Оценка перспектив использования методов машинного обучения при решении задачи мониторинга выведения тераностических флуоресцентных нанокомпозитов из организма*

***О. Э. Сарманова*[1], *С. А. Буриков*[1,2], *С. А. Доленко*[2], *И. В. Исаев*[2], *В. А. Светлов*[2], *К. А. Лаптинский*[1,2], *Т. А. Доленко*[1,2]**

helga-sharman@rambler.ru

[1]Физический факультет МГУ им. М. В. Ломоносова, Россия, г. Москва, Ленинские горы, 1/2

[2]НИИ ядерной физики им. Д. В. Скобельцына МГУ им. М. В. Ломоносова, Россия, г. Москва, Ленинские горы, 1/2

Представлены результаты оценки перспектив применения методов машинного обучения для разработки мониторинга в человеческой урине выведенных из организма тераностических нанокомпозитов и их компонентов по спектрам их флуоресценции. Решалась задача определения в урине компонентов нанокомпозитов — флуоресцирующих углеродных точек (УТ), покрытых сополимером (СП) и лигандами фолиевой кислоты (ФК). Задача решалась в рамках двух подходов — как задача кластеризации (с использованием алгоритма k-means и разрабатываемого авторами алгоритма адаптивного построения иерархических нейросетевых классификаторов) с анализом состава полученных кластеров и как задача классификации. Ни одно из полученных разбиений на кластеры не продемонстрировало явно выраженной чувствительности или специфичности по отношению к типам содержащихся в суспензии наночастиц, что привело к необходимости использовать обучение с учителем (классификацию данных). При этом использовался набор различных архитектур нейронных сетей и 4 альтернативные процедуры отбора существенных входных признаков: по кросс-корреляции, по кросс-энтропии, по стандартному отклонению и с помощью анализа весов нейронной сети. Наилучшие результаты решения задачи классификации нанокомпозитов и их компонентов в урине обеспечивает персептрон с 8 нейронами в единственном скрытом слое, натренированный на наборе существенных входных признаков, выделенных с помощью кросс-корреляции. Точность распознавания, усредненная по всем 5 классам, составила 72,3%.

Ключевые слова: *распознавание образов; кластеризация; искусственные нейронные сети; классификация; углеродные нанокомпозиты; флуоресцентная спектроскопия*

DOI: 10.21469/22233792.3.4.01

1 Введение

В настоящее время в наномедицине чрезвычайно актуальным является создание принципиально новых наноматериалов, которые могут использоваться одновременно и для

*Работы по решению задачи кластеризации были выполнены при финансовой поддержке РФФИ в рамках научного проекта № 15-07-08975-а (С.Д., В.С.). Все остальные работы выполнены за счет гранта Российского научного фонда (проект № 17-12-01481) — постановка и проведение экспериментов, решение задачи классификации (О.С., С.Б., И.И., К.Л., Т.Д.). Авторы благодарны О. А. Шендеровой (Adamas Nanotechnologies, Inc., США), Джессике Розенхольм, Еве фон Хаартман и Дидем Сен Караман (Университет Або Академии Финляндии, г. Турку) за синтез углеродных точек и нанокомпозитов на их основе.

диагностики, и для лечения заболеваний [1–4]. Такие тераностические агенты могут одновременно выполнять в организме следующие функции: (1) по изменению своих флуоресцентных свойств «показывать» больные ткани; (2) после загрузки их поверхности лекарственными препаратами осуществлять адресную доставку этих лекарств, препятствуя попаданию их в здоровые ткани; (3) контролировать локализацию лекарств в организме на клеточном уровне по спектрам флуоресценции агентов. В отличие от обычного введения лекарственного препарата и его распространения по всему организму направленная доставка позволяет снизить дозу вводимого лекарства и минимизировать его воздействие на другие клетки (побочное действие) [2, 4]. При агрессивной терапии опухолей аспект адресной доставки высокотоксичных онкологических препаратов приобретает особое значение [3, 4].

Современные носители лекарственных препаратов представляют собой многослойные композиты, состоящие из флуоресцентной наночастицы-носителя, модифицированного полимером, к которому прикреплено лекарство. Благодаря способности к стабильной флуоресценции, возможности целенаправленной функционализации поверхности и иммобилизации лекарственных средств на ней, нетоксичности и высокой биосовместимости, углеродные наночастицы лучше многих других наночастиц подходят для таких применений в наномедицине [5–7]. При разработке таких тераностических наноагентов необходимо уделять внимание контролю над их выведением из организма, причем выводиться могут не только носители, но и компоненты нанокомпозитов. Поскольку сами носители и их компоненты обладают интенсивными флуоресцентными свойствами, оптимальным методом их визуализации в биотканях является распознавание наночастиц по спектрам флуоресценции. Такой метод контроля является неразрушающим и экспрессным.

В данной работе предлагается новый метод распознавания и классификации в человеческой урине выведенных нанокомпозитов и их компонентов. Распознавание осуществляется по спектрам флуоресценции с помощью методов машинного обучения. Сложность поставленной задачи заключается в том, что спектры флуоресценции углеродных нанокомпозитов и их компонентов существенно перекрываются со спектрами флуоресценции естественных флуорофоров биологических тканей, т. е. с аутофлуоресценцией [8]. Спектр аутофлуоресценции ткани является результатом наложения полос флуоресценции большого количества естественных флуорофоров и занимает диапазон от 250 до 700 нм. В этом же диапазоне расположены полосы флуоресценции углеродных частиц (рис. 1) [9], поэтому задача оптической визуализации углеродных наночастиц в биоткани заключается в разработке эффективных методов выделения их флуоресценции на фоне аутофлуоресценции биообъекта. Для решения таких многопараметрических обратных задач лазерной спектроскопии нами используются искусственные нейронные сети (ИНС) [10], которые зарекомендовали себя как мощный инструмент решения разнообразных задач распознавания образов и анализа данных, в том числе при решении спектроскопических задач [9, 11–16].

Ранее подобные задачи были успешно решены нами с применением ИНС для распознавания наноалмазов и УТ в яичном белке [9] и в человеческой урине [11]. Было показано, что ИНС позволяют распознавать флуоресценцию детонационных наноалмазов и УТ на фоне собственной флуоресценции яичного белка и урины и определять концентрацию наноалмазов в этих биообъектах с достаточно высокой точностью — не хуже 2 мкг/мл [16], а УТ в яичном белке — с точностью 4 мкг/мл [9].

В данной работе с помощью алгоритмов машинного обучения разрабатывается метод оптической визуализации/распознавания в биоткани новых тераностических наноагентов,

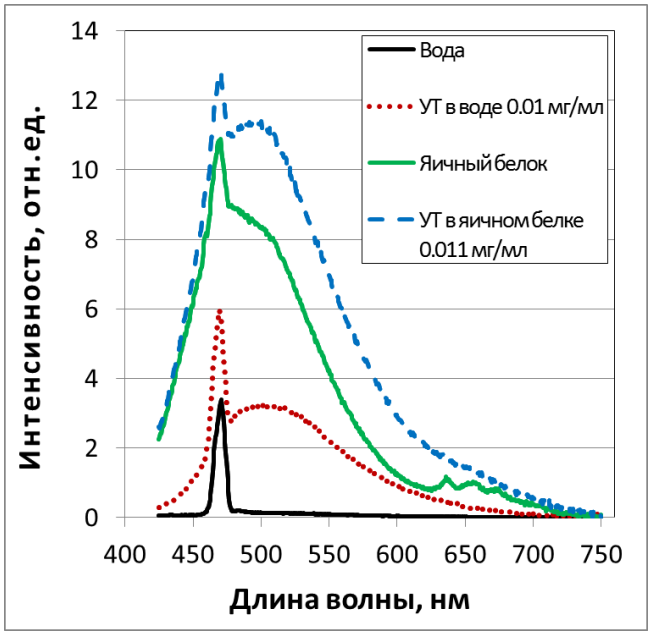

Рис. 1 Спектры флуоресценции и комбинационного рассеяния воды, яичного белка, водной суспензии УТ и яичного белка, содержащего УТ [9]

которые одновременно могут использоваться в качестве флуоресцентных маркеров и носителей лекарственных препаратов. Эти агенты представляют собой нанокомпозиты, состоящие из УТ, покрытых СП и лигандами ФК [7]. Сложность задачи состоит в том, что из организма выводятся не только нанокомпозиты целиком, но и их компоненты. Спектры всех этих компонентов перекрываются со спектром аутофлуоресценции. Описанная задача распознавания решалась в рамках двух подходов — как задача кластеризации с анализом состава полученных кластеров и как задача многозначной классификации (посредством обучения ИНС) спектров флуоресценции в 5 классов в соответствии с количеством возможных выведенных с уриной компонентов нанокомпозитов.

Ввиду новизны работы, дороговизны нанокомпозитов и трудоемкости эксперимента данная работа имеет разведочный характер. Таким образом, настоящая работа преследует следующие цели:

1. Показать принципиальную возможность использования спектроскопических методов для решения задачи мониторинга выведения тераностических флоуресцентных нанокомпозитов из организма.
2. Показать принципиальную возможность и перспективность применения методов машинного обучения для обработки таких данных.
3. Определить сложности, возникающие при решении данной задачи, и наметить пути их преодоления.
4. Воспользовавшись полученной информацией, спланировать последующие более масштабные эксперименты.

2 Эксперимент

В работе использовались нанокомпозиты УТ+СП+ФК — УТ, покрытые СП полиэтиленгликолем и полиэтиленимином и прикрепленными к нему лигандами ФК [7]. Как известно, ФК необходима организму для развития и роста новых клеток, при этом растут

и онкологические клетки [17]. В опухоли происходит экспрессия рецепторов фолатов, в результате чего клетки опухоли активно «забирают» из организма свободную ФК, которую используют для своего роста [17]. В связи с этим в качестве терапии используют такие лиганды ФК, которые блокируют экспрессию рецепторов фолатов в опухоли и прекращают отток ФК из организма на рост онкологических клеток. Именно такие лиганды были прикреплены к поверхности нанокомпозитов [7]. При введении таких нанокомпозитов в организм возможны следующие ситуации: (1) нанокомпозит не отдал лекарство и выводится в неизменном виде; (2) нанокомпозит отдал со своей поверхности лиганды ФК, и выводится компонент УТ+СП; (3) от нанокомпозита отделились лекарство и СП, выводится УТ; (4) выводятся отдельно отделившийся СП и излишек ФК. Таким образом, в урине возможно наличие следующих 5 классов веществ: УТ+СП+ФК, УТ+СП, УТ, СП и ФК, которые могут образовывать $2^5 = 32$ возможных сочетаний с одновременным присутствием в исследуемом объекте от 0 до всех 5 компонентов.

В работе моделировались образцы со всеми 32 возможными сочетаниями указанных классов нанокомпозитов и их компонентов в уринах от трех различных доноров в возрасте от 18 до 25 лет. Были приготовлены суспензии всех сочетаний компонентов в урине в диапазоне концентраций каждого компонента от 2,1 до 2,7 мг/л.

Экспериментально были получены спектры флуоресценции всех приготовленных суспензий нанокомпозитов и их компонентов в урине. Для возбуждения флуоресценции использовался диодный лазер с длиной волны 405 нм. Система регистрации состояла из монохроматора Acton (решетка 1800 штрихов/мм, фокусное расстояние 500 мм) и фотоэлектронного умножителя (Hamamatsu, H-8259-01). Спектры флуоресценции регистрировались в 341 канале в диапазоне 410–750 нм. Обработка спектров заключалась в вычитании пьедестала, обусловленного упругим рассеянием света, и нормировке спектров на площадь валентных колебаний OH-групп комбинационного рассеяния (КР) воды. На рис. 2 приведены экспериментально полученные спектры флуоресценции и КР урины и суспензий

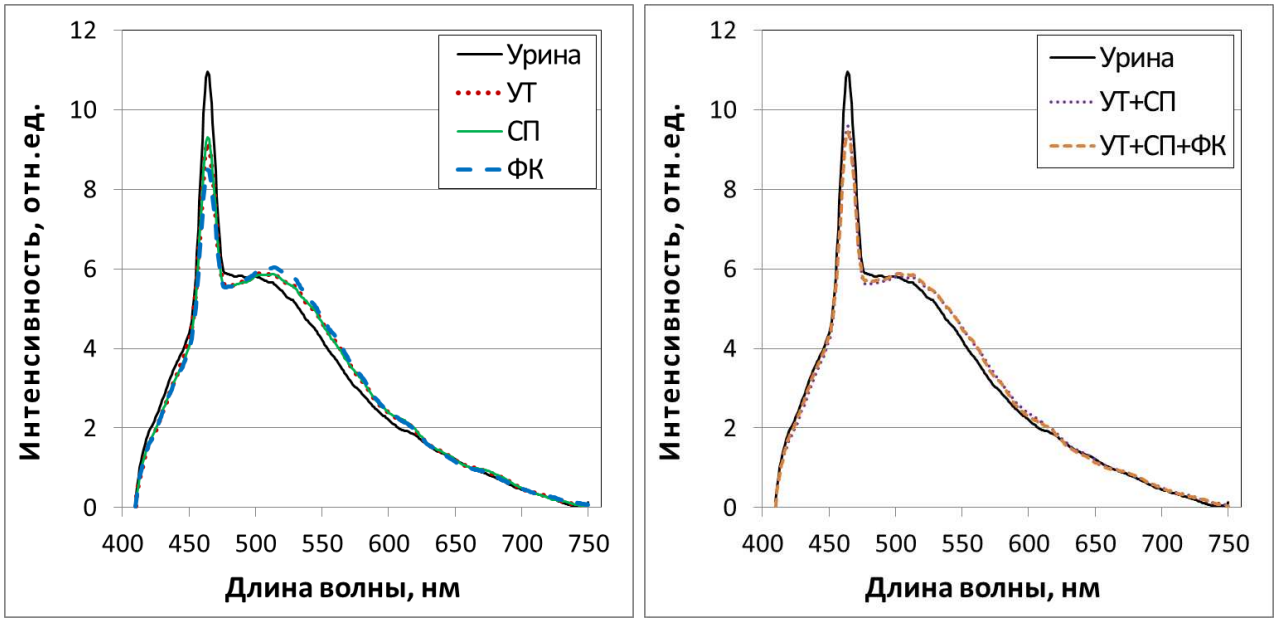

Рис. 2 Спектры флуоресценции урины и суспензий нанокомпозитов и их компонентов в урине. Концентрация 2,7 мг/л

всех рассматриваемых классов компонентов. В общий массив спектров входили суспензии нанокомпозитов и их компонентов в трех различных образцах урины, а также спектры дистиллированной воды и урины без наночастиц — сразу и спустя 6,5 ч после отбора пробы. Всего было получено 373 спектра флуоресценции — три серии для одних и тех же сочетаний компонентов с одними и теми же концентрациями в уринах от трех доноров.

Таким образом, исходный массив данных состоял из 373 примеров, которые описывались 5 параметрами и 341 входным признаком. Каждый параметр принимал значение 0 или 1, которое соответствовало отсутствию или наличию соответствующего компонента в суспензии. Значение каждого признака равнялось интенсивности в соответствующем канале спектра, прошедшего описанную выше предобработку.

3 Алгоритмы анализа данных и результаты их применения

3.1 Применение алгоритмов кластеризации

На исходном массиве данных проводилась кластеризация с целью соотнести результаты, полученные после кластерного анализа, с априорно известными классами. Ожидалось получить чувствительность или специфичность по отношению к типам содержащихся в суспензии наночастиц или количеству этих типов. Была проведена кластеризация исследуемого массива спектров на 4 кластера тремя способами: разрабатываемым авторами алгоритмом адаптивного построения иерархических нейросетевых классификаторов (ИНК) [18], используемым в режиме кластеризации с двумя слегка отличающимися комплектами параметров, и стандартным алгоритмом кластеризации k-средних (k-means).

Алгоритм адаптивного построения ИНК основан на использовании для решения задачи множественной классификации (multiple classification) персептрона с единственным скрытым слоем, содержащим небольшое количество нейронов («мелкого» и «узкого»). Первоначально данная архитектура обучается с учителем на распознавание всех требуемых классов, каждому из которых соответствует один из выходов сети, а желаемым ответом является единица на «своем» выходе и нули на всех остальных (как это часто делается при нейросетевом решении задач классификации). Спустя некоторое количество эпох обучения тренировка останавливается и производится анализ ответов сети на тренировочном наборе. Ввиду заведомой простоты архитектуры нейронная сеть оказывается неспособной правильно распознать все требуемые классы, и на значительном количестве примеров максимальная амплитуда на выходе, хотя и превышает требуемое пороговое значение, соответствует не тому классу i, к которому относится данный пример, а некоторому другому классу k. Если такая картина наблюдается более чем для половины примеров данного класса i, то производится модификация желаемого ответа: для всех примеров этого класса желаемым ответом будет теперь отнесение примера к классу k, т. е. классы i и k оказываются слитыми в один класс. После этого обучение продолжается с модифицированным массивом желаемых ответов, а цикл «обучение – анализ статистики ответов сети – модификация желаемого выхода» повторяется многократно, пока слияние классов не прекратится, а статистика ответов сети не перестанет улучшаться в процессе обучения.

Таким образом, получается система с положительной обратной связью, которая быстро формирует небольшое количество групп классов с высокой долей правильного распознавания. Дальнейшее распознавание классов, попавших в одну группу, производится на последующих этапах рекурсивным применением того же алгоритма. Построение иерархического дерева заканчивается, когда все классы перестают объединяться и оказываются распознанными. Более подробно алгоритм построения ИНК, процедура его применения и его особенности описаны в [18].

Таблица 1 Результаты кластеризации массива спектров разрабатываемым алгоритмом ИНК и алгоритмом k-means

			Серия						Количество типов наночастиц						Тип нанокомпонента				
	Всего	373	123	125	125	3	9	3	15	59	120	119	48	12	192	178	180	179	179
Алгоритм	Кластер	Кол-во	1	2	3	У0	У6	Д	0	1	2	3	4	5	СП	УТ	ФК	УТ + СП	УТ + СП + ФК
ИНК №1	1	303	122	58	123	2	9	0	11	49	97	93	41	12	152	148	145	149	152
	2	23	0	22	1	0	0	0	0	1	6	13	3	0	18	7	15	10	14
	3	3	1	1	1	0	0	3	3	0	0	0	0	0	0	0	0	0	0
	4	44	0	44	0	1	0	0	1	9	17	13	4	0	22	23	20	20	13
ИНК №2	1	8	0	8	0	0	0	0	0	1	3	2	2	0	3	7	2	4	5
	2	63	1	60	2	1	0	3	4	9	19	25	6	0	37	28	31	25	25
	3	31	0	31	0	0	0	0	0	5	10	12	4	0	13	14	17	18	15
	4	271	122	26	123	2	9	0	11	44	88	80	36	12	139	129	130	132	134
k-means	1	87	46	0	41	0	2	0	2	7	16	31	23	8	49	60	48	50	57
	2	126	0	120	6	1	3	0	4	22	42	41	16	1	63	58	61	58	58
	3	3	1	1	1	0	0	3	3	0	0	0	0	0	0	0	0	0	0
	4	157	76	4	77	2	4	0	6	30	62	47	9	3	80	60	71	71	64

Помещая первоначально каждый пример или небольшую группу примеров в свой собственный исходный «класс», получаем, что процедура обучения корневого узла ИНК фактически представляет собой работу специфического алгоритма кластеризации, объединяющего исходные классы в небольшое количество групп. Отметим, что такая кластеризация основана не на подсчете расстояния между примерами в каком-либо признаковом пространстве, а на схожести примеров друг с другом с точки зрения персептрона, который пытается обучиться их различать.

Результаты кластеризации представлены в табл. 1 и на рис. 3.

В табл. 1 для всего массива данных (верхняя строка), а также для каждого из кластеров каждого из полученных разбиений приведены количества примеров, попавших в следующие категории: (1) по типу среды (Серия) — номер образца урины (1, 2, 3), урина без наночастиц сразу после отбора образца (У0) и спустя 6,5 ч (У6), дистиллированная вода (Д); (2) по количеству различных типов наночастиц, одновременно присутствующих в суспензии (от 0 до 5); (3) по конкретным типам наночастиц, присутствующих в суспензии (СП — сополимер; УТ — углеродные точки без покрытия; ФК — фолиевая кислота; УТ + СП — УТ, покрытые СП; УТ + СП + ФК — УТ, покрытые СП и ФК).

На рис. 3 для одного из разбиений, полученных с помощью алгоритма ИНК, и для разбиения, полученного с помощью k-means, представлено изображение, на котором для каждого из кластеров приведен спектр, соответствующий центроиду этого кластера (яркая сплошная линия); поканальное стандартное отклонение для этого кластера соответствует ширине светлой полосы того же оттенка.

Видно, что сильнее всего от всех других спектров отличаются спектры дистиллированной воды, в которой полностью отсутствует флуоресценция. Все три таких спектра, снятые по одному в каждой серии, мало отличаются друг от друга и сильно от всех остальных спектров; для одного из разбиений, построенных разрабатываемым алгоритмом, и разбиения, построенного алгоритмом k-means, эти спектры выделяются в отдельный клас-

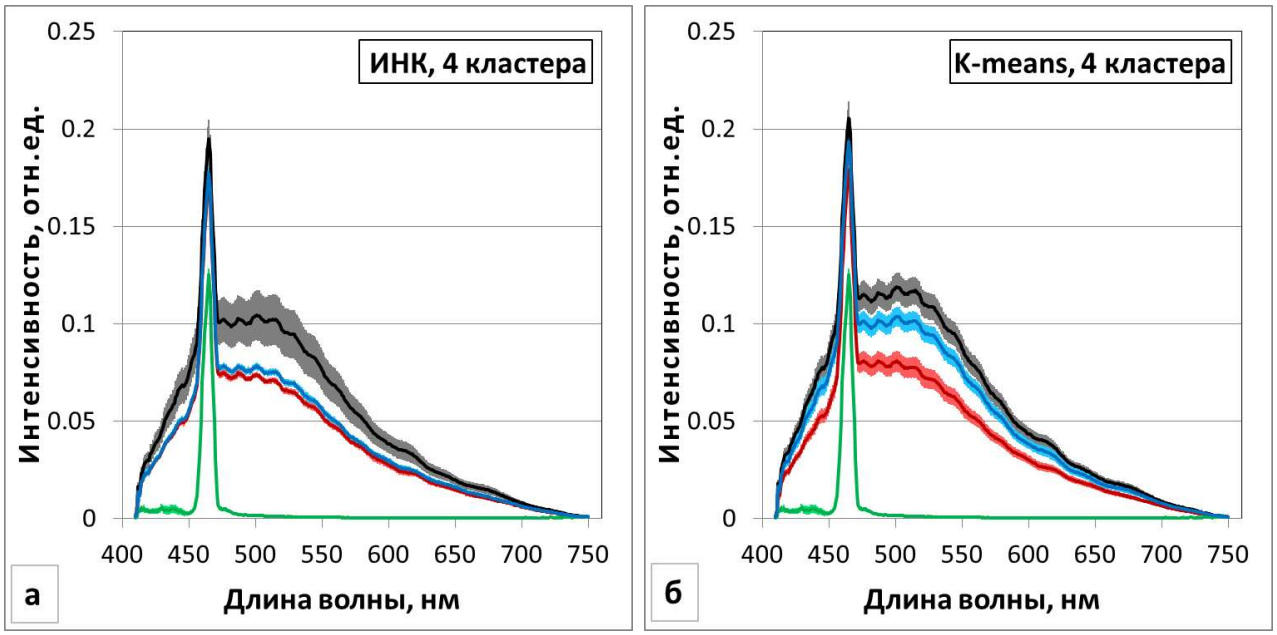

Рис. 3 Спектры-центроиды четырех кластеров, полученных разными алгоритмами (сплошные яркие линии) и стандартное отклонение для каждого кластера (светлая заливка): (*а*) эксперимент ИНК № 1; (*б*) *k*-means

тер, состоящий только из этих примеров (№ 3 в обоих разбиениях) (см. рис. 3, зеленый спектр).

Спектры серий № 1 и № 3 характеризуются более сильной флуоресценцией и обоими вариантами разбиения, построенного разрабатываемым алгоритмом ИНК, помещаются в один кластер (№ 1, черный спектр на рис. 3, *а* для ИНК № 1). Алгоритм *k*-means выделил для спектров первой и третьей серии два кластера, однако в каждом из этих кластеров обе серии представлены практически поровну. Серия № 2 почти полностью помещена алгоритмом *k*-means в отдельный кластер № 2 (красный на рис. 3, *б*), а разрабатываемым алгоритмом ИНК разделена между разными кластерами.

К сожалению, ни одно из рассматриваемых разбиений не продемонстрировало явно выраженной чувствительности или специфичности по отношению к типам содержащихся в суспензии наночастиц или количеству этих типов. Таким образом, можно сделать вывод, что объекты, принадлежащие к одному классу, не образуют локальных сгущений ни в исходном признаковом пространстве, ни в пространстве, полученном после преобразования исходного скрытым слоем ИНК. Это означает, что задачу мониторинга выведения наночастиц из организма следует решать посредством обучения с учителем как задачу классификации и использовать для этого нелинейные методы классификации. В настоящей работе в качестве таких нелинейных методов были использованы ИНС.

Кроме того, поскольку в соответствии с приведенными выше результатами серия № 2 оказалась существенно отличающейся от серий № 1 и № 3, было принято решение оценить возможное качество решения задачи классификации при использовании только серий № 1 и № 3.

3.2 Решение задачи классификации

В соответствии с описанным выше, в настоящей работе решалась задача определения качественного состава суспензии, где каждый компонент может присутствовать или

отсутствовать в исследуемом объекте независимо от остальных. Таким образом, рассматривалась задача многозначной классификации (multilabel classification).

Одним из базовых подходов к решению является сведение ее к задаче множественной классификации, где каждому сочетанию компонентов будет соответствовать отдельный «комбинаторный» класс. Однако в рамках рассматриваемой задачи он практически неприменим ввиду очень малого размера тренировочной выборки.

Другим базовым подходом к ее решению является независимое использование бинарной классификации для каждого компонента. Именно он и применялся в настоящей работе. К использованию этого подхода помимо малого размера тренировочной выборки также располагает тот факт, что для каждого компонента представительность классов «компонент присутствует»/«компонент отсутствует» в тренировочной выборке примерно одинаковая. В рамках такого подхода возможно использование нескольких независимых бинарных классификаторов (по количеству компонентов) или одного классификатора, решающего все рассматриваемые задачи бинарной классификации совместно. В ситуации, когда влияние присутствия отдельных компонентов на спектр не является независимым, второй вариант предпочтительнее. По этой причине в данной работе для совместного решения задач бинарной классификации для нескольких компонентов были выбраны нейронные сети.

Ввиду одинаковой важности правильного установления как наличия, так и отсутствия компонента в суспензии для выставления порога классификатора и в качестве критерия оценки качества решения использовалась точность (accuracy), равная сумме верных положительных и верных отрицательных ответов, отнесенной к общему количеству примеров. Равная представительность классов позволяет получить неискаженный ответ.

Весь массив выбранных спектров флуоресценции (серии №1 и №3) был разбит на тренировочный, валидационный и экзаменационный наборы случайным образом в соотношении 70 : 20 : 10. В результате в тренировочный набор вошли 175 примеров (спектров), в валидационный — 49 примеров, в экзаменационный — 24 примера. Тренировочный набор использовался для подстройки весов сети в процессе обучения, валидационный — для определения момента остановки обучения по минимуму ошибки на этом наборе, экзаменационный — для оценки результатов работы сети на независимых данных.

Ввиду малого размера обучающей выборки использовались архитектуры нейронных сетей, содержащие 1 или 2 скрытых слоя и небольшое количество слоев и нейронов.

3.2.1 Тренировка искусственной нейронной сети на полном наборе входных признаков

Использовались следующие архитектуры многослойных персептронов (МСП):

- с одним скрытым слоем: N01 с 8, 16, 32 и 64 нейронами в скрытом слое;
- с двумя скрытыми слоями: N02 с (8+2), (8+4), (12+3), (16+8), (32+16) нейронами в скрытых слоях.

В работе также использовались нейронные сети с общей регрессией [19] с последовательным поиском параметра сглаживания (НСОР, посл.) и с поиском дополнительных поправок к параметру сглаживания для каждого входного признака генетическим алгоритмом (НСОР, ген.), реализованные в пакете NeuroShell 2 [20]. Во всех случаях тренировались три одинаковые нейронные сети с различными начальными значениями весовых коэффициентов. Результаты их применения усреднялись для того, чтобы уменьшить влияние выбора начальных весовых коэффициентов. Все ИНС имели 5 выходов, соответствующих каждому из 5 классов: УТ+СП+ФК, УТ+СП, УТ, СП и ФК. В качестве желаемого

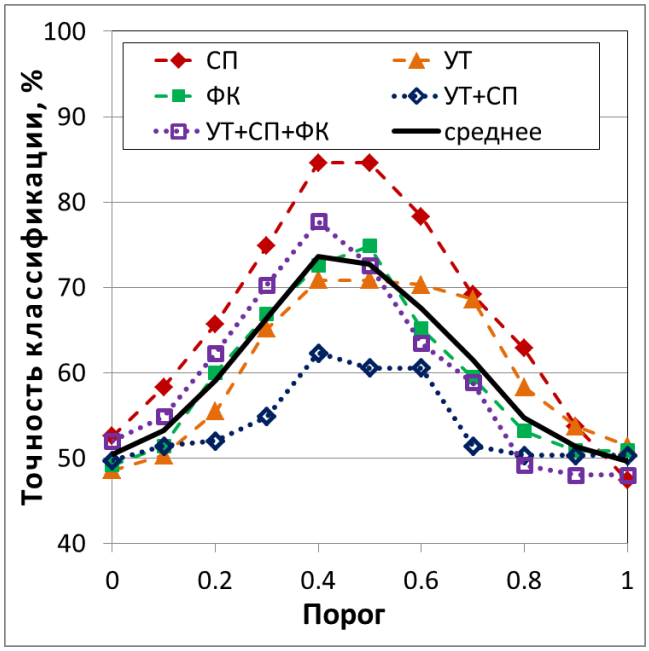

Рис. 4 Зависимость точности классификации на тренировочном наборе от значения порога для каждого из 5 классов отдельно и по всем классам вместе. Результаты получены с помощью персептрона с 8 нейронами в единственном скрытом слое (МСП N01 (8))

Таблица 2 Наилучшие значения точности классификации компонентов нанокомпозитов в урине, полученные использованными архитектурами ИНС

Архитектура, (кол-во нейронов)	СП	УТ	ФК	УТ + СП	УТ + СП + ФК
N01 (64)	76,7%	75,0%	66,7%	64,9%	78,5%
N02 (32+16)	77,5%	72,5%	65,0%	67,7%	79,2%
N01 (32)	77,8%	75,8%	66,7%	64,1%	82,3%
N02 (16+8)	79,2%	69,8%	64,6%	68,8%	79,2%
N01 (16)	75,7%	75,0%	66,7%	64,1%	79,6%
N02 (8+4)	79,2%	70,8%	63,9%	64,6%	75,4%
N01 (8)	**79,2%**	**75,0%**	**69,4%**	**63,5%**	**83,3%**
N02 (8+2)	75,0%	68,1%	58,8%	65,1%	76,8%
N02 (12+3)	79,2%	70,8%	63,5%	70,8%	79,2%
НСОР, посл.	75,0%	75,0%	66,7%	66,7%	79,2%
НСОР, ген.	62,5%	62,5%	62,5%	62,5%	83,3%

ответа сети на выходе, соответствующем каждому компоненту, ожидалось значение 1, если пример соответствовал спектру раствора, в котором данный компонент присутствовал, и значение 0, если отсутствовал. Для обученной нейронной сети значение порога, по которому производилось принятие решения о присутствии компонента, рассчитывалось по тренировочному набору. На рис. 4 приведена зависимость точности классификации от значения порога для каждого из 5 классов отдельно и по всем классам вместе. В качестве значения порога выбиралось такое, на котором точность по всем классам вместе была максимальной.

В табл. 2 представлены наилучшие результаты классификации всех использованных архитектур МСП, натренированных на полном наборе входных признаков.

Как видно из полученных результатов, наилучшую классификацию продемонстрировал персептрон с 8 нейронами в единственном скрытом слое — 67,9% (усредненное значение по всем 5 классам). Ухудшение в ряде случаев качества решения при использовании архитектур с большим количеством слоев или нейронов свидетельствует о недостаточном размере тренировочной выборки. Из анализа данных табл. 2 следует, что все использованные архитектуры ИНС хуже всего решают задачу классификации для ФК и нанокомпозита УТ+СП. Можно предположить, что подобный результат связан с особой формой спектра флуоресценции данных классов.

3.2.2 Тренировка искусственной нейронной сети после отбора существенных входных признаков

Качество обучения и работы ИНС при заданном количестве примеров в обучающей выборке существенно зависит от входной размерности задачи. Очевидно, что не все каналы спектра являются одинаково информативными (см. рис. 2), поэтому уменьшение количества используемых каналов спектра может привести к улучшению нейросетевой аппроксимации искомой зависимости благодаря уменьшению количества весовых коэффициентов и упрощению аппроксимирующей функции (нейронной сети), при этом существенные входные признаки должны отбираться объективным образом, а не вручную. Для уменьшения входной размерности задачи и возможного переучивания ИНС были проведены 4 альтернативные процедуры отбора существенных входных признаков: по кросс-корреляции, по кросс-энтропии, по стандартному отклонению и с помощью анализа весов нейронной сети [21]. Для сравнения эффективности способов отбора признаков при использовании каждого из них путем изменения порога отбора были, по возможности, сформированы наборы с количеством значимых входных признаков около 250, около 150 и около 50. На отобранных наборах входных признаков во всех случаях тренировался персептрон с 8 нейронами в единственном скрытом слое (МСП N01 (8)), продемонстрировавший наилучшие результаты при тренировке ИНС на полном наборе входных признаков (см. п. 3.2.1). Результаты решения задачи классификации компонентов нанокомпозитов в урине, полученные с помощью МСП N01 (8), натренированном на наборах отобранных существенных признаков, представлены на рис. 5.

Кросс-корреляция (рис. 5, *а*). Вычислялись значения кросс-корреляции величин в каждом спектральном канале со значениями каждого выхода. Для каждого выхода значимые входные признаки определялись отдельно, а затем все признаки, значимые хотя бы для одного выхода, использовались для последующей тренировки персептрона. Наилучший результат продемонстрировала ИНС, натренированная на наборе спектров с 252 значимыми признаками, — точность классификации 72,3% на экзаменационном наборе (это значение усреднено по всем классам).

Кросс-энтропия (рис. 5, *б*). Вычислялись значения кросс-энтропии величин в каждом спектральном канале со значениями каждого выхода. Для каждого выхода значимые входные признаки определялись отдельно, а затем все признаки, значимые хотя бы для одного выхода, использовались для последующей тренировки персептрона. В данном случае не удалось подобрать параметры, обеспечивающие выбор 250, 150 и 50 значимых признаков, что обусловлено структурой самих данных, поэтому были натренированы сети на наборах спектров с 318, 85 и 52 входными признаками. Наилучший результат продемонстрировал МСП N01 (8), натренированный на наборе спектров с 318 входными признаками, — точность классификации 69,7% на экзаменационном наборе (это значение усреднено по всем классам).

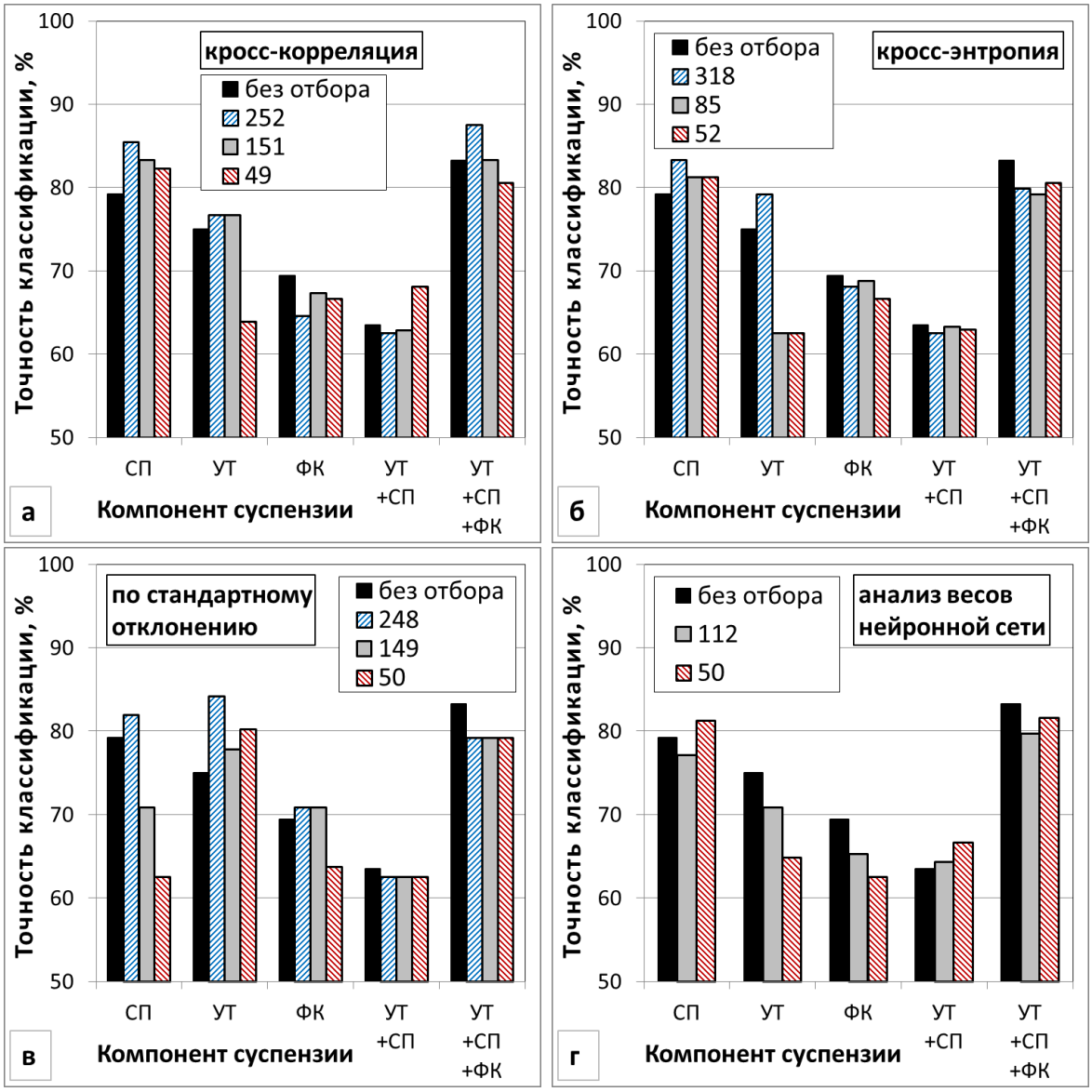

Рис. 5 Точность классификации, полученная МСП N01 (8) с применением процедур отбора признаков: (*а*) кросс-корреляция; (*б*) кросс-энтропия; (*в*) отбор по стандартному отклонению; (*г*) анализ весов нейронной сети

Стандартное отклонение (рис. 5, *в*). Вычислялось значение стандартного отклонения значений в каждом спектральном канале по всем примерам, пропорциональное значению энтропии, т. е. количеству информации в данном канале. Наилучший результат продемонстрировал МСП N01 (8), натренированный на наборе спектров с 248 значимыми признаками, — точность классификации 71,6% на экзаменационном наборе (это значение усреднено по всем классам).

Анализ весов нейронной сети. Этот метод отбора существенных признаков [22] основан на значениях весов нейронных сетей, натренированных на полном наборе данных. Идея метода заключается в том, что существенный входной признак обычно имеет большие значения весовых коэффициентов хотя бы для некоторых из связей, соединяющих его с выходным слоем. Исходя из этого, определяется показатель существенности для каждого входного признака (канала спектра). Для каждого класса веществ определение значимых

Таблица 3 Наилучшие результаты распознавания каждого класса компонентов в урине, полученные с помощью МСП N01 (8) и отбора существенных признаков

Номер класса	Нанокомпозит	Точность распознавания, %	Процедура отбора признаков (количество значимых признаков)
1	СП	85,7	КЭ (318)
2	УТ	84,2	СТО (248)
3	ФК	76,2	КЭ (52)
4	УТ + СП	68,1	КК (49)
5	УТ + СП + ФК	87,5	КК (252)

каналов производилось отдельно следующим образом: на полном наборе данных тренировались 5 идентичных нейронных сетей с 32 нейронами в единственном скрытом слое, различающихся начальными значениями весов. Далее производился анализ весов каждой из полученных нейронных сетей, и для каждой из них вычислялось среднее по всем каналам значение показателя существенности, а также его стандартное отклонение. Если значение показателя существенности в данном канале превышало значение «среднее + $+k$ стандартных отклонений», то канал признавался значимым. Далее, если канал внутри класса был значимым хотя бы для трех сетей из пяти, то канал полагался значимым для класса. Если канал являлся значимым хотя бы для одного класса, то канал признавался значимым и использовался для последующей тренировки ИНС. В зависимости от параметра k можно было варьировать количество значимых каналов.

Наилучший результат продемонстрировал МСП N01 (8), натренированный на наборе спектров со 112 значимыми признаками, — точность классификации 66% на экзаменационном наборе (это значение усреднено по всем классам).

Таким образом, точность классификации компонентов нанокомпозитов в урине в результате обучения ИНС на отобранных существенных признаках повысилась (до 72.3%) по сравнению с обучением на полном наборе признаков (67,9%).

В табл. 3 представлены наилучшие результаты (точность распознавания) каждого класса компонентов в урине с помощью персептрона с 8 нейронами в единственном скрытом слое и указан способ отбора существенных признаков, обеспечивший эти результаты.

Сравнивая полученные результаты классификации компонентов нанокомпозитов в урине для различных подходов, можно сделать вывод, что точность распознавания каждого класса зависит от используемой архитектуры ИНС, а также от процедуры отбора существенных признаков. Например, в зависимости от указанных параметров точность распознавания для класса 1 (СП) изменяется от 62,5% (см. табл. 2) до 85,7% (см. табл. 3). Таким образом, имеется несколько подходов к решению поставленной задачи классификации.

1. Если необходимо быстро идентифицировать все классы, то целесообразно использовать одну архитектуру ИНС, продемонстрировавшую лучший результат, усредненный по всем классам веществ. Улучшить результаты возможно посредством отбора существенных признаков. Именно такой подход был реализован в данной работе. Результаты представлены в табл. 3.
2. Если необходимо распознать все классы с более высокой точностью, то для идентификации каждого класса необходимо использовать свою оптимальную архитектуру

Таблица 4 Наилучшие результаты распознавания каждого класса компонентов в урине

Номер класса	Нанокомпозит	Точность распознавания, %	Архитектура и процедура отбора признаков (количество значимых признаков)
1	СП	87,5	N01 8, КЭ (318)
2	УТ	84,2	N01 8, СТО (248)
3	ФК	76,2	N01 8, КЭ (52)
4	УТ + СП	70,8	N02 12+3, на полном наборе
5	УТ + СП + ФК	87,5	N01 8, К (252)

ИНС и применять свою оптимальную процедуру отбора существенных признаков. Естественно, время, затраченное на получение результата, будет пропорционально количеству классов веществ в исследуемом растворе. Для такого подхода результаты решения нашей задачи представлены в табл. 4.

Как следует из сравнения табл. 3 и 4, точность классификации компонентов нанокомпозитов в урине во втором подходе выше. Однако время решения задачи таким методом больше.

Очевидно, что чрезвычайно важной с точки зрения биомедицины является идентификация нанокомпозитов УТ+СП, УТ+СП+ФК и УТ. Как видно из табл. 3, точность распознавания нанокомпозита УТ+СП+ФК и УТ составляет около 85%, что удовлетворяет требованиям медицины. Несмотря на то что точность распознавания нанокомпозита УТ+СП составляет 70,8%, такая точность определения принадлежности указанного компонента к своему классу также вполне удовлетворяет потребностям современной наномедицины. Отметим, что в настоящее время способы контроля доставки лекарств носителями, их распределения в организме и выведения из организма практически не развиты. Точность идентификации сополимера и ФК является также удовлетворительной в контексте значимости данных классов.

4 Заключение

В настоящей работе была показана принципиальная возможность использования спектроскопических методов для решения задачи мониторинга выведения тераностических флоуресцентных нанокомпозитов из организма, а также принципиальная возможность и перспективность применения нейронных сетей для обработки таких данных.

На основании того, что попытки получить кластеризацию исходного массива данных, в которой разбиение устойчиво соотносится с присутствием или отсутствием в объекте какого-либо из типов исследуемых нанокомпозитов и их компонентов, успехом не увенчались, был сделан вывод о том, что объекты, принадлежащие к одному классу, не образуют локальных сгущений ни в исходном признаковом пространстве, ни в пространстве, полученном после преобразования исходного скрытым слоем ИНК. Это приводит к необходимости решать задачу в режиме обучения с учителем (как задачу классификации) и использовать для этого нелинейные алгоритмы классификации.

Наилучшие результаты решения задачи классификации нанокомпозитов и их компонентов в урине показал персептрон с 8 нейронами в единственном скрытом слое, натренированный на наборе существенных входных признаков, выделенных с помощью кросс-

корреляции. Доля правильного распознавания, усредненная по всем пяти классам, составила 72,3%.

Таким образом, полученные результаты свидетельствуют о перспективности работ в этом направлении. Первоочередной задачей здесь является получение массива данных большего размера, с использованием урины от большего количества разных доноров, для чего планируется новый более масштабный эксперимент. Ожидается, что использование увеличенного массива данных позволит использовать архитектуры больших размеров и повысить качество решения, а также придаст результатам бóльшую статистическую значимость.

Литература

[1] *Kim J., Piao Y., Hyeon T.* Multifunctional nanostructured materials for multimodal imaging, and simultaneous imaging and therapy // Chem. Soc. Rev., 2009. Vol. 38. P. 372–390. doi: 10.1039/B709883A.

[2] *Doane T. L., Burda C.* The unique role of nanoparticles in nanomedicine: Imaging, drug delivery and therapy // Chem. Soc. Rev., 2012. Vol. 41. No. 7. P. 2885–2911. doi: 10.3390/ijms18051102.

[3] *Cheng L., Wang C., Feng L., Yang K., Liu Z.* Functional nanomaterials for phototherapies of cancer // Chem. Rev., 2014. Vol. 114. P. 10869–10939. doi: 10.1021/cr400532z.

[4] *Elgqvist J.* Nanoparticles as theranostic vehicles in experimental and clinical applications — focus on prostate and breast cancer (review) // Int. J. Mol. Sci., 2017. Vol. 18. P. 1102. 53 p. doi: 10.3390/ijms18051102.

[5] *Bartelmess J., Quinn S. J., Giordani S.* Carbon nanomaterials: Multi-functional agents for biomedical fluorescence and Raman imaging // Chem. Soc. Rev., 2015. Vol. 11. P. 4672–4698. doi: 10.1039/C4CS00306C.

[6] *Hong G., Diao S., Antaris A. L., Dai H.* Carbon nanomaterials for biological imaging and nanomedicinal therapy // Chem. Rev., 2015. Vol. 115. No. 19. P. 10816–10906. doi: 10.1021/acs.chemrev.5b00008.

[7] *Prabhakar N., Nareoja T., von Haartman E., Karaman D., Burikov S., Dolenko T., Deguchi T., Mamaeva V., Hanninen P., Vlasov I., Shenderova O., Rosenholm J.M.* Functionalization of graphene oxide nanostructures improves photoluminescence and facilitates their use as optical probes in preclinical imaging // Nanoscale, 2015. Vol. 7. P. 10410–10420. doi: 10.1039/c5nr01403d.

[8] *Zellweger M.* Fluorescence spectroscopy of exogenous, exogenously-induced and endogenous fluorofores for the photodetection and photodynamic therapy of cancer. — Lausanne: Fevrier, 2000. 224 p.

[9] *Dolenko T., Burikov S., Vervald A., Vlasov I., Dolenko S., Laptinskiy K., Rosenholm J. M., Shenderova O.* Optical imaging of fluorescent carbon biomarkers using artificial neural networks // J. Biomed. Opt., 2014. Vol. 19. No. 11. P. 117007. 9 p. doi: 10.1117/1.JBO.19.11.117007.

[10] *Хайкин. С.* Нейронные сети: полный курс / Пер. с англ. — 2-е изд. — М.: Изд. дом «Вильямс», 2006. 1104 с. (*Haykin S. S.* Neural networks: A comprehensive foundation. — Macmillan, 1994. 696 p.)

[11] *Dolenko S.A., Gerdova I.V., Dolenko T.A., Fadeev V.V.* Laser fluorimetry of mixtures of polyatomic organic compounds using artificial neural networks // Quantum Electron., 2001. Vol. 31. No. 9. P. 834–838. doi: 10.1070/QE2001v031n09ABEH002056.

[12] *Li M., Verma B., Fan X., Tickle K.* RBF neural networks for solving the inverse problem of backscattering spectra // Neural Comput. Appl., 2008. Vol. 17. No. 4. P. 391–397. doi: 10.1007/s00521-007-0138-2.

[13] *Lenhardt L., Zeković I., Dramićanin T., Dramićanin M. D.* Artificial neural networks for processing fluorescence spectroscopy data in skin cancer diagnostics // Phys. Scripta, 2013. Vol. 2013. No. T157. P. 014057. doi: 10.1088/0031-8949/2013/T157/014057.

[14] *Мандрикова О. В., Жижикина Е. А.* Оценка состояния геомагнитного поля на основе совмещения вейвлет-преобразования с радиальными нейронными сетями // Машинное обучение и анализ данных, 2014. Т. 1. №10. С. 1335–1344.

[15] *Baharifar H., Amani A.* Cytotoxicity of chitosan/streptokinase nanoparticles as a function of size: an artificial neural networks study. Nanomed. Nanotechnol., 2016. Vol. 12. No. 1. P. 171–180. doi: 10.1016/j.nano.2015.09.002.

[16] *Laptinskiy K., Burikov S., Dolenko S., Efitorov A., Sarmanova O., Shenderova O., Vlasov I., Dolenko T.* Monitoring of nanodiamonds in human urine using artificial neural networks // Phys. Status Solidi A, 2016. Vol. 231. No. 10. P. 2614–2622. doi: 10.1002/pssa.201600178.

[17] *Kim Y.* Role of folate in colon cancer development and progression // J. Nutr., 2003. Vol. 133. P. 3731S–3739S.

[18] *Svetlov V. A., Dolenko S. A.* Development of the algorithm of adaptive construction of hierarchical neural network classifiers // Opt. Mem. Neural Networks, 2017. Vol. 26. No. 1. P. 40–46. doi: 10.3103/S1060992X17010076.

[19] *Specht D.* A General regression neural network // IEEE T. Neural Networ., 1991. Vol. 2. No. 6. P. 568–576. doi: 10.1109/72.97934.

[20] NeuroShell 2. `http://www.neuroproject.ru/aboutproduct.php?info=ns2info`.

[21] *Efitorov A., Burikov S., Dolenko T., Laptinskiy K., Dolenko S.* Significant feature selection in neural network solution of an inverse problem in spectroscopy // Procedia Comp. Sci., 2015. Vol. 66. P. 93–102. doi: 10.1016/j.procs.2015.11.012.

[22] *Gevrey M., Dimopoulos I., Lek S.* Review and comparison of methods to study the contribution of variables in artificial neural network models // Ecol. Model., 2003. Vol. 160. P. 249–264. doi: 10.1016/S0304-3800(02)00257-0.

Поступила в редакцию 01.09.2017

Estimation of the perspective of using machine learning methods for the purpose of monitoring of the excretion of theranostic fluorescent nanocomposites out of the organism*

***O. E. Sarmanova*[1], *S. A. Burikov*[1,2], *S. A. Dolenko*[2], *I. V. Isaev*[2], *V. A. Svetlov*[2], *K. A. Laptinskiy*[1,2], *T. A. Dolenko*[1,2]**

`helga-sharman@rambler.ru`

[1]Faculty of Physics, Lomonosov Moscow State University, 1/2 Leninskie Gory, Moscow, Russia

[2]Skobeltsyn Institute of Nuclear Physics, Lomonosov Moscow State University, 1/2 Leninskie Gory, Moscow, Russia

*This study was supported by the following foundations: grant of the Russian Foundation for Basic Research No. 15-07-08975-a (HNC algorithm) and the grant of Russian Science Foundation No. 17-12-01481 (all other studies). The authors are grateful to O. A. Shenderova (Adamas Nanotechnologies, Inc., USA), Jessica Rosenholm, Eva von Haartman and Didem Sen Karaman (University of Abo Academy of Finland, Turku) for the synthesis of carbon dots and nanocomposites based on them.

Background: At present, development of new nanomaterials that can be used for diagnostics and medical treatment simultaneously is utterly relevant in biomedicine. While using such agents, one has to control their excretion out of the body.
Methods: The results of the estimation of the perspective for application of machine learning methods for monitoring of the excreted theranostic nanocomposites (carbon dots, covered by copolymer and folic acid) and their components by their fluorescence spectra in urine are presented. The problem was solved as a clusterization problem (by k-means and by the algorithm of adaptive construction of hierarchical neural classifiers, developed by the authors), and as a classification problem (by neural networks). None of the clusterings revealed sensitivity to the types of nanoparticles contained in the suspension.
Results: The best results of the solution of the classification problem were provided by a perceptron with 8 neurons in the single hidden layer, trained on the set of significant input features selected by cross-correlation. Recognition accuracy averaged over all five classes was 72.3%.
Keywords: *pattern recognition; clusterization; artificial neural networks; classification; carbon nanocomposites; fluorescent spectroscopy*

DOI: 10.21469/22233792.3.4.01

References

[1] Kim, J., Y. Piao, and T. Hyeon. 2009. Multifunctional nanostructured materials for multimodal imaging, and simultaneous imaging and therapy. *Chem. Soc. Rev.* 38:372–390. doi: 10.1039/B709883A.

[2] Doane, T. L., and C. Burda. 2012. The unique role of nanoparticles in nanomedicine: Imaging, drug delivery and therapy. *Chem. Soc. Rev.* 41(7):2885–2911. doi: 10.3390/ijms18051102.

[3] Cheng, L., C. Wang, L. Feng, K. Yang, and Z. Liu. 2014. Functional nanomaterials for phototherapies of cancer. *Chem. Rev.* 114:10869–10939. doi: 10.1021/cr400532z.

[4] Elgqvist, J. 2017. Nanoparticles as theranostic vehicles in experimental and clinical applications — focus on prostate and breast cancer (review). *Int. J. Mol. Sci.* 18:1102. 53 p. doi: 10.3390/ijms18051102.

[5] Bartelmess, J., S. J. Quinn, and S. Giordani. 2015. Carbon nanomaterials: Multi-functional agents for biomedical fluorescence and Raman imaging. *Chem. Soc. Rev.* 11:4672–4698. doi: 10.1039/C4CS00306C.

[6] Hong, G., S. Diao, A. L. Antaris, and H. Dai. 2015. Carbon nanomaterials for biological imaging and nanomedicinal therapy. *Chem. Rev.* 115(19):10816–10906. doi: 10.1021/acs.chemrev.5b00008.

[7] Prabhakar, N., T. Nareoja, E. von Haartman, D. Karaman, S. Burikov, T. Dolenko, T. Deguchi, V. Mamaeva, P. Hanninen, I. Vlasov, O. Shenderova, and J. M. Rosenholm. 2015. Functionalization of graphene oxide nanostructures improves photoluminescence and facilitates their use as optical probes in preclinical imaging. *Nanoscale* 7:10410–10420. doi: 10.1039/c5nr01403d.

[8] Zellweger, M. 2000. *Fluorescence spectroscopy of exogenous, exogenously-induced and endogenous fluorofores for the photodetection and photodynamic therapy of cancer*. Lausanne: Fevrier, 2000. 224 p.

[9] Dolenko, T., S. Burikov, A. Vervald, I. Vlasov, S. Dolenko, K. Laptinskiy, J. M. Rosenholm, and O. Shenderova. 2014. Optical imaging of fluorescent carbon biomarkers using artificial neural networks. *J. Biomed. Opt.* 19(11):117007. 9 p. doi: 10.1117/1.JBO.19.11.117007.

[10] Haykin, S. S. 1994. *Neural networks: A comprehensive foundation*. Macmillan. 696 p.

[11] Dolenko, S. A., I. V. Gerdova, T. A. Dolenko, and V. V. Fadeev. 2001. Laser fluorimetry of mixtures of polyatomic organic compounds using artificial neural networks. *Quantum Electron.* 31(9):834–838. doi: 10.1070/QE2001v031n09ABEH002056.

[12] Li, M., B. Verma, X. Fan, and K. Tickle. 2008. RBF neural networks for solving the inverse problem of backscattering spectra. *Neural Comput. Appl.* 17(4):391–397. doi: 10.1007/s00521-007-0138-2.

[13] Lenhardt, L., I. Zeković, T. Dramićanin, and M. D. Dramićanin. 2013. Artificial neural networks for processing fluorescence spectroscopy data in skin cancer diagnostics. *Phys. Scripta* 2013(T157):014057. doi: 10.1088/0031-8949/2013/T157/014057.

[14] Mandrikova, O. V., and E. A. Zhizhikina. 2014. Otsenka sostoyaniya geomagnitnogo polya na osnove sovmeshcheniya veyvlet-preobrazovaniya s radial'nymi neyronnymi setyami [Estimation of the state of the geomagnetic field on the basis of combining the wavelet transform with radial neural networks]. *Machine Learning Data Anal.* 1(10):1335–1344.

[15] Baharifar, H., and A. Amani. 2016. Cytotoxicity of chitosan/streptokinase nanoparticles as a function of size: an artificial neural networks study. *Nanomed. Nanotechnol.* 12(1):171–180. doi: 10.1016/j.nano.2015.09.002.

[16] Laptinskiy, K., S. Burikov, S. Dolenko, A. Efitorov, O. Sarmanova, O. Shenderova, I. Vlasov, and T. Dolenko. 2016. Monitoring of nanodiamonds in human urine using artificial neural networks. *Phys. Status Solidi A* 231(10):2614–2622. doi: 10.1002/pssa.201600178.

[17] Kim, Y. 2003. Role of folate in colon cancer development and progression. *J. Nutr.* 133:3731S–3739S.

[18] Svetlov, V. A., and S. A. Dolenko. 2017. Development of the algorithm of adaptive construction of hierarchical neural network classifiers. *Opt. Mem. Neural Networks* 26(1):40–46. doi: 10.3103/S1060992X17010076.

[19] Specht, D. 1991. A general regression neural network. *IEEE T. Neural Networ.* 2(6):568–576. doi: 10.1109/72.97934.

[20] NeuroShell 2. Available at: `http://www.wardsystems.com/neuroshell2.asp` (accessed December 29, 2017).

[21] Efitorov, A., S. Burikov, T. Dolenko, K. Laptinskiy, and S. Dolenko. 2015. Significant feature selection in neural network solution of an inverse problem in spectroscopy. *Procedia Comp. Sci.* 66:93–102. doi: 10.1016/j.procs.2015.11.012.

[22] Gevrey, M., I. Dimopoulos, and S. Lek. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.* 160:249–264. doi: 10.1016/S0304-3800(02)00257-0.

Received September 01, 2017