www.volsu.ru

DOI: https://doi.org/10.15688/mpcm.jvolsu.2025.2.6

УДК 519.68
ББК 56.12

Дата поступления статьи: 08.06.2025
Дата принятия статьи: 18.06.2025

ИНТЕРПРЕТАЦИЯ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ ПО ДАННЫМ МИКРОВОЛНОВОЙ РАДИОТЕРМОМЕТРИИ [1]

Илларион Евгеньевич Попов

Аспирант кафедры математического анализа и теории функций,
Волгоградский государственный университет
popov.larion@volsu.ru
https://orcid.org/0000-0002-0997-8721
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Аннотация. В статье рассматривается задача повышения интерпретируемости решений математических моделей при сохранении высокой точности предсказаний. Основное внимание уделяется интеграции нейронных сетей, обладающих высокой точностью, с ансамблем интерпретируемых моделей, механизмы которых прозрачны и поддаются аналитическому описанию. Предложен метод формирования обоснования на основе решений ансамбля, согласованных с предсказанием нейросети. Согласование достигается путем сравнения степеней уверенности моделей после предварительной калибровки, необходимость которой обусловлена эффектом избыточной уверенности (overconfidence), характерным для некоторых моделей машинного обучения. Разработан метод выбора интерпретируемых моделей ансамбля классификаторов, чьи оценки на конкретном объекте максимально близки по степени уверенности к выходу нейронной сети. Это позволяет формировать обоснования, содержащие как аргументы в пользу принятого решения, так и возможные альтернативные мнения. Для повышения гибкости интерпретации введено расширенное определение высокоинформативного признака, включающее категоризацию значений по степени их характерности для различных классов. Показано, что переход от бинарных к категориальным признакам способствует росту точности классификации и увеличивает ее общую эффективность. Дополнительно разработан метод построения информативных интервалов признаков, позволяющий повысить их информативность – разделяющую способность. На основе полученных интервалов предложены алгоритмы уточнения полуопределенных меток и коррекции обучающей выборки с целью повышения ее качества и репрезентативности. Предложенные подходы протестированы на задаче диагностики рака молочной железы по данным микроволновой

радиотермометрии. Результаты вычислительных экспериментов подтверждают, что использование категориальных интерпретируемых признаков в сочетании с модельной калибровкой позволяет существенно повысить точность классификации и обоснованность принимаемых решений.

Ключевые слова: задача интерпретации, математическая модель, машинное обучение, набор данных, классификация.

Введение

В настоящее время в исследованиях искусственного интеллекта (ИИ) наблюдается значительный рост интереса к объяснимому искусственному интеллекту (eXplainable Artificial Intelligence, XAI) [27; 29]. Этот интерес обусловлен все более глубокой интеграцией ИИ-систем в различные сферы деятельности, где принимаемые решения оказывают критически важное влияние, например, в медицине, экономике, праве, а также в системах безопасности. В этих областях требуется не только высокая точность предсказаний, но и возможность объяснить, почему система приняла то или иное решение. Это необходимо для повышения доверия со стороны экспертов, принимающих решения, и пользователей, а также для обеспечения прозрачности работы алгоритмов и выявления потенциальных ошибок или предвзятости. Прозрачность работы моделей, в свою очередь, способствует их регулированию на соответствие нормам закона и этики, что позволяет внедрять системы ИИ в сферы деятельности с высоким уровнем контроля.

Сложность объяснимости моделей ИИ обусловлена тем, что современные алгоритмы машинного обучения, особенно нейронные сети, представляют собой комплексные нелинейные системы с большим числом параметров [10]. Такие модели зачастую функционируют как «черные ящики», в которых сложно понять внутреннюю логику принятия решений. Однако в области XAI принято выделять два уровня прозрачности моделей:

- **Объяснимые модели** [13] – представляют собой «черный ящик», где невозможно напрямую интерпретировать процесс принятия решений, однако существуют методы, позволяющие анализировать вклад признаков в итоговый результат. Примеры таких методов включают SHAP (SHapley Additive exPlanations [14]), Grad-CAM (Gradient-weighted Class Activation Mapping [33]) и LIME (Local Interpretable Model-agnostic Explanations [11]). Данные методы помогают выявлять, какие факторы наиболее сильно влияют на предсказания модели, что особенно полезно в медицинской диагностике и финансовом анализе.
- **Интерпретируемые модели** [30] – представляют собой «белый ящик», где процесс принятия решений полностью прозрачен и поддается анализу. К таким моделям относятся деревья решений, линейные модели, наивный байесовский классификатор и некоторые виды ансамблевых методов. Достоинство этих моделей заключается в простоте их интерпретации, что делает их востребованными в критически важных приложениях, таких как медицина, юридический анализ и оценка рисков.

Основным недостатком интерпретируемых моделей является их сравнительно более слабая предсказательная способность по сравнению с глубокими нейронными сетями. В свою очередь, последние демонстрируют высокую эффективность в широком

спектре задач, таких как компьютерное зрение, обработка естественного языка и прогнозирование временных рядов [16; 17]. Таким образом, выбор между объяснимыми и интерпретируемыми моделями ИИ сводится к нахождению компромисса между точностью предсказаний и степенью прозрачности. В задачах, где ошибки не оказывают критического влияния, например, в рекомендательных системах, можно использовать нейронные сети без механизма объяснения решений. Однако во многих областях прозрачность играет ключевую роль, что делает оправданным применение менее точных, но более интерпретируемых моделей.

Основные направления исследований в области XAI на текущий день связаны с повышением прозрачности «черных ящиков», а именно нейронных сетей. Исследуются архитектуры, способствующие выделению интерпретируемых признаков и областей интереса, алгоритмы контекстного объяснения при работе с языковыми данными [20; 21; 26; 31]. Однако фундаментальная проблема нейронных сетей, заключенная в их непрозрачности, остается, из-за чего продолжают быть актуальными вопросы этики и безопасности в применении систем ИИ.

В данной работе предлагается модель, объединяющая преимущества обоих подходов, что позволяет одновременно достичь высокой точности и прозрачности системы ИИ. Известно, что объединение классических моделей машинного обучения (например, деревьев решений) в ансамбль может повысить общую точность предсказаний. Это достигается за счет того, что каждая модель ансамбля обучается на различных подмножествах признаков и данных, что обеспечивает их разнообразие. Высока вероятность, что среди ансамбля решений найдется модель, чьи предсказания будут схожи с результатами высокоточных нейронных сетей [9]. Включение интерпретируемых методов объяснения в такую систему позволяет не только повысить прозрачность работы алгоритма, но и обеспечить согласованность решений, что критически важно для практических приложений в ответственных сферах. Таким образом, рассматриваемый подход направлен на разработку гибридных решений, объединяющих возможности глубокого обучения и интерпретируемости традиционных методов.

Актуальной областью применения рекомендательных систем ИИ, особенно требующей прозрачности используемых моделей, является медицинская диагностика. Поскольку такие системы не несут ответственности за свои рекомендации, врачу необходимо не только получать предлагаемый вывод, но и понимать его предпосылки. В настоящее время системы ИИ активно разрабатываются совместно с методом микроволновой радиотермометрии [24]. Данный метод позволяет измерять глубинные температуры тела, на основе чего фиксируются температурные аномалии биологических тканей и органов, свидетельствующие о наличии различных заболеваний [6]. По измеренным температурам строятся модели машинного обучения, которые показывают свою точность при обследовании различных органов и заболеваний [4]. При этом для более эффективного применения систем были разработаны математические модели, формализующие знания врачей. Модели описывают характеристики температурных полей органов, на основе которых врачи определяют диагноз. Например, можно выделить характеристику, описывающую наличие области с повышенной температурой. Как правило, повышенная температура свидетельствует о воспалительных процессах, вызванных заболеванием.

Данные математические модели не только повышают точность машинного обучения, но и качество обоснования. Сложность обоснования напрямую зависит от информативности входных признаков: чем она меньше, тем в большей степени необходимо увеличение сложности моделей машинного обучения. Так, например, для сохранения точно-

сти модели увеличивается глубина дерева решений при обработке большого количества малоинформативных признаков. И напротив, при использовании высокоинформативных признаков требуется меньшая глубина дерева решений. В данной работе рассматривается модель с использованием высокоинформативных признаков [5; 19] и их модификация для задач обоснования, а именно: модификация метода построения интервалов, а также дискретизация признаков на несколько категорий с учетом характерности для каждого из рассматриваемых классов.

Другим фактором, влияющим на качество систем ИИ, является качество обучающего набора данных [25]. А именно, его корректность и полнота, что зачастую слабо представлено в медицинских данных. Постановка корректного диагноза является нетривиальной задачей и медицинские наборы данных могут нести в себе нехарактерные метки, то есть такие, которые не свойственны измерениям. Полнота же может быть слабо представлена в силу малого объема данных, вследствие чего он является нерепрезентативным, что негативно влияет как на точность моделей машинного обучения, так и на адекватность обоснования, которое может делать акценты на второстепенных характеристиках. В работе рассматриваются подходы, корректирующие данные характеристики.

1. Материалы и предшествующие результаты

В работе исследуется набор данных, полученных в результате обследований методом микроволновой радиотермометрии. В ходе обследований производилось измерение температур органа и тела по определенной схеме (см. рис. 1).

В каждой пронумерованной области измерялись глубинные и кожные температуры. Таким образом, набор данных по каждому пациенту содержит следующие температурные измерения:

- $T_r^{mw} = (t_{0,r}^{mw}, t_{1,r}^{mw}, ..., t_{9,r}^{mw})$ – множество глубинных температур правой молочной железы;
- $T_l^{mw} = (t_{0,l}^{mw}, t_{1,l}^{mw}, ..., t_{9,l}^{mw})$ – множество глубинных температур левой молочной железы;
- $T_r^{ir} = (t_{0,r}^{ir}, t_{1,r}^{ir}, ..., t_{9,r}^{ir})$ – множество кожных температур правой молочной железы;
- $T_l^{ir} = (t_{0,l}^{ir}, t_{1,l}^{ir}, ..., t_{9,l}^{ir})$ – множество кожных температур левой молочной железы;
- $T_a^{mw} = (t_{1,a}^{mw}, t_{2,a}^{mw})$ – множество глубинных температур опорной области (точки $T1$ и $T2$ на схеме);
- $T_a^{ir} = (t_{1,a}^{ir}, t_{2,a}^{ir})$ – множество кожных температур опорной области.

По проведенным измерениям врачами осуществляется анализ температурных полей молочных желез (пример полей на рис. 2) на наличие и степень выраженности температурных отклонений от нормы. В результате анализа врач ставит в соответствие пациенту или молочным железам оценку – уровень выраженности температурных аномалий – по шкале от 0 до 5. Где 0 – температурные аномалии не выявлены; 5 – температурные аномалии характерны для рака молочной железы. Как правило, решается задача бинарной классификации, а именно определения группы риска, поэтому данные классы объединяются в два основных: метки 0, 1, 2 – в группу здоровых; 3, 4, 5 – в группу риска [1].

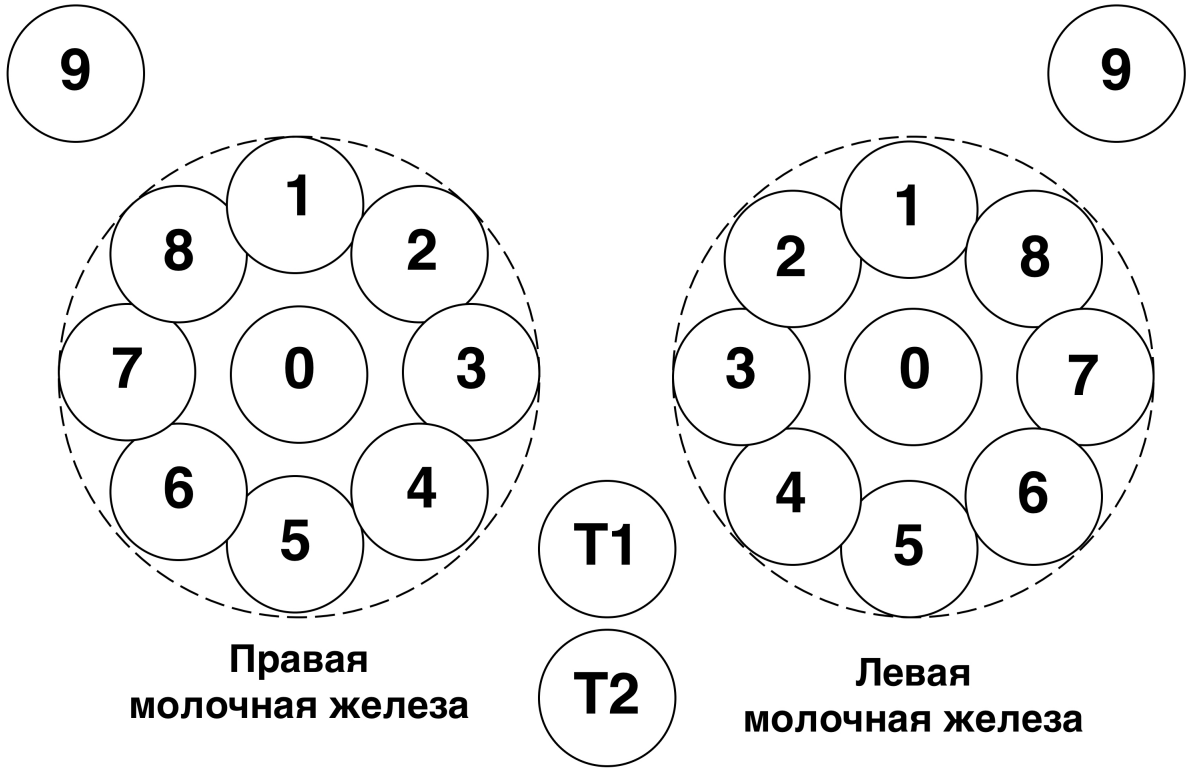


Рис. 1. Схема обследования. Здесь точки 0–8 – температуры молочной железы (правой или левой); 9 – аксиллярная область; $T1$ и $T2$ – опорные точки

В связи с тем, что в некоторых случаях врачи ставят метку пациентам, становится затруднительной задача определения состояния отдельной молочной железы, так как в таких случаях метка считается полуопределенной. Известно, что одна из молочных желез обладает характеристиками заданного класса, но неизвестно, какая из них. При этом исходя из предметной области полагается, что парная молочная железа не может принадлежать к более высокому классу, так как врачи ставят метку наивысшего класса уровня выраженности температурных аномалий. В таблице 1 представлен состав набора данных с известными и полуопределенными метками молочных желез.

Таблица 1

Состав набора данных

Класс	0	1	2	3	4	5
Количество объектов с известными метками	5 744	1 512	1 221	27	370	194
Количество объектов с полуопределенными метками	0	1 721	3 254	527	290	124

Ранее исследователями были построены математические модели диагностического состояния пациентов, количественно описывающие различные температурные аномалии органа. При этом было показано, что метод построения математической модели носит универсальный характер и применим к различным органам и заболеваниям (молочные железы и рак, нижние конечности и венозные заболевания, легкие и пневмония и т. д. [7; 18; 23]). В качестве примера элементов математических моделей выделим несколько:

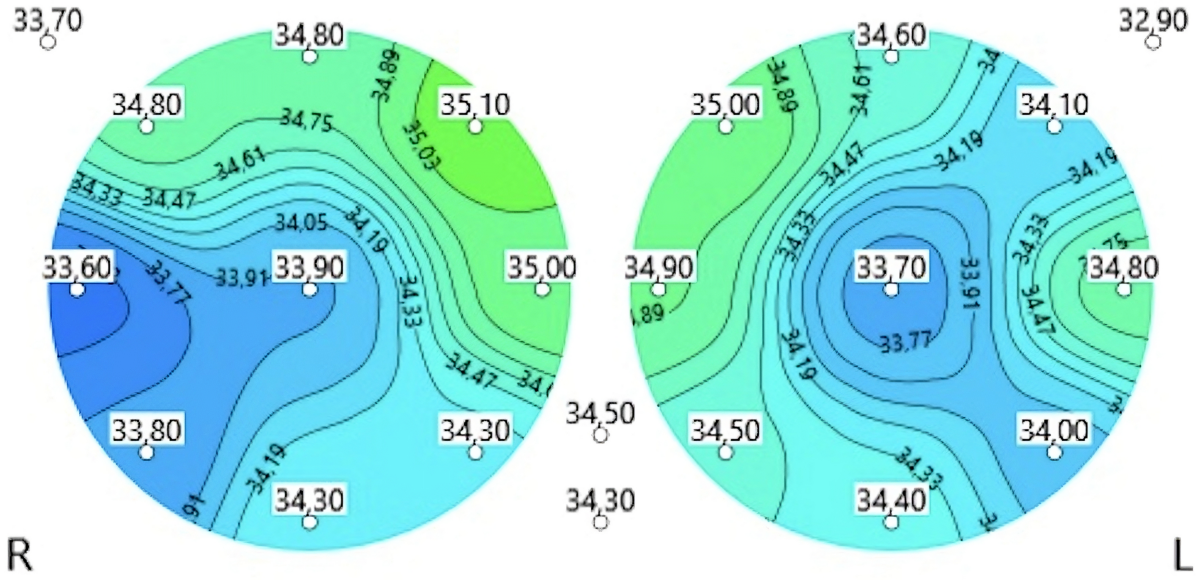


Рис. 2. Поле кожных температур молочных желез с отклонениями. Цвет соответствует значению температуры: синий – наиболее холодные области, зеленый – средние температуры молочной железы, красный – наиболее горячие области. Наличие синих областей и асимметрии полей правой и левой молочных желез может свидетельствовать о нарушении физиологических процессов

- Положение о симметрии температурных полей парных органов или систем. Наличие термоасимметрии свидетельствует о нарушении процессов в одном из органов. Симметрия может характеризоваться разностью средних значений глубинных температур

$$\overline{T_r^{mw}} - \overline{T_l^{mw}}.$$

- Положение о наличии областей с повышенным уровнем температур, что, как правило, свидетельствует о воспалительных процессах в этой области. Наличие области может характеризоваться величиной разброса глубинных температур

$$\sqrt{\frac{\sum_{t \in T_r^{mw}} (t - \overline{T_r^{mw}})^2}{|T_r^{mw}| - 1}},$$

где $|\cdot|$ – мощность множества.

При этом элементами математической модели являются термометрические признаки: тройка значений $\phi = (f, I, W)$, где f – функция (например, разброс глубинных температур или разность средних значений, описанные ранее); I – интервал; W – информативность функции f на интервале I. Информативность характеризует то, насколько хорошо с помощью рассматриваемого интервала можно разделить объекты различных классов. На основе этого можно построить простой предикат, определяющий, входит ли значение функции f, полученное при данном обследовании, в интервал I. Если входит, значит значение характерно для мажоритарного класса в данном интервале. А на основе W возможно формирование математической модели, содержащей высокоинформативные термометрические признаки.

Использование термометрических признаков модели в задаче классификации позволяет повысить точность моделей машинного обучения, а также их интерпретируемость.

2. Интерпретация

Основные критерии качества интерпретации математических моделей следующие.

- **Понятность и простота интерпретации.** Чем менее информативны признаки, используемые математической моделью, тем более объемной и сложной в понимании будет интерпретация. Несмотря на то что описанные математические модели повышают информативность входных признаков, актуальным остается разработка алгоритма нахождения интервалов, максимизирующих значение W, так как в работе [3] используется жадный алгоритм, который не всегда оптимален с точки зрения построения интервала для конкретного класса. Другим актуальным направлением является категоризация признака. Так как при бинаризации теряется большое количество информации, включающей в себя принадлежность к характерным значениям классов здоровых или группы риска.
- **Интерпретируемость и точность.** Модели машинного обучения, которые строятся по элементам математической модели, должны быть точными и в то же время интерпретируемыми.
- **Адекватность.** Адекватность характеризует, насколько решение и объяснение модели соответствует действительности. В первую очередь на адекватность влияет качество обучающего набора данных. А именно количество шумов (как в метках, так и в признаках), объем и репрезентативность данных. При слабой репрезентативности обучающего набора данных модели могут не учитывать некорректные закономерности или опираться на второстепенные факторы.

2.1. Информативность термометрических признаков

Опишем алгоритм построения интервала I по признаку f для класса 0.

1) На входе алгоритму поступают множества f_0 и f_1, содержащие значения функции f объектов класса 0 и 1 соответственно, а также указывается значение s – шаг расширения интервала.
2) Левая (I_l) и правая (I_r) границы интервала I определяются как медиана множества f_0.
3) Шаг расширения интервала вправо s_r принимается равным значению s.
4) Пока $s_r > \epsilon$, интервал I расширяется вправо на шаг s_r

$$I_r = I_r + s_r.$$

Если информативность W интервала $(I_l; I_r)$ меньше, чем информативность интервала $(I_l; I_r - s_r)$, то возвращаются предыдущие границы интервала и шаг s_r уменьшается вдвое

$$I_r = I_r - s_r, \quad s_r = \frac{s_r}{2}.$$

5) Шаг расширения интервала влево s_l принимается равным значению s.
6) Пока $s_l > \epsilon$, интервал I расширяется влево на шаг s_l

$$I_l = I_l - s_l.$$

Если информативность W интервала $(I_l; I_r)$ меньше, чем информативность интервала $(I_l + s_l; I_r)$, то возвращаются предыдущие границы интервала и шаг s_l уменьшается вдвое

$$I_l = I_l + s_l, \quad s_l = \frac{s_l}{2}.$$

7) В результате формируется интервал $I = (I_l; I_r)$.

Значение ϵ является настраиваемым параметром, рекомендуемое значение – минимальное расстояние между элементами множества $f_0 \cup f_1$

$$\min_{\substack{x,y \in f_0 \cup f_1 \\ x \neq y}} |x - y|,$$

так как при меньшем значении шага расширение интервала не всегда будет изменять соотношение количества элементов в нем.

В работе информативность W оценивается по формуле

$$W(f, I) = \sqrt{l_0 \cdot (1 - l_1)},$$

где

$$l_n = \frac{|\{x \in f_n \mid I_l < x < I_r\}|}{|f_0 \cup f_1|} \quad -$$

доля объектов класса n, попавших в интервал $(I_l; I_r)$, характерный для данного класса.

Чтобы выявить эффективность данного метода, было проведено сравнение с методом жадного объединения [3]. В качестве признакового пространства были выбраны функции из работы [2, прил. А], интервалы строились для каждого класса (0, 1, ... 5) и проводилось сравнение средних информативностей по каждому из них. В таблице 2 представлены результаты. Как видно, в среднем описанный выше метод значительно повышает информативность признаков, хотя в частных случаях она может падать на 1–2 %.

Таблица 2

Информативность интервалов

Номер класса	W по жадному алгоритму	W по описанному алгоритму
0	0,56	0,60
1	0,38	0,51
2	0,41	0,48
3	0,49	0,64
4	0,36	0,59
5	0,41	0,61

3. Интерпретируемость и точность

Сформулируем задачу гибридного обоснования следующим образом: имеются обученные на задачу классификации модели высокоточной нейронной сети и ансамбля интерпретируемых алгоритмов. Каждая из них определяется функцией классификации

$$C : \mathbb{R}^n \to [0; 1],$$

для объектов, описываемых n признаками. Пусть C_{nn} – функция нейронной сети; C_a – функция ансамбля алгоритмов. При этом последняя формируется на основе множества моделей ансамбля

$$C_a = A(C_{a1}, ..., C_{am}),$$

где A – некоторая агрегирующая функция, принимающая итоговое решение о принадлежности объекта к какому-либо классу; $C_{a1}, ..., C_{am}$ – модели ансамбля. Так как нейронная сеть является более точной моделью, необходимо сформировать обоснование, наиболее согласованное с ней. Также для специалистов может иметь смысл формирование краевых обоснований, приводящих аргументы в пользу как предлагаемого класса, так и противоположных.

Для этого в данной работе предлагается следующий метод:

1) Пусть для рассматриваемого объекта x определены $C_a(x)$ и $(C_{a1}, ..., C_{am})$.
2) Определяется модель ансамбля со схожим решением

$$i = \arg\min_{j=1...m}(|C_{nn}(x) - C_{aj}|). \quad (1)$$

И краевые случаи

$$i_{\min} = \arg\min_{j=1...m} C_{aj}, \quad (2)$$

$$i_{\max} = \arg\max_{j=1...m} C_{aj}. \quad (3)$$

3) По определенным интерпретируемым моделям 1, 2, 3 формируется обоснование.

Опишем пункты 2 и 3 подробнее.

3.1. Сравнение степеней уверенности

Ключевым элементом предлагаемого в работе подхода является сравнение выходных значений моделей классификации. В общем случае выходное значение C определяет степень уверенности классификатора в принадлежности объекта к заданному классу. Как правило, к классу с меткой 1. Зададим для этого степень уверенности следующим образом:

$$\gamma(x|y) = \begin{cases} C(x), & y = 1 \\ 1 - C(x), & y = 0 \end{cases},$$

где x — признаковое описание объекта, y — класс. Здесь и далее рассматривается случай бинарной классифик;ции.

Если γ близка к 1, то это значение можно интерпретировать, как высокую уверенность модели в принадлежности объекта к классу y. Если же степень уверенности близка к 0,5, то объект является трудно определимым для модели и проявляет характеристики обоих классов.

Имея несколько моделей классификации, предполагается возможным сравнение их степеней уверенности в определении класса рассматриваемого объекта. Пусть γ_{nn} – степень уверенности для модели нейронной сети, а

$$Y = \{\gamma_1, \gamma_2, ..., \gamma_m\} -$$

степени уверенности для моделей ансамбля классификатора в принадлежности объекта к классу 1. Тогда наиболее согласованной моделью ансамбля будет та, степень уверенности которой ближе к γ_{nn}

$$i = \arg \min_{j=1\ldots m} (|\gamma_{nn} - Y_j|). \tag{4}$$

На основе данного сравнения становится возможным выявление наиболее адекватного обоснования, если предполагать, что модель нейронной сети является наиболее точной.

Однако степени уверенности не всегда являются сопоставимыми из-за излишней уверенности (**overconfidence**) моделей [12]. Это явление, при котором значения $\hat{\gamma}$ почти всюду находятся в окрестностях чисел 0 или 1 (рис. 3).

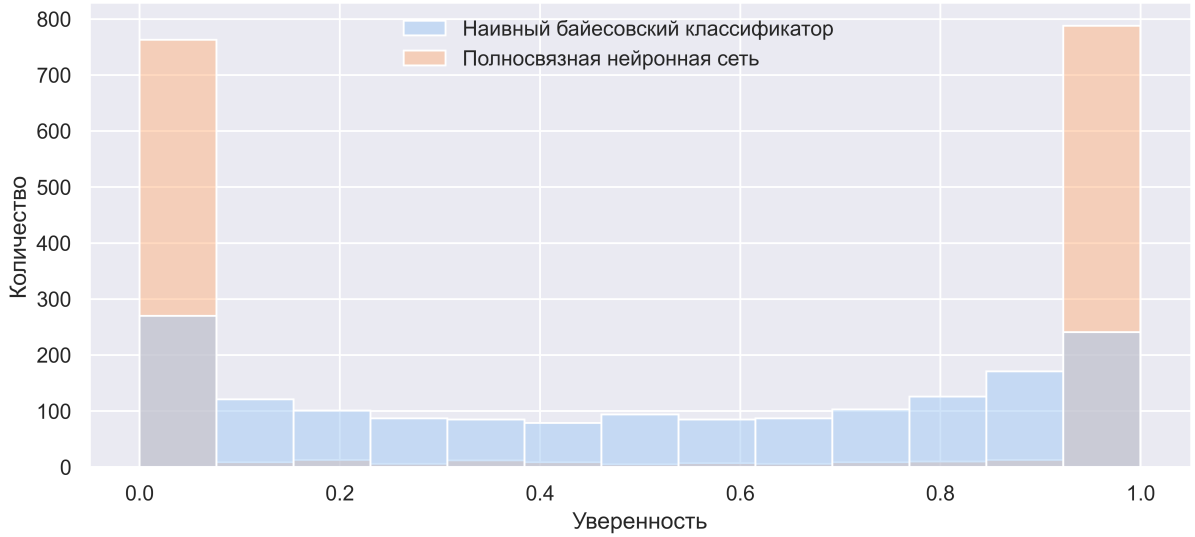


Рис. 3. Распределение параметра γ для наивного байесовского классификатора и полносвязной нейронной сети. Видно, что у байесовского классификатора более равномерное распределение по степени уверенности. Если применять формулу 4, то всегда будут выбираться такие же излишне уверенные модели. Модели обучались и тестировались на синтетическом наборе данных [32]

Чтобы модели были сравнимы, необходимо смягчить степень уверенности моделей, данный процесс называется калибровкой. Существует множество стратегий калибровки, приведем здесь лишь несколько [8; 22]:

- Логит-преобразование с уменьшением уверенности. Если функцией активации нейронной сети является сигмоида, можно применить логит-функцию для перевода вероятности в пространство логит-модели

$$\text{logit}(\gamma) = \log \frac{\gamma}{1-\gamma}.$$

Затем применяется обратное преобразование, используя функцию сигмоиды с коэффициентом α, уменьшающим уверенность модели

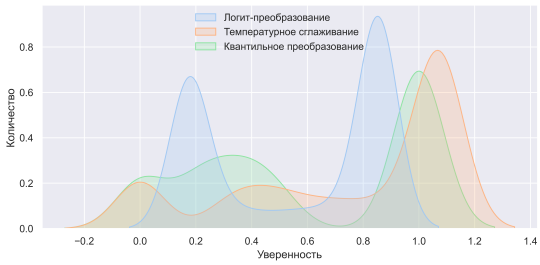
$$\gamma' = \sigma(\alpha * \text{logit}(\gamma)).$$

- Температурное сглаживание. Если функцией активации является функция softmax, аналогичным образом можно применить обратную функцию с коэффициентом α, сглаживающим распределение γ

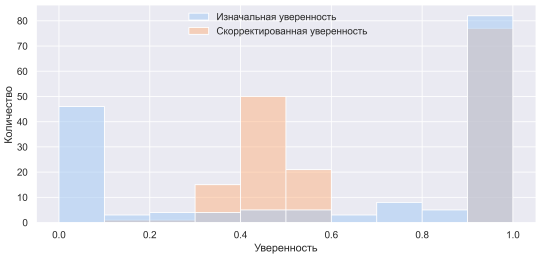
$$\gamma' = \frac{\exp(\gamma(x,1)/\alpha)}{\sum_{j\in\{0,1\}}\exp(\gamma(x|j)/\alpha)}.$$

- Квантильное преобразование. Метод делит распределение величины γ на квантили и отображает их на интервал соответствующим заданному распределению образу.

На рисунке 4(a) приведены сравнения данных методов, применяемых к распределению величины γ_{nn} из рисунка 3. Подобные преобразования позволяют сделать результаты модели более адекватными. Так, ошибочные решения более близки к состоянию «неуверенности» (рис. 4(b)). Синим показаны значения уверенности модели нейронной сети на объектах, определяемых некорректно. Практически всегда модель присваивает им значение 0 или 1. После калибровки уверенность приближается к значению 0,5.



(a) Распределения, полученные в результате калибровки описанными методами



(b) Неуверенность модели

Рис. 4. Калибровка уверенности модели

3.2. Формирование обоснования на основе информативных областей

Формирование обоснования по выбранной интерпретируемой модели машинного обучения не представляет сложности из-за прозрачности механизмов принятия решений. Однако актуальным остается вопрос качества обоснования, а именно его простота и адекватность для конечного пользователя. В первую очередь обозначенные критерии зависят от формируемого признакового пространства. Чем полнее каждый из признаков описывает характеристические особенности классов, тем выше его информативность при обосновании.

Одним из подходов обоснования является сравнение значения признака объекта с нормой у объектов основного класса, то есть такого, от которого происходит отделение остальных [28]. В таком случае значения признака разбиваются на интервалы, определяющие степень отклонения признака от нормы (значение признака выше или ниже нормы, незначительно или значительно). В работах [5; 19] описан подход, который заключается в определении информативной области нормы. В данной работе предлагается подход, развивающий его за счет определения информативных областей обоих классов, в которых объекты в высокой степени отделимы от остальных.

Для этого определим высокоинформативный признак

$$F = (f, I, W),$$

где f – функция; $I = (I_1, I_2)$ – вектор информативных зон для каждого из классов; W – агрегированная информативность признака. В свою очередь, I_n – интервал, либо вектор интервалов, определяющих области характерных значений класса n (рис. 5). В рамках данной предметной области I_n – интервал.

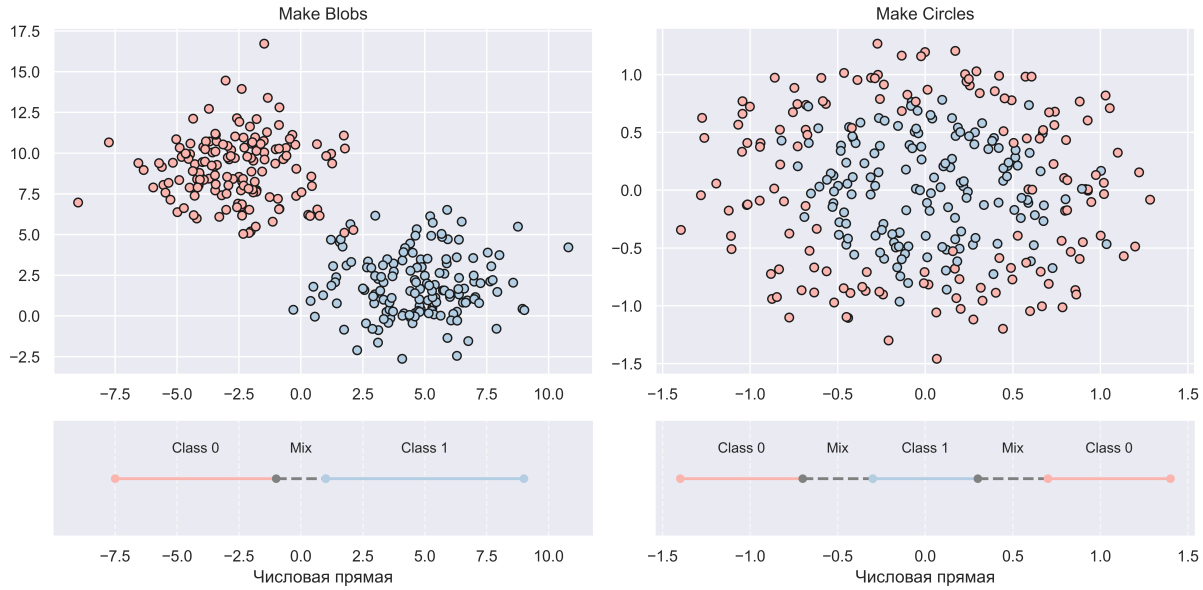


Рис. 5. Информативные зоны. Слева демонстрируется пример, в котором по оси абсцисс каждый из классов характеризуется одним интервалом. Справа демонстрируется пример, в котором объекты класса 0 характеризуются двумя интервалами. Наборы данных являются синтетическими [32]

На основе вектора I определим четыре области обоснования:

1) I_{12} – интервал, характерный для смешанных данных (пунктирная серая область на рисунке 5). Если в результате построения интервалы I_1 и I_2 пересекаются, то I_{12} будет являться продуктом их пересечения

$$I_{12} = I_1 \cap I_2,$$

а сами интервалы уменьшаются до границ пересечения

$$I_1 = I_1 \setminus I_2, \quad I_2 = I_2 \setminus I_1.$$

2) I_1 – область, характерная для основного класса.
3) I_2 – область, характерная для второго класса.
4) $Outlier = \mathbb{R} \setminus [\min(I_{1,l}, I_{2,l}), \max(I_{1,r}, I_{2,r})]$ – области, определяющие нехарактерные данные, то есть выбросы. Здесь $I_{n,l}$ – левая граница интервала; $I_{n,r}$ – правая граница интервала.

Определим значения признака f^*

$$f^* = \begin{cases} 0, f \in I_1, \\ 0.5, f \in I_{12}, \\ 1, f \in I_2, \\ 2, f \in Outlier\,. \end{cases}$$

На основе данных категорий упрощается процесс обоснования. Так, зная, за какую из характеристик отвечает признак, возможно оценить степень характерности и его влияния на принимаемое решение. Если значение признака равно нулю или единице, признак влияет в высокой степени, если 0,5 или 2 – в меньшей из-за нехарактерности. Также такая категоризация признака повышает точность классификации относительно бинаризации. В таблице 3 представлены результаты бинарной классификации по бинарным термометрическим признакам и по категориальным. Результаты оценивались по трем метрикам:

- Специфичность – доля верноопределенных здоровых молочных желез.
- Чувствительность – доля верноопределенных молочных желез группы риска.
- Эффективность – среднегеометрическое из специфичности и чувствительности.

Таблица 3

Результаты классификации

Признаки	Специфичность	Чувствительность	Эффективность
Бинарные	0,945	0,693	0,809
Категориальные	0,946	0,738	0,833

4. Адекватность объяснения

В настоящее время методы корректировки шумов в признаках широко исследованы, однако актуальной остается задача корректировки шумов в метках. При поиске ошибочных меток используются методы, определяющие степень характерности признакового описания объекта соответствующей метке [15]. Однако в них предлагается удаление объектов с нехарактерными метками, что негативно влияет на качество данных в условиях их малого объема. Поэтому для медицинских данных малого объема имеет смысл изменить метку на ту, чей класс наиболее характерен данному признаковому описанию. В качестве нахождения наиболее характерной метки можно использовать как многоклассовый классификатор, так и построенные интервалы. В интервалы какого класса объект чаще всего попадает, тому классу он и наиболее характерен. Такой подход позволяет не только исправлять нехарактерные метки, но и уточнять полуопределенные (см. табл. 1). Вычислительные эксперименты показали, что при уточнении полуопределенных объектов с меткой 3 повышается точность его определения на тестовой выборке на 13 %.

Другой описанной проблемой данных малого объема является их слабая репрезентативность – степень, с которой обучающий набор данных отражает характеристики реального распределения объектов. Зная область характерных значений объектов по

каждому признаку, возможно повышать репрезентативность данных, синтезируя объекты в них. Таким образом повышается качество обучающей выборки, что положительно влияет на адекватность математических моделей и их интерпретации.

Заключение

Рассмотренные в работе подходы демонстрируют повышение качества математических моделей и их интерпретации. Сравнение степеней уверенности позволяет выявлять наиболее согласованные с нейросетевыми предсказаниями решения деревьев решений, на основе которых осуществляется интерпретация результатов. При этом предложенный подход позволяет также регулировать степень и сложность интерпретации, привлекая большее количество моделей ансамбля, согласованных с нейронной сетью, или же, наоборот, наиболее несогласованные модели для приведения аргументов в пользу противоположного диагностического решения.

Категоризация термометрических признаков с выделением характерных зон для каждого класса позволила не только улучшить точность классификации, но и сделать интерпретацию более гибкой с определением пограничных случаев. Актуальным остается вопрос о категоризации признаков в многоклассовом случае, так как в таком случае возможно пересечение интервалов по нескольким классам, что приводит к снижению их информативности.

Предложенные методы построения информативных интервалов и корректировки обучающего набора на их основе позволяют повысить качество данных — как за счет устранения некорректных меток, так и за счет улучшения репрезентативности выборки. Это подтверждает значимость не только выбора модели, но и качества данных при построении систем объяснимого ИИ.

ПРИМЕЧАНИЕ

[1] Исследование выполнено за счет гранта Российского научного фонда № 25-21-00330, https://rscf.ru/project/25-21-00330/.

СПИСОК ЛИТЕРАТУРЫ

1. Замечник, Т. В. Управляемый классификатор в диагностике рака молочной железы по данным микроволновой радиотермометрии / Т. В. Замечник, А. Г. Лосев, А. Ю. Петренко // Математическая физика и компьютерное моделирование. — 2019. — № 22 (3). — С. 52–66. — DOI: 10.15688/mpcm.jvolsu.2019.3.5

2. Левшинский, В. В. Математические методы анализа и интерпретации термометрических данных в медицинской диагностике : дис. . . . канд. техн. наук / В. В. Левшинский. — Волгоград, 2022. — 202 с.

3. Лекции по логическим алгоритмам классификации. — URL: http://www.ccas.ru/voron/download/LogicAlgs.pdf

4. Лосев, А. Г. Интеллектуальный анализ данных микроволновой радиотермометрии в диагностике рака молочной железы / А. Г. Лосев, В. В. Левшинский // Математическая физика и компьютерное моделирование. — 2017. — Т. 20, № 5. — С. 49–62. — DOI: 10.15688/mpcm.jvolsu.2017.5.6

5. Лосев, А. Г. Интеллектуальный анализ термометрических данных в диагностике молочных желез / А. Г. Лосев, В. В. Левшинский // Управление большими системами:

сб. тр. — М. : Институт проблем управления им. В.А. Трапезникова РАН, 2017. — № 70. — С. 113–135.

6. Перспективы микроминиатюризации многоканальных многочастотных радиотермографов / А. Г. Гудков, С. Г. Веснин, Ю. В. Соловьев, В. Г. Тихомиров, В. В. Попов // Инфокоммуникационные и радиоэлектронные технологии. — 2022. — № 4. — С. 531–547. — DOI: 10.29039/2587-9936.2022.05.4.38

7. Попов, И. Е. Анализ термометрических данных головного мозга, полученных методом микроволновой радиотермометрии / И. Е. Попов, А. Е. Крылова // Математическая физика и компьютерное моделирование. — 2023. — № 26 (2). — С. 32–42. — DOI: 10.15688/mpcm.jvolsu.2023.2.3

8. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias / B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed // Bioinformatics. — 2003. — № 2. — P. 185–193. — DOI: 10.1093/bioinformatics/19.2.185

9. A Survey of Commonly Used Ensemble-Based Classification Techniques / A. Jurek, Y. Bi, S. Wu, C. Nugent // The Knowledge Engineering Review. — 2014. — № 29 (5). — P. 551–581. — DOI: 10.1017/S0269888913000155

10. A Survey of Methods for Explaining Black Box Models / R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi // ACM Comput. Surv. — 2018. — № 5. — P. 1–42. — DOI: 10.1145/3236009

11. BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations / X. Zhao, W. Huang, X. Huang, V. Robu, D. Flynn // Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence. — 2021. — № 161. — P. 887–896.

12. Don't Just Blame Over-Parametrization for Over-Confidence: Theoretical Analysis of Calibration in Binary Classification / Y. Bai, S. Mei, H. Wang, C. Xiong // ArXiv Preprint. — 2021. — P. 1–31.

13. Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction / M. Mersha, K. Lam, J. Wood, A. K. AlShami, J. Kalita // Neurocomputing. — 2024. — № 599. — P. 128111. — DOI: 10.1016/j.neucom.2024.128111

14. Explaining Anomalies Detected by Autoencoders Using Shapley Additive Explanations / L. Antwarg, R. M. Miller, B. Shapira, L. Rokach // Expert Systems with Applications. — 2021. — № 186. — P. 115736. — DOI: 10.1016/j.eswa.2021.115736

15. Feng, W. Label Noise Cleaning with an Adaptive Ensemble Method Based on Noise Detection Metric / W. Feng, Y. Quan, G. Dauphin // Sensors. — 2020. — № 20. — P. 6718. — DOI: https://doi.org/10.3390/s20236718

16. Krizhevsky, A. ImageNet Classification with Deep Convolutional Neural Networks / A. Krizhevsky, I. Sutskever, G. E. Hinton // Commun. ACM. — 2017. — Vol. 60, № 6. — P. 84–90. — DOI: 10.1145/3065386

17. LeCun, Y. Deep Learning / Y. LeCun, Y. Bengio, G. Hinton // Nature. — 2015. — № 521. — P. 436–444. — DOI: 10.1038/nature14539

18. Levshinskii, V. Using AI and Passive Medical Radiometry for Diagnostics (MWR) of Venous Diseases / V. Levshinskii, A. Losev, T. Zamechnik // Computer Methods and Programs in Biomedicine. — 2022. — № 215. — P. 106611. — DOI: 10.1016/j.cmpb.2021.106611

19. Levshinskii, V. V. Mathematical Models for Analyzing and Interpreting Microwave Radiometry Data in Medical Diagnosis / V. V. Levshinskii // Journal of Computational and Engineering Mathematics. — 2021. — № 1. — P. 3–14. — DOI: 10.14529/jcem210101

20. Lundberg, S. M. A Unified Approach to Interpreting Model Predictions / S. M. Lundberg, S. Lee // Advances in Neural Information Processing Systems. — 2017. — Vol. 30. — P. 1–10. — DOI: arXiv:1705.07874

21. Neural Additive Models: Interpretable Machine Learning with Neural Nets / R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, G. E. Hinton // Advances in Neural Information Processing Systems. — 2021. — Vol. 34. — P. 1–23. — DOI: arXiv:2004.13912

22. On Calibration of Modern Neural Networks / C. Guo, G. Pleiss, Y. Sun,

K. Q. Weinberger // ArXiv Preprint. — 2017. — P. 1–14. — DOI: arXiv:1706.04599

23. Passive Microwave Radiometry for the Diagnosis of Coronavirus Disease 2019 Lung Complications in Kyrgyzstan / B. Osmonov, L. Ovchinnikov, C. Galazis, B. Emilov, M. Karaibragimov, M. Seitov, S. Vesnin, A. Losev, V. Levshinskii, I. Popov // Diagnostics. — 2021. — № 11. — P. 259. — DOI: 10.3390/diagnostics11020259

24. Passive Microwave Radiometry in Biomedical Studies / I. Goryanin, S. Karbainov, O. Shevelev, A. Tarakanov, K. Redpath, S. Vesnin, Y. Ivanov // Drug Discovery Today. — 2020. — № 4. — P. 757–763. — DOI: 10.1016/j.drudis.2020.01.016

25. Priestley, M. A Survey of Data Quality Requirements That Matter in ML Development Pipelines / M. Priestley, F. O'Donnell, E. Simperl // Journal of Data and Information Quality. — 2023. — № 15 (2). — P. 1–39.

26. Ribeiro, M. T. «Why Should I Trust You?»: Explaining the Predictions of Any Classifier / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2016. — P. 1–10. — DOI: 10.1145/2939672.2939778

27. Saeed, W. Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities / W. Saeed, C. Omlin // Knowledge-Based Systems. — 2023. — № 263. — P. 110273. — DOI: https://doi.org/10.1016/j.knosys.2023.110273

28. Some Methods for Substantiating Diagnostic Decisions Made Using Machine Learning Algorithms / A. G. Losev, I. E. Popov, A. Yu. Petrenko, A. G. Gudkov, S. G. Vesnin, S. V. Chizhikov // Biomedical Engineering. — 2022. — № 6. — P. 442. — DOI: 10.1007/s10527-022-10153-y

29. Tjoa, E. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI / E. Tjoa, C. Guan // IEEE Transactions on Neural Networks and Learning Systems. — 2021. — № 11. — P. 4793–4813. — DOI: 10.1109/TNNLS.2020.3027314

30. Towards Compositional Interpretability for XAI / S. Tull, R. Lorenz, S. Clark, I. Khan, B. Coecke // arXiv. — 2024. — P. 1–106. — DOI: 10.48550/arXiv.2406.17583

31. This Looks Like That: Deep Learning for Interpretable Image Recognition / C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. K. Su // Advances in Neural Information Processing Systems. — 2019. — Vol. 32. — P. 1–12. — DOI: arXiv:1806.10574

32. Utilities to Load Popular Datasets and Artificial Data Generators. — URL: https://scikit-learn.org/stable/api/sklearn.datasets.html

33. Vinogradova, K. Towards Interpretable Semantic Segmentation via Gradient-Weighted Class Activation Mapping / K. Vinogradova, A. Dibrov, G. Myers // Proceedings of the AAAI Conference on Artificial Intelligence. — 2020. — Vol. 34, № 10. — P. 13943–13944. — DOI: 10.1609/aaai.v34i10.7244

REFERENCES

1. Zamechnik T.V., Losev A.G., Petrenko A.Yu. Upravlyaemyy klassifikator v diagnostike raka molochnoy zhelezy po dannym mikrovolnovoy radiotermometrii [Supervised Classifier for Breast Cancer Diagnosis Based on Microwave Radiothermometry Data]. *Matematicheskaya fizika i kompyuternoe modelirovanie* [Mathematical Physics and Computer Simulation], 2019, no. 22 (3), pp. 52-66. DOI: 10.15688/mpcm.jvolsu.2019.3.5

2. Levshinskij V. V. *Matematicheskie metody analiza i interpretacii termometricheskih dannyh v medicinskoj diagnostike*: dis. ... kand. tekhn. nauk [Mathematical Methods for Analysis and Interpretation of Thermometric Data in Medical Diagnostics: PhD Thesis]. Volgograd, 2022. 202 p.

3. *Lekcii po logicheskim algoritmam klassifikacii* [Lectures on Logical Classification Algorithms]. URL: http://www.ccas.ru/voron/download/LogicAlgs.pdf

4. Losev A.G., Levshinskiy V.V. Intellektualnyy analiz dannykh mikrovolnovoy radiotermometrii v diagnostike raka molochnoy zhelezy [Data Mining of Microwave Radiometry Data in the Diagnosis of Breast Cancer]. *Matematicheskaya fizika i kompyuternoe*

modelirovanie [Mathematical Physics and Computer Simulation], 2017, vol. 20, no. 5, pp. 49-62. DOI: 10.15688/mpcm.jvolsu.2017.5.6

5. Losev A.G., Levshinskiy V.V. Intellektualnyy analiz termometricheskikh dannykh v diagnostike molochnykh zhelez [The Thermometry Data Mining in the Diagnostics of Mammary Glands]. *Upravlenie bolshimi sistemami: sb. tr.* [Large-Scale Systems Control] Moscow, Institut problem upravleniya im. V.A. Trapeznikova RAN, 2017, no. 70, pp. 113-135.

6. Gudkov A.G., Vesnin S.G., Solovev Yu.V., Tikhomirov V.G., Popov V.V. Perspektivy mikrominiatyurizatsii mnogokanalnykh mnogochastotnykh radiotermografov [Prospects of Microminiaturization of Multichannel Multi-Frequency Radiothermographs]. *Infokommunikatsionnye i radioelektronnye tekhnologii* [Infocommunications and Radio Technologies], 2022, no. 4, pp. 531-547. DOI: 10.29039/2587-9936.2022.05.4.38

7. Popov I.E., Krylova A.E. Analiz termometricheskikh dannykh golovnogo mozga, poluchennykh metodom mikrovolnovoy radiotermometrii [Analysis of Brain Thermometric Data Obtained by Microwave Radiothermometry]. *Matematicheskaya fizika i kompyuternoe modelirovanie* [Large-Scale Systems Control], 2023, no. 26 (2), pp. 32-42. DOI: 10.15688/mpcm.jvolsu.2023.2.3

8. Bolstad B.M., Irizarry R.A., Astrand M., Speed T.P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, 2003, no. 2, pp. 185-193. DOI: 10.1093/bioinformatics/19.2.185

9. Jurek A., Bi Y., Wu S., Nugent C. A Survey of Commonly Used Ensemble-Based Classification Techniques. *The Knowledge Engineering Review*, 2014, no. 29 (5), pp. 551-581. DOI: 10.1017/S0269888913000155

10. Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 2018, no. 5, pp. 1-42. DOI: 10.1145/3236009

11. Zhao X., Huang W., Huang X., Robu V., Flynn D. BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations. *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2021, no. 161, pp. 887-896.

12. Bai Y., Mei S., Wang H., Xiong C. Don't Just Blame Over-Parametrization for Over-Confidence: Theoretical Analysis of Calibration in Binary Classification. *ArXiv Preprint*, 2021, pp. 1-31.

13. Mersha M., Lam K., Wood J., AlShami A.K., Kalita J. Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction. *Neurocomputing*, 2024, no. 599, p. 128111. DOI: 10.1016/j.neucom.2024.128111

14. Antwarg L., Miller R.M., Shapira B., Rokach L. Explaining Anomalies Detected by Autoencoders Using Shapley Additive Explanations. *Expert Systems with Applications*, 2021, no. 186, p. 115736. DOI: 10.1016/j.eswa.2021.115736

15. Feng W., Quan Y., Dauphin G. Label Noise Cleaning with an Adaptive Ensemble Method Based on Noise Detection Metric. *Sensors*, 2020, no. 20, p. 6718. DOI: https://doi.org/10.3390/s20236718

16. Krizhevsky A., Sutskever I., Hinton G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM*, 2017, vol. 60, no. 6, pp. 84-90. DOI: 10.1145/3065386

17. LeCun Y., Bengio Y., Hinton G. Deep Learning. *Nature*, 2015, no. 521, pp. 436-444. DOI: 10.1038/nature14539

18. Levshinskii V., Losev A., Zamechnik T. Using AI and Passive Medical Radiometry for Diagnostics (MWR) of Venous Diseases. *Computer Methods and Programs in Biomedicine*, 2022, no. 215, p. 106611. DOI: 10.1016/j.cmpb.2021.106611

19. Levshinskii V.V. Mathematical Models for Analyzing and Interpreting Microwave Radiometry Data in Medical Diagnosis. *Journal of Computational and Engineering Mathematics*, 2021, no. 1, pp. 3-14. DOI: 10.14529/jcem210101

20. Lundberg S.M., Lee S. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 1-10. DOI: arXiv:1705.07874

21. Agarwal R., Melnick L., Frosst N., Zhang X., Lengerich B., Caruana R., Hinton G.E. Neural Additive Models: Interpretable Machine Learning with Neural Nets. *Advances in Neural*

Information Processing Systems, 2021, vol. 34, pp. 1-23. DOI: arXiv:2004.13912

22. Guo C., Pleiss G., Sun Y., Weinberger K.Q. On Calibration of Modern Neural Networks. *ArXiv Preprint*, 2017, pp. 1-14. DOI: arXiv:1706.04599

23. Osmonov B., Ovchinnikov L., Galazis C., Emilov B., Karaibragimov M., Seitov M., Vesnin S., Losev A., Levshinskii V., Popov I. Passive Microwave Radiometry for the Diagnosis of Coronavirus Disease 2019 Lung Complications in Kyrgyzstan. *Diagnostics*, 2021, no. 11, p. 259. DOI: 10.3390/diagnostics11020259

24. Goryanin I., Karbainov S., Shevelev O., Tarakanov A., Redpath K., Vesnin S., Ivanov Y. Passive Microwave Radiometry in Biomedical Studies. *Drug Discovery Today*, 2020, no. 4, pp. 757-763. DOI: 10.1016/j.drudis.2020.01.016

25. Priestley M., O'Donnell F., Simperl E. A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *Journal of Data and Information Quality*, 2023, no. 15 (2), pp. 1-39.

26. Ribeiro M.T., Singh S., Guestrin C. «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1-10. DOI: 10.1145/2939672.2939778

27. Saeed W., Omlin C. Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. *Knowledge-Based Systems*, 2023, no. 263, p. 110273. DOI: https://doi.org/10.1016/j.knosys.2023.110273

28. Losev A.G., Popov I.E., Petrenko A.Yu., Gudkov A.G., Vesnin S.G., Chizhikov S.V. Some Methods for Substantiating Diagnostic Decisions Made Using Machine Learning Algorithms. *Biomedical Engineering*, 2022, no. 6, p. 442. DOI: 10.1007/s10527-022-10153-y

29. Tjoa E., Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, no. 11, pp. 4793-4813. DOI: 10.1109/TNNLS.2020.3027314

30. Tull S., Lorenz R., Clark S., Khan I., Coecke B. Towards Compositional Interpretability for XAI. *arXiv*, 2024, pp. 1-106. DOI: 10.48550/arXiv.2406.17583

31. Chen C., Li O., Tao D., Barnett A., Rudin C., Su J.K. This Looks Like That: Deep Learning for Interpretable Image Recognition. *Advances in Neural Information Processing Systems*, 2019, vol. 32, pp. 1-12. DOI: arXiv:1806.10574

32. *Utilities to Load Popular Datasets and Artificial Data Generators*. URL: https://scikit-learn.org/stable/api/sklearn.datasets.html

33. Vinogradova K., Dibrov A., Myers G. Towards Interpretable Semantic Segmentation Via Gradient-Weighted Class Activation Mapping. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 10, pp. 13943-13944. DOI: 10.1609/aaai.v34i10.7244

INTERPRETATION OF MATHEMATICAL MODELS BASED ON MICROWAVE RADIOTHERMOMETRY DATA

Illarion E. Popov

Postgraduate Student, Department of Mathematical Analysis and Function Theory,
Volgograd State University
popov.larion@volsu.ru
https://orcid.org/0000-0002-0997-8721
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Abstract. The article considers the problem of increasing the interpretability of solutions of mathematical models while maintaining high prediction accuracy. The main attention is paid to the integration of highly accurate neural networks with an ensemble of interpretable models whose mechanisms are transparent and amenable to analytical description. A method for forming a justification based on ensemble decisions consistent with the prediction of the neural network is proposed. The agreement is achieved by comparing the confidence levels of the models after preliminary calibration, the need for which is due to the overconfidence effect characteristic of some machine learning models. A method has been developed for selecting interpretable models of an ensemble of classifiers whose estimates on a specific object are as close as possible in terms of confidence to the output of the neural network. This allows forming justifications containing both arguments in favor of the decision made and possible alternative opinions. To increase the flexibility of interpretation, an extended definition of a highly informative feature has been introduced, including categorization of values by the degree of their characteristic for different classes. It is shown that the transition from binary to categorical features contributes to the growth of classification accuracy and increases its overall efficiency. Additionally, a method for constructing informative intervals of features has been developed, which allows increasing their informativeness – separating ability. Based on the obtained intervals, algorithms for refining semi-definite labels and correcting the training sample in order to improve its quality and representativeness have been proposed. The proposed approaches have been tested on the problem of breast cancer diagnostics using microwave radiothermometry data. The results of computational experiments confirm that the use of categorical interpretable features in combination with model calibration allows for a significant increase in classification accuracy and the validity of decisions made.

Key words: interpretation, mathematical model, machine learning, dataset, classification.