Уровень pH также показал различное влияние в зависимости от задачи. При регрессии влияние pH на люциферазный индекс биолуминесцентной системы характеризуется значительной вариабельностью, особенно при увеличении уровня pH. В классификации динамика влияния pH также остается неопределенной, что затрудняет четкое понимание его роли в изменении состояния организма, несмотря на большую значимость для прогноза уровня ЛИ. Лактат показал четкое значение перехода (8 ммоль/л), после которого его влияние на люциферазный индекс значительно снижается. Это указывает на возможность существования порогового значения, после которого лактат перестает быть значимым фактором, что может быть полезно для практического применения в мониторинге состояния работников. Несмотря на низкое значение R^2 (0,175), модель успешно выделяет значимые факторы и демонстрирует высокую классификационную точность. Это указывает на сильную нелинейность и потенциальное наличие скрытых переменных, не учтенных в текущем наборе данных. Исследование не включало

в анализ ряд потенциально значимых факторов, таких как физическая активность, стрессовые ситуации вне работы и индивидуальные особенности метаболизма, хотя включение этих параметров могло бы значительно повысить точность и полезность моделей. В дальнейшем планируется провести сравнение эффективности метода Random Forest с классической линейной регрессией, логистической регрессией, XGBoost и нейронными сетями.

Метод случайного леса подходит для выбора информативных факторов, характеризующих особенности персональных характеристик организма в нормальном состоянии и при отклонениях от нормы, позволяет минимизировать количество факторов влияния на ЛИ в случае предварительной настройки теста на персонифицированную работу. Метод Random Forest может быть составной частью персонифицированного биолуминесцентного биосенсора для оценки влияния стрессовых и трудовых нагрузок на организм человека. Предложенный анализ данных может быть применен для обработки и интерпретации результатов различных методик исследования, позволяет анализировать большие объемы данных и выявлять скрытые закономерности, а также позволяет минимизировать риски неверного диагноза и корректировать схему мониторинга исследуемых, что является одним из критических аспектов в биотехнологических и медицинских исследованиях.

Финансирование. Работа поддержана грантом Министерства науки и высшего образования Российской Федерации на реализацию крупных научных проектов по приоритетным направлениям научного-технологического развития (проект № 075-15-2024-638).

Конфликт интересов. Авторы сообщают об отсутствии конфликта интересов.

Список литературы

1. Tabak L.A. Point-of-care diagnostics enter the mouth // *Ann. N Y Acad. Sci.* – 2007. – Vol. 1098. – P. 7–14. DOI: 10.1196/annals.1384.043
2. A Review of Selected Studies That Determine the Physical and Chemical Properties of Saliva in the Field of Dental Treatment / E. Kubala, P. Strzelecka, M. Grzegocka, D. Lietz-Kijak, H. Gronwald, P. Skomro, E. Kijak // *BioMed Res. Int.* – 2018. – Vol. 2018. – P. 6572381. DOI: 10.1155/2018/6572381
3. Saliva diagnostics – Current views and directions / K.E. Kaczor-Urbanowicz, C. Martin Carreras-Presas, K. Aro, M. Tu, F. Garcia-Godoy, D.T. Wong // *Exp. Biol. Med. (Maywood)*. – 2017. – Vol. 242, № 5. – P. 459–472. DOI: 10.1177/1535370216681550
4. Ферментативное биотестирование: научные основы и применение / Е.Н. Есимбекова, И.Г. Торгашина, В.П. Калябина, В.А. Кратасюк // *Сибирский экологический журнал*. – 2021. – Т. 28, № 3. – С. 364–382. DOI: 10.15372/SEJ20210308
5. A noninvasive and qualitative bioluminescent assay for express diagnostics of athletes' responses to physical exertion / V.A. Kratasyuk, L.V. Stepanova, R. Ranjan, O.S. Sutormin, S. Pande, G.V. Zhukova, O.M. Miller, N.V. Maznyak, O.A. Kolenchukova // *Luminescence*. – 2021. – Vol. 36, № 2. – P. 384–390. DOI: 10.1002/bio.3954
6. Jordan M.I., Mitchell T.M. Machine learning: Trends, perspectives, and prospects // *Science*. – 2015. – Vol. 349, № 6245. – P. 255–260. DOI: 10.1126/science.aaa8415
7. Бельская Л.В., Сарф Е.А. Применение ИК-Фурье-спектроскопии слюны для экспресс-оценки уровня продуктов перекисного окисления липидов // *Biomedical Chemistry: Research and Methods*. – 2019. – Т. 2, № 2. – С. e00094. DOI: 10.18097/BMCRM00094
8. Бельская Л.В., Сарф Е.А., Косенок В.К. Корреляционные взаимосвязи состава слюны и плазмы крови в норме // *Клиническая лабораторная диагностика*. – 2018. – Т. 63, № 8. – С. 477–482. DOI: 10.18821/0869-2084-2018-63-8-477-482
9. Никонорова М.Л. Построение оптимальных моделей анализа биомедицинских данных // *Вестник новых медицинских технологий*. – 2021. – Т. 28, № 1. – С. 55–59. DOI: 10.24412/1609-2163-2021-1-55-59
10. Берестнева О.Г., Осадчая И.А., Немерово Е.В. Методы исследования структуры медицинских данных // *Вестник науки Сибири*. – 2012. – Т. 1, № 2. – С. 333–338.
11. Клименко А.В., Слащев И.С. Кластерный анализ данных // *Вестник науки*. – 2019. – № 1 (10), Т. 1. – С. 159–163.

12. Using machine learning and an electronic tongue for discriminating saliva samples from oral cavity cancer patients and healthy individual / D.C. Braz, M. Popolin Neto, F.M. Shimizu, A.C. Sá, R.S. Lima, A.L. Gobbi, M.E. Melendez, L.M.R.B. Arantes [et al.] // *Talanta*. – 2022. – Vol. 243. – P. 123327. DOI: 10.1016/j.talanta.2022.123327
13. Microbiome of Saliva and Plaque in Children According to Age and Dental Caries Experience / E. Lee, S. Park, S. Um, S. Kim, J. Lee, J. Jang, H.-O. Jeong, J. Shin [et al.] // *Diagnostics (Basel)*. – 2021. – Vol. 11, № 8. – P. 1324. DOI: 10.3390/diagnostics11081324
14. Finding distinctions between oral cancer and periodontitis using saliva metabolites and machine learning / V.L. Kouznetsova, J. Li, E. Romm, I.F. Tsigelny // *Oral Dis.* – 2021. – Vol. 27, № 3. – P. 484–493. DOI: 10.1111/odi.13591
15. Zarrin P.S., Roeckendorf N., Wenger C. In-Vitro Classification of Saliva Samples of COPD Patients and Healthy Controls Using Machine Learning Tools // *IEEE Access*. – 2020. – Vol. 8. – P. 168053–168060. DOI: 10.1109/ACCESS.2020.3023971
16. Salivary metabolomics with alternative decision tree-based machine learning methods for breast cancer discrimination / T. Murata, T. Yanagisawa, T. Kurihara, M. Kaneko, S. Ota, A. Enomoto, M. Tomita, M. Sugimoto [et al.] // *Breast Cancer Res. Treat.* – 2019. – Vol. 177, № 3. – P. 591–601. DOI: 10.1007/s10549-019-05330-9
17. Applications of machine learning in routine laboratory medicine: Current state and future directions / N. Rabbani, G.Y.E. Kim, C.J. Suarez, J.H. Chen // *Clin. Biochem.* – 2022. – Vol. 103. – P. 1–7. DOI: 10.1016/j.clinbiochem.2022.02.011
18. The oral microbiome of pregnant women facilitates gestational diabetes discrimination / X. Li, J. Zheng, X. Ma, B. Zhang, J. Zhang, W. Wang, C. Sun, Y. Wang [et al.] // *J. Genet. Genomics*. – 2021. – Vol. 48, № 1. – P. 32–39. DOI: 10.1016/j.jgg.2020.11.006
19. Salivary metabolomics with machine learning for colorectal cancer detection / H. Kuwabara, K. Katsumata, A. Iwabuchi, R. Udo, T. Tago, K. Kasahara, J. Mazaki, M. Enomoto [et al.] // *Cancer Sci.* – 2022. – Vol. 113, № 9. – P. 3234–324. DOI: 10.1111/cas.15472
20. Alam M.Z., Rahman M.S., Rahman M.S. A Random Forest based predictor for medical data classification using feature ranking // *Informatics in Medicine Unlocked*. – 2019. – Vol. 15. – P. 100180. DOI: 10.1016/j.imu.2019.100180
21. Применение биолуминесцентного ферментативного биотеста для анализа слюны работников железнодорожного транспорта с целью мониторинга функционального состояния организма в условиях трудовой деятельности / Л.В. Степанова, О.А. Коленчукова, Г.В. Жукова, О.С. Сутормин, В.А. Кратасюк // *Биофизика*. – 2024. – Т. 69, № 3. – С. 674–683. DOI: 10.31857/S0006302924030224

Метод Random Forest в задачах интерпретации результатов биолуминесцентного анализа слюны при персонализированной диагностике / Г.В. Жукова, П.А. Мартышук, Е.Р. Афер, А.Н. Шуваев, Н.А. Розанова, Д.В. Сергеев, В.А. Кратасюк // Анализ риска здоровью. – 2025. – № 2. – С. 166–174. DOI: 10.21668/health.risk/2025.2.14

UDC 519.254:519.257

DOI: 10.21668/health.risk/2025.2.14.eng



Research article

RANDOM FOREST METHOD FOR INTERPRETING RESULTS OBTAINED BY BIOLUMINESCENCE ANALYSIS OF SALIVA IN PERSONALIZED DIAGNOSTICS

**G.V. Zhukova¹, P.A. Martyshuk¹, E.R. Afer¹, A.N. Shuvaev¹,
N.A. Rozanova², D.V. Sergeev², V.A. Kratasyuk^{1,3}**

¹Siberian Federal University, 79 Svobodnyi Av., Krasnoyarsk, 660041, Russian Federation

²Research Center for Neurology, 80 Volokolamskoe highway, Moscow, 125367, Russian Federation

³Institute of Biophysics, Siberian Department of the RAS, separate division of the Federal Research Center Krasnoyarsk Scientific Center of the Siberian Department of the Russian Academy of Sciences, 50 Akademgorodok, build. 50, Krasnoyarsk, 660036, Russian Federation

Development of personalized medicine and biotechnologies is directly linked to obtaining relevant data, which largely depend on individual characteristics of examined patients. Permissible ranges of analyzed indicators that are commonly used in conventional medicine do not always describe a patient's health adequately. It seems necessary to search for such data analysis techniques, which allow considering variable individual peculiarities of patients' bodies and their lifestyles.

The aim of this study is to determine whether it is possible to use the Random Forest method for biomedical data analysis in order to achieve correct interpretation of results obtained by personalized diagnostic tests. Bioluminescent testing is used as an example since it estimates effects produced by various characteristics of examined patients and their living conditions. The method allows minimizing risks of incorrect diagnosis and adjusting monitoring schemes for specific patients.

This study relies on using the results obtained by diagnosing workloads of railway workers using the bioluminescence method. A patient's health is assessed by examining effects produced by a patient's saliva on intensity of the bi-enzyme system luminescence: NAD(P)H:FMN oxidoreductase + luciferase. This analysis is integral and responses to many factors,

each of which can influence the analysis result. Effectiveness of various methods for data analysis is assessed on an example group made of traffic controllers employed by the Krasnoyarsk Branch of Russian Railways JSC. Both statistical methods and the Random Forest machine learning algorithm were used for data analysis.

As a result, our study has revealed that it is advisable to use the Random Forest method for assessing significance of some biochemical saliva indicators to predict health of railway workers. The method makes it possible to identify the most significant factors and create graphs to show partial influence exerted by various factors on the target variable. This study allows optimizing the system for health diagnostics using integral bioluminescence analysis. The Random Forest method can become a component of a personalized bioluminescent biosensor for assessing effects produced by stress and workloads on the body.

Keywords: personalized diagnostics, machine learning, data analysis, multifactorial analysis, saliva, bioluminescence, biosensor, signal systems.

References

1. Tabak L.A. Point-of-care diagnostics enter the mouth. *Ann. N Y Acad. Sci.*, 2007, vol. 1098, pp. 7–14. DOI: 10.1196/annals.1384.043
2. Kubala E., Strzelecka P., Grzegocka M., Lietz-Kijak D., Gronwald H., Skomro P., Kijak E. A Review of Selected Studies That Determine the Physical and Chemical Properties of Saliva in the Field of Dental Treatment. *BioMed Res. Int.*, 2018, vol. 2018, pp. 6572381. DOI: 10.1155/2018/657238
3. Kaczor-Urbanowicz K.E., Carreras-Presas C.M., Aro K., Tu M., Garcia-Godoy F., Wong D.T. Saliva diagnostics – Current views and directions. *Exp. Biol. Med. (Maywood)*, 2017, vol. 242, no. 5, pp. 459–472. DOI: 10.1177/1535370216681550
4. Esimbekova E.N., Torgashina I.G., Kalyabina V.P., Kratasyuk V.A. Enzymatic Biotesting: Scientific Basis and Application. *Contemp. Probl. Ecol.*, 2021, vol. 14, pp. 290–304. DOI: 10.1134/S1995425521030069
5. Kratasyuk V.A., Stepanova L.V., Ranjan R., Sutormin O.S., Pande S., Zhukova G.V., Miller O.M., Maznyak N.V., Kolenchukova O.A. A noninvasive and qualitative bioluminescent assay for express diagnostics of athletes' responses to physical exertion. *Luminescence*, 2021, vol. 36, no. 2, pp. 384–390. DOI: 10.1002/bio.3954
6. Jordan M.L., Mitchell T.M. Machine learning: Trends, perspectives, and prospects. *Science*, 2015, vol. 349, no. 6245, pp. 255–260. DOI: 10.1126/science.aaa8415
7. Bel'skaya L., Sarf E. The use of Fourier transform IR spectroscopy of saliva for rapid assessment of the level of lipid peroxidation products. *Biomedical Chemistry: Research and Methods*, 2019, vol. 2, no. 2, pp. e00094. DOI: 10.18097/BMCRM00094 (in Russian).
8. Bel'skaya L.V., Sarf E.A., Kosenok V.K. Correlation interrelations between the composition of saliva and blood plasmain norm. *Klinicheskaya laboratornaya diagnostika*, 2018, vol. 63, no. 8, pp. 477–482. DOI: 10.18821/0869-2084-2018-63-8-477-482 (in Russian).
9. Nikonorova M.L. Building optimal models for analysis of biomedical data. *Vestnik novykh meditsinskikh tekhnologii*, 2021, vol. 28, no. 1, pp. 55–59. DOI: 10.24412/1609-2163-2021-1-55-59 (in Russian).
10. Berestneva O.G., Osadchaya I.A., Nemerov E.V. Methods for studying the medical data structure. *Vestnik nauki Sibiri*, 2012, vol. 1, no. 2, pp. 333–338 (in Russian).
11. Klimenko A.V., Slashchev I.S. Klasternyi analiz dannykh [Cluster data analysis]. *Vestnik nauki*, 2019, no. 1 (10), vol. 1, pp. 159–163 (in Russian).
12. Braz D.C., Popolin Neto M., Shimizu F.M., Sá A.C., Lima R.S., Gobbi A.L., Melendez M.E., Arantes L.M.R.B. [et al.]. Using machine learning and an electronic tongue for discriminating saliva samples from oral cavity cancer patients and healthy individual. *Talanta*, 2022, vol. 243, pp. 123327. DOI: 10.1016/j.talanta.2022.123327
13. Lee E., Park S., Um S., Kim S., Lee J., Jang J., Jeong H.-O., Shin J. [et al.]. Microbiome of Saliva and Plaque in Children According to Age and Dental Caries Experience. *Diagnostics (Basel)*, 2021, vol. 11, no. 8, pp. 1324. DOI: 10.3390/diagnostics11081324

© Zhukova G.V., Martyshuk P.A., Afer E.R., Shuvaev A.N., Rozanova N.A., Sergeev D.V., Kratasyuk V.A., 2025

Galina V. Zhukova – Junior Researcher of Bioluminescent Technologies Laboratory (e-mail: gvivanova@sfu-kras.ru; tel.: +7 (913) 047-88-37; ORCID: <https://orcid.org/0000-0002-8646-1224>).

Polina A. Martishuk – student of the Institute of Fundamental Biology and Biotechnology (e-mail: p.martishuk@yandex.ru; tel.: +7 (904) 896-97-34).

Egor R. Afer – student of the Institute of Fundamental Biology and Biotechnology (e-mail: egorafer@yandex.ru; tel.: +7 (902) 957-64-89).

Andrey N. Shuvaev – Candidate of Physical and Mathematical Sciences, Head of the Basic Department of Biomedical Systems and Complexes of the Institute of Fundamental Biology and Biotechnology (e-mail: ashuvaev@sfu-kras.ru; tel.: +7 (950) 997-77-92; ORCID: <https://orcid.org/0000-0002-3887-1413>).

Natalia A. Rozanova – graduate student, laboratory research assistant at Brain Institute's Neurobiology and Tissue Engineering Laboratory (e-mail: nataliarozanovaa@gmail.com; tel.: +7 (968) 030-47-28; ORCID: <https://orcid.org/0000-0001-9619-4679>).

Dmitry V. Sergeev – Candidate of Medical Sciences, Senior Researcher, Neurologist (e-mail: sergeev@neurology.ru; tel.: +7 (926) 860-08-97; ORCID: <https://orcid.org/0000-0002-9130-1292>).

Valentina A. Kratasyuk – Doctor of Biological Sciences, Professor, Head of Biophysical Department of the Institute of Fundamental Biology and Biotechnology; Leading Researcher (e-mail: valkrat@mail.ru; ORCID: <https://orcid.org/0000-0001-6764-5231>).

14. Kouznetsova V.L., Li J., Romm E., Tsigelny I.F. Finding distinctions between oral cancer and periodontitis using saliva metabolites and machine learning. *Oral Dis.*, 2021, vol. 27, no. 3, pp. 484–493. DOI: 10.1111/odi.13591
15. Zarrin, P.S., Roeckendorf N., Wenger C. In-Vitro Classification of Saliva Samples of COPD Patients and Healthy Controls Using Machine Learning Tools. *IEEE Access*, 2020, vol. 8, pp. 168053–168060. DOI: 10.1109/ACCESS.2020.3023971
16. Murata T., Yanagisawa T., Kurihara T., Kaneko M., Ota S., Enomoto A., Tomita M., Sugimoto M. [et al.]. Salivary metabolomics with alternative decision tree-based machine learning methods for breast cancer discrimination. *Breast Cancer Res. Treat.*, 2019, vol. 177, no. 3, pp. 591–601. DOI: 10.1007/s10549-019-05330-9
17. Rabbani N., Kim G.Y.E., Suarez C.J., Chen J.H. Applications of machine learning in routine laboratory medicine: Current state and future directions. *Clin. Biochem.*, 2022, vol. 103, pp. 1–7. DOI: 10.1016/j.clinbiochem.2022.02.011
18. Li X., Zheng J., Ma X., Zhang B., Zhang J., Wang W., Sun C., Wang Y. [et al.]. The oral microbiome of pregnant women facilitates gestational diabetes discrimination. *J. Genet. Genomics*, 2021, vol. 48, no. 1, pp. 32–39. DOI: 10.1016/j.jgg.2020.11.006
19. Kuwabara H., Katsumata K., Iwabuchi A., Udo R., Tago T., Kasahara K., Mazaki J., Enomoto M. [et al.]. Salivary metabolomics with machine learning for colorectal cancer detection. *Cancer Sci.*, 2022, vol. 113, no. 9, pp. 3234–3243. DOI: 10.1111/cas.15472
20. Alam M.Z., Rahman M.S., Rahman M.S. A Random Forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*, 2019, vol. 15, pp. 100180. DOI: 10.1016/j.imu.2019.100180
21. Stepanova L.V., Kolenchukova O.A., Zhukova G.V., Sutormin O.S., Kratasyuk V.A. The Use of the Bioluminescent Enzyme Bioassay for the Analysis of Saliva of Railway Transport Workers to Monitor the Functional State of the Body in the Conditions of Labor Activity. *Biofizika*, 2024, vol. 69, no. 3, pp. 674–683. DOI: 10.31857/S0006302924030224 (in Russian).

Zhukova G.V., Martyshuk P.A., Afer E.R., Shuvaev A.N., Rozanova N.A., Sergeev D.V., Kratasyuk V.A. Random Forest method for interpreting results obtained by bioluminescence analysis of saliva in personalized diagnostics. Health Risk Analysis, 2025, no. 2, pp. 166–174. DOI: 10.21668/health.risk/2025.2.14.eng

Получена: 22.12.2024

Одобрена: 07.05.2025

Принята к публикации: 14.06.2025