

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В БИОМЕДИЦИНСКИХ СИСТЕМАХ INFORMATION TECHNOLOGIES IN BIOMEDICAL SYSTEMS

УДК 004.89

DOI: 10.17587/it.26.46-55

Я. Е. Львович, д-р техн. наук, проф., e-mail: office@vivt.ru,
Воронежский институт высоких технологий, Воронеж,
И. Л. Каширина, д-р техн. наук, проф., e-mail: kash.irina@mail.ru,
М. В. Демченко, аспирант, e-mail: masha-vrn@yandex.ru,
Воронежский государственный университет, г. Воронеж

Использование методов машинного обучения для исследования маркеров атеросклероза магистральных артерий

На основе методов интеллектуального анализа данных и машинного обучения анализируются маркеры атеросклероза магистральных артерий. Основная цель данного исследования — поиск факторов и их ассоциаций, определяющих высокую вероятность наличия атеросклероза магистральных артерий на основании данных многоканальной объемной сфигмографии, и разработка прикладного программного обеспечения для ранней диагностики этого заболевания.

Ключевые слова: машинное обучение, нейронные сети, деревья решений, карты Кохонена, атеросклероз артерий конечностей, бинарная классификация, чувствительность, специфичность

Введение

Заболевания сердечно-сосудистой системы являются наиболее распространенной причиной смертности во всем мире, вызывающей более 30 % от общего числа всех фатальных исходов. Одной из ведущих причин ишемических поражений органов является такое заболевание, как атеросклероз, т. е. патология, приводящая к деформации и сужению просвета сосудов вплоть до обтурации (закупорки сосуда). Во всем мире непрерывно ведутся исследования, направленные на предупреждение и лечение этого заболевания [1].

Согласно новым исследованиям простым и в то же время эффективным критерием атеросклеротического поражения сосудов является нарушение баланса систолического артериального давления (САД) на верхних и нижних конечностях [2]. В то время как позиции лодыжечно-плечевого индекса (ЛПИ), характеризующего отношение САД на руках и ногах и являющегося маркером атеросклеротического поражения конечностей, в современных руководствах достаточно хорошо определены, значение асимметрии АД на руках (Δ САД_р) или

ногах (Δ САД_н) для диагностики и прогноза атеросклероза еще изучается [2–4]. В качестве коэффициента значимой асимметрии артериального давления на верхних или нижних конечностях многими медицинскими экспертами принимается значение 15 мм рт. ст. Однако данная величина не является общепринятым показателем и часто во многих исследованиях при ее определении встречаются расхождения. В частности, авторы исследования [2] рассматривают значимую разницу давлений 10 мм рт. ст. на верхних и 20 мм рт. ст., на нижних конечностях. Опубликованные в 2014 г. данные Фрамингемского эпидемиологического проекта показали, что разница САД на руках, превышающая 10 мм рт. ст., встречалась в 26,2 % случаев от общего числа наблюдений за жителями старше 40 лет и была связана с повышенным риском развития различных сердечно-сосудистых заболеваний [3]. В одном из современных мета-анализов было продемонстрировано, что разница САД на руках, превышающая 15 мм рт. ст., является не только предиктором атеросклеротического поражения периферических артерий, но и значимо ассоциируется с увеличением риска сердечно-сосудистой смертности [4].

В ходе данного исследования с помощью методов машинного обучения предпринята попытка оценить эффективность использования многоканальной объемной сфигмографии (МОС) при проведении скрининга взрослого населения. Выбор данной методики объясняется необходимостью дополнить набор достаточно простых диагностических инструментов (анамнез, антропометрия, шкала SCORE и т. д.) тестом, способным обнаруживать признаки бессимптомного атеросклеротического поражения артерий без существенных дополнительных затрат [5–7].

Исследование включает следующие этапы.

1. Анализ взаимосвязи маркеров атеросклероза (полученных на основе результатов МОС) и их предикторов с использованием методов машинного обучения. Поиск зависимостей маркеров атеросклероза от гемодинамических параметров пациента и построение моделей зависимостей маркеров от социально-демографических, антропометрических и клинических факторов.

2. Отыскание значений асимметрии САД на руках и ногах, являющихся наиболее информативными признаками атеросклероза магистральных артерий.

3. Разработка комплекса программ для диагностики атеросклероза с помощью нейросетевых моделей.

4. Разработка веб-сайта и мобильного приложения для пациентов и врачей, позволяющих пройти ряд диагностических тестов по обнаружению атеросклероза артерий конечностей.

Исходные данные

В данном исследовании для выявления нарушений артериального кровообращения используется комбинация нескольких признаков (маркеров), основанных на значениях САД и представленных следующими формулами:

$$ArmsIndex = \begin{cases} 1, & \text{если } |\Delta САД_p| \geq \Delta_1; \\ 0, & \text{иначе;} \end{cases} \quad (1)$$

$$LegsIndex = \begin{cases} 1, & \text{если } |\Delta САД_n| \geq \Delta_2; \\ 0, & \text{иначе;} \end{cases} \quad (2)$$

$$ЛПИ = \begin{cases} 1, & \text{если } ЛПИ_l \leq \Delta_3 \text{ или } ЛПИ_p \leq \Delta_3; \\ 0, & \text{иначе.} \end{cases} \quad (3)$$

В формулах (1)–(3) присутствуют три параметра Δ_1 , Δ_2 , Δ_3 , характеризующих диагно-

стически значимое расхождение САД на руках или ногах, а также значение ЛПИ на левых или правых конечностях. Значение этих параметров предлагается определять с помощью алгоритмов машинного обучения, на основе которых будет строиться диагностическая модель, а затем будут отыскиваться параметры, обеспечивающие максимальную достоверность построенной модели [8].

В качестве обучающей выборки для задачи машинного обучения на начальном этапе использовалась неперсонифицированная база данных обследования пациентов Богучарского района Воронежской области, проведенного в рамках программы всеобщей диспансеризации. В ходе программы было обследовано 522 жителя, у которых помимо обязательных процедур методом МОС выполнялось синхронное измерение систолического и диастолического артериального давления на верхних и нижних конечностях, вычислялась их разница, а также автоматически рассчитывалось значение ЛПИ и показателей асимметрии на конечностях. Кроме того, фиксировались антропометрические, клинические, гемодинамические и другие показатели пациентов, что в совокупности составило 45 параметров. Перечень из 28 наиболее значимых показателей представлен в табл. 1. Фрагмент исходной вы-

Таблица 1

Входные признаки

Категория признаков	Признаки
Гемодинамические	Систолическое/диастолическое/пульсовое артериальное давление на правой/левой руке/ноге (САД _{пр} , ДАД _{пр} , ПД _{пр} , САД _{лр} , ДАД _{лр} , ПД _{лр} , САД _{пн} , САД _{лн}), частота сердечных сокращений (ЧСС), скорость пульсовой волны (cfPWV, baPWV)
Социально-демографические	Пол, возраст, курительщик
Антропометрические	Рост, вес, индекс массы тела (ИМТ)
Лабораторные	Глюкоза, общий холестерин (ОХ)
Клинические	Артериальная гипертензия (АГ), стенокардия, инфаркт миокарда (ИМ), острое нарушение мозгового кровообращения (ОНМК), аортокоронарное шунтирование/чрескожное вмешательство (АКШ/ЧКВ), сахарный диабет (СД), хроническая сердечная недостаточность (ХСН), фибрилляция/трепетание предсердий (ФП/ТП), ожирение

Фрагмент исходного набора данных

Код	Возраст	Пол	Курит	АГ	ИМ	ОНМК	СД	ХСН	ФП/ТП	Рост	Вес	Глюкоза	ОХ
1	51	М	Нет	Нет	Да	Нет	Нет	Нет	Нет	174	84	5,14	6,80
2	51	Ж	Нет	Нет	Нет	Нет	Нет	Нет	Нет	152	53	4,26	7,68
3	64	М	Нет	Да	Нет	Да	Да	Да	Нет	165	80	8,50	7,00
4	52	Ж	Нет	Да	Нет	Нет	Нет	Нет	Нет	158	62	4,20	4,50
5	66	Ж	Нет	Да	Нет	Нет	Нет	Нет	Нет	160	90	5,01	4,61
6	50	М	Да	Да	Нет	Нет	Нет	Нет	Нет	172	69	5,70	4,30
7	58	М	Нет	Нет	Нет	Нет	Да	Нет	Да	175	138	11,35	6,40

борки представлен в табл. 2. Детали данного обследования отражены в источниках [9, 10]. Исследование проводилось по инициативе областного кардиологического диспансера при ВОКБ № 1 г. Воронежа.

Исследуемый набор данных представляет собой выборку достаточно малого объема, так как содержит 522 примера, что соответствует числу обследованных пациентов. В условиях малого объема дополнительное разбиение выборки на обучающее и тестовое подмножества способно отрицательно сказаться на ее общей репрезентативности и, следовательно, качестве обучения.

Для решения этой проблемы в данном исследовании была применена процедура n -кратной кросс-проверки в сочетании с механизмом бутстрепа при формировании подвыборок [11]. Такая процедура позволяет, с одной стороны, задействовать весь имеющийся набор данных в обучении, а с другой стороны, максимально точно оценить производительность модели на произвольных наборах данных.

Помимо малого объема другим критичным недостатком исходной выборки, типичным для задач медицинской диагностики и способным отрицательно повлиять на чувствительность диагностических моделей, является ее несбалансированность, выражающаяся в том, что число здоровых пациентов в выборке значительно превосходит число пациентов с атеросклерозом. Классическая стратегия минимизации общего числа ошибок классификации в данном случае способна отрицательным образом сказаться на качестве распознавания класса, представленного меньшинством объектов.

Для решения этой проблемы была задействована матрица стоимости ошибок классификации C , отражающая издержки, связанные со всеми возможными исходами, где C_{ij} — затраты на ошибку отнесения элемента класса

i к классу j . Большие веса назначаются тем классам, ошибка по которым является более критичной, в данной задаче — классу пациентов с атеросклерозом. Таким образом, в результате анализа исследуемого набора данных были выявлены проблемы, требующие применения особых методик разработки и обучения классифицирующих моделей.

Метрики качества

Значения метрик рассчитывались на основании матрицы ошибок классификации (табл. 3), строкам которой соответствуют ожидаемые результаты классификации, а столбцам — результаты, предсказанные моделью. Главная и побочная диагонали матрицы содержат число верных (True Positives, True Negatives) и ошибочных ответов (False Positives, False Negatives) соответственно.

На основании истинных и ложных ответов модели можно вычислить множество метрик качества классификации, в частности, Accuracy, Precision, Recall и True Negative Rate, представленных следующими формулами (при этом Recall и True Negative Rate в случае задач медицинской диагностики приобретают смысл чувствительности и специфичности, т.е. характеризуют способность модели верно классифи-

Таблица 3

Матрица ошибок классификации

Фактический класс	Предсказанный класс	
	Класс "+" (больные)	Класс "-" (здоровые)
Класс "+" (больные)	TP	FN
Класс "-" (здоровые)	FP	TN

цировать больных и здоровых пациентов соответственно):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}; \quad (4)$$

$$Precision = \frac{TP}{TP + FP}; \quad (5)$$

$$Recall = \frac{TP}{TP + FN}; \quad (6)$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}. \quad (7)$$

В медицинской диагностике оптимален метод исследования, который был бы как высоко специфичен, так и высоко чувствителен. Однако в реальности это труднодостижимо, так как повышение чувствительности алгоритма, как правило, сопровождается снижением его специфичности и наоборот. В основном выбор между чувствительностью и специфичностью зависит от условий конкретной задачи и от специфики заболевания. Например, в случае дорогого и травматичного лечения критичным становится риск "ложных тревог", однако в то же время пропуск опасных заболеваний на ранних стадиях способствует прогрессированию и развитию потенциально угрожающих здоровью состояний.

Чтобы создать оптимальную диагностическую систему, необходимо найти компромисс между показателями чувствительности и специфичности. Для решения этой проблемы хорошо подходит такой инструмент, как характеристические ROC-кривые. Графическое представление построенных в ходе данного исследования ROC-кривых, отражающих качество некоторых из разработанных классификаторов, представлено на рис. 1.

Графики кривых на рис. 1 отражают взаимосвязь между истинноположительными (TPR, чувствительность) и ложноположительными

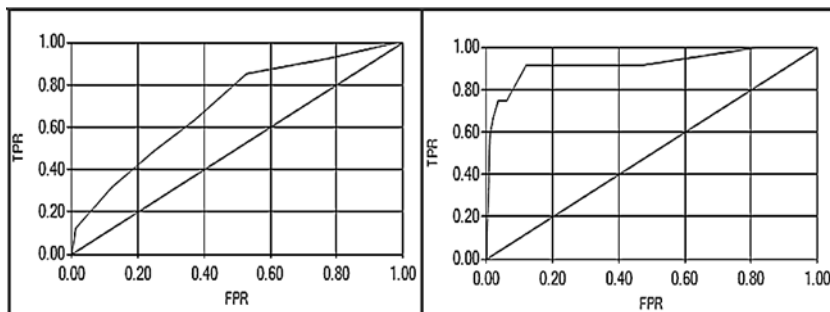


Рис. 1. Характеристические ROC-кривые

Таблица 4

Шкала классификации AUC

Интервал AUC	Качество модели
0,9...1	Отличное
0,8...0,9	Очень хорошее
0,7...0,8	Хорошее
0,6...0,7	Среднее
0,5...0,6	Неудовлетворительное

(FPR, 1-специфичность) результатами классификации, а площадь AUC под кривой является интегральным критерием, позволяющим отнести построенную модель к определенному классу качества, что демонстрирует табл. 4.

Площадь под кривой AUC часто рассматривается в задачах медицинской диагностики как оптимальный и достоверный критерий качества.

Метод деревьев решений

С учетом особенностей исходного набора данных в первую очередь в качестве используемого метода машинного обучения были рассмотрены деревья решений, широко применяемые в медицинских исследованиях.

Построение моделей выполнялось с использованием библиотеки машинного обучения Scikit-Learn для языка Python, в которой реализована как настройка весов матрицы стоимости ошибок классификации, так и процедура кросс-проверки.

Среди неоспоримых достоинств метода деревьев решений следует отметить его высокую интерпретируемость, поскольку в качестве результата мы получаем на выходе интуитивно понятную графическую иллюстрацию классификационной модели, которая может быть преобразована в набор правил. Например, одно

из правил, построенных для предсказания маркера ЛПИ по социально-демографическим, антропометрическим и клиническим факторам, имеет вид: "Если вес > 114,5 и (возраст > 66,5 или сахарный диабет (СД)), то ЛПИ < 0,9".

Метод деревьев решений позволяет не только разрабатывать классификационные диагностические алгоритмы, но и оценивать значимость используемых признаков.

Важность признака в методе деревьев решений вычисляется как нормализованный итог сокращения критерия ветвления, вызванный этим признаком. Критерий ветвления измеряет меру неопределенности в узлах дерева. В качестве такого критерия обычно используется индекс Джини или информационная энтропия. С помощью данного метода был проведен отбор наиболее значимых признаков, выбранных далее в качестве предикторов, т. е. независимых переменных анализа, на основании которых выполняется предсказание значений маркеров в целях выявления зависимостей между маркерами атеросклероза и социально-демографическими, антропометрическими, клиническими и гемодинамическими параметрами пациента.

На рис. 2 приведены значимые социально-демографические, антропометрические и клинические параметры для маркера ЛПИ (значимость масштабирована в интервал [0;100]).

Благодаря использованному механизму корректировки ошибок для классов, представленных меньшинством объектов, с помощью метода деревьев решений были построены модели, обладающие высокой точностью на обучающей выборке. Однако результаты кросс-проверки выявили сильное ухудшение чувствительности классификационных деревьев в процессе тестирования. В табл. 5 приведены значения чувствительности, специфичности и доли правильных ответов деревьев решений для всех маркеров и набора гемодинамических параметров.

В целях устранения недостатка низкой чувствительности классификационных деревьев было принято решение о необходимости по-

строения модели, обладающей лучшими показателями критериев качества, а именно нейронной сети архитектуры многослойный перцептрон, обученной по алгоритму обратного распространения ошибки [13]. При этом метод деревьев решений позволил существенно снизить размерность признакового пространства для построения нейросетевых моделей.

Нейронные сети архитектуры MLP

Нейронная сеть архитектуры многослойный перцептрон (MLP) состоит из нескольких слоев нейронов, причем каждый нейрон предыдущего слоя связан с каждым нейроном последующего слоя. Во многих практических приложениях оказывается достаточным рассмотрение двухслойной нейронной сети, имеющей скрытый и выходной слою. Именно такая конфигурация сети была использована для моделирования зависимостей маркеров атеросклероза от признаков, отобранных на этапе построения деревьев решений.

Классический алгоритм обучения сетей архитектуры MLP ставит задачу минимизации целевой функции ошибок сети. В задачах классификации, как правило, в качестве функции потерь используется кросс-энтропия (перекрестная энтропия), которая характеризует отклонение правильного ответа $y^*(x)$ от полученного сетью ответа y .

К сожалению, стандартный алгоритм обратного распространения ошибки не позволяет достигнуть желаемой чувствительности сети на несбалансированных выборках. Для решения данной проблемы во многих исследованиях предлагаются различные модификации классического алгоритма [12], одна из которых была использована в данной работе.

Модификация алгоритма заключалась во внедрении дополнительных конфигурационных параметров — коэффициентов матрицы затрат, настраиваемых пользователем наряду с такими основными параметрами, как скорость обучения, момент, число эпох. На основании матрицы затрат формировались коэффициенты целевой функции, являющиеся корректирующими факторами. Таким образом, стандартная функция ошибок приобрела смысл функции затрат, гибко учитывающей даже ошибки малочисленных классов. Основные модификации отражены в следующих формулах:

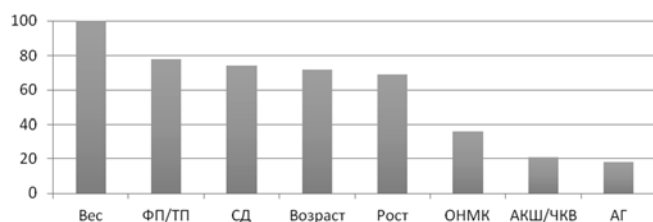


Рис. 2. Значимость предикторов для ЛПИ

Таблица 5

Качество деревьев решений по результатам кросс-проверки

Маркер	Чувствительность, % Recall	Специфичность, % True Negative Rate	Доля верных ответов, % Accuracy
ArmsIndex	39,7	94,7	90
LegsIndex	25,2	92,8	87,9
ЛПИ	35	97,6	92,5

$$CostFunction = -\sum_{j=1}^m y_{kj}^* \log y_{kj} K[class(k), j]; \quad (8)$$

$$K[i, j] = \begin{cases} CostVector[i], & i = j; \\ Cost[i, j], & i \neq j; \end{cases} \quad (9)$$

$$Cost[i, j] = \begin{cases} c_{ij}, & i \neq j; \\ 0, & i = j; \end{cases} \quad (10)$$

$$CostVector[i] = \begin{cases} 0, & \text{если } P(i) = 1; \\ \frac{1}{1 - P(i)} \sum_{j \neq i} P(j) Cost[i, j], & \text{иначе,} \end{cases} \quad (11)$$

где $CostFunction$ — функция затрат; m — число выходов сети; y_{kj}^* — правильное значение j -го выхода сети на k -м входном векторе; y_{kj} — полученное значение j -го выхода сети на k -м входном векторе; $K[i, j]$ — корректирующий фактор; $Cost[i, j]$ — матрица затрат на ошибку определения класса i к классу j ; $CostVector[i]$ — вектор затрат ошибочного определения объекта класса i к любому классу, отличному от i ; $P(i)$ — априорная вероятность принадлежности объекта классу i ; $class(k)$ — ожидаемый класс k -го примера выборки.

В качестве инструмента программной реализации была выбрана библиотека, реализованная на платформе .Net — Aforge.Neuro, содержащая набор базовых функций, необходимых для работы с различными нейросетевыми архитектурами. На базе платформы .Net было разработано приложение для диагностики атеросклероза с помощью нейронных сетей архи-

тектуры MLP, пример работы которого отражен на рис. 3.

Помимо основного обучающего модуля разработанный программный продукт предоставляет функционал определения основных критериев качества нейросетевых моделей, а также построения ROC-кривых.

На рис. 3 представлены результаты построения нейронной сети для маркера ЛПИ по набору гемодинамических параметров. Проиллюстрированная модель обладает высокой чувствительностью (75 %), специфичностью (94,12 %) и долей верных ответов (93,68 %), что в целом обеспечивает высокое качество по шкале AUC — 0,93.

Таким образом, с помощью применения классических подходов нейросетевого обучения и с учетом оптимизации базового алгоритма удалось достигнуть высокой эффективности построенных классификаторов. Однако очевидным недостатком построенных нейросетевых моделей является их низкая интерпретируемость.

Карты Кохонена

Методом, в определенном смысле объединяющим преимущества нейросетевых и интерпретируемых алгоритмов и предоставляющим возможность визуализации закономерностей между маркерами и предикторами, обнаруженных в исходном наборе данных, является метод самоорганизующихся карт Кохонена [11].

Существенным отличием карт Кохонена от рассмотренных ранее сетей MLP и деревьев решений является то, что данный метод представляет собой метод нейросетевой кластеризации (а не классификации). Данная сеть обучается без учителя на основе алгоритма самоорганизации (Self Organizing Map).

Для построения карт Кохонена для каждого из маркеров по набору гемодинамических, клинических и других показателей также был разработан программный продукт, выполненный на технологиях .NET [13]. Набор раскрасок карт Кохонена по входным признакам позволяет выявить ряд закономерностей в исходном наборе данных.

Например, при построении карты Кохонена для исследования связи

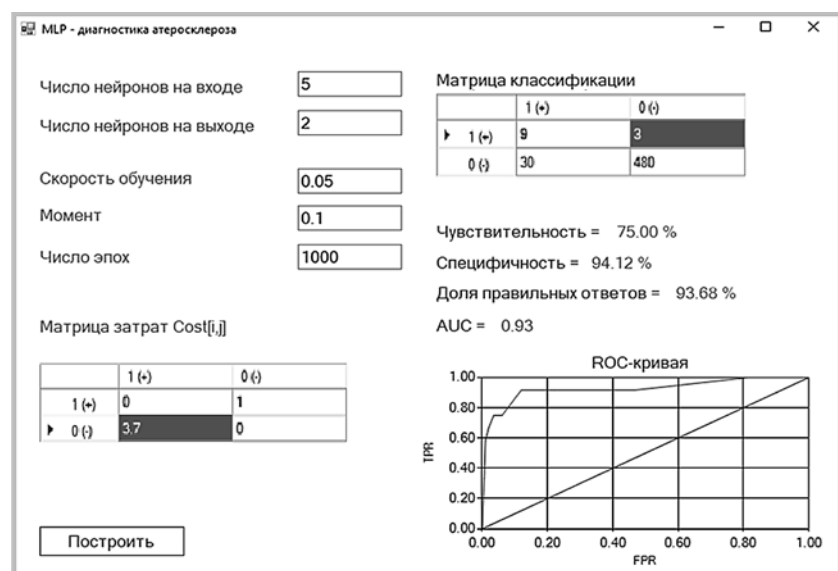


Рис. 3. Диагностика атеросклероза с помощью нейронных сетей MLP

Характеристики нейронных сетей для маркера ArmsIndex

№ сети	Δ_1	Набор факторов	Чувствительность	Специфичность	Доля верных ответов	AUC
1	14	Антропометрические и социально-демографические	75,61	74,45	85,33	0,87
2	14	Клинические	73,41	72,14	81,46	0,84
3	13	Гемодинамические	70,73	65,93	78,09	0,74

маркера ЛПИ с отобранными ранее социально-демографическими, антропометрическими и клиническими факторами сформировалось семь кластеров, в одном из которых наблюдалось значимое повышение числа пациентов с низким значением ЛПИ. В данном кластере наблюдались повышенные значения показателей веса (среднее значение — 92 кг), возраста (в среднем — 60 лет), наличие в анамнезе стенокардии (80 % пациентов), сахарного диабета (85 % пациентов), хронической сердечной недостаточности (74 % пациентов). При построении карт Кохонена по признакам, значимым для маркера LegsIndex, также было выделено семь кластеров, в одном из которых наблюдалось значимое повышение LegsIndex. В данном кластере преобладали данные лиц мужского пола (90 % пациентов), со статусом курильщика (65 % пациентов), с отягощенным анамнезом (67 % пациентов), артериальной гипертонией (87 % пациентов), фибрилляцией/трепетанием предсердий (56 % пациентов), хронической сердечной недостаточностью (56 % пациентов). При этом разброс значений внутри кластеров (коэффициент вариации) не превышал 20 %, что позволяет считать кластеры достаточно однородными. Таким образом, карты Кохонена позволяют не только предсказывать риск атеросклероза для новых пациентов, но и позволяют сформировать обобщенный портрет пациента с наличием того или иного маркера атеросклероза. Подробно результаты этого анализа описаны в работе [13], там же представлены результаты визуализации.

Определение коэффициентов асимметрии артериального давления

На начальном этапе анализа в качестве коэффициентов асимметрии артериального давления были выбраны значения $\Delta_1 = \Delta_2 = 15$ мм рт. ст., $\Delta_3 = 0,9$, предложенные экспертами области медицины и представленные

в формулах (1) и (2). Однако если значение Δ_3 является хорошо установленным в медицинской практике, то Δ_1 и Δ_2 не являются общепринятыми показателями и нередко во многих исследованиях при их определении встречаются расхождения. В рамках данного исследования была проведена калибровка данных величин.

Решение этой задачи подразумевало построение набора нейронных сетей различных конфигураций с зависимыми переменными-маркерами, сформированными с учетом различных значений асимметрии от 10 до 15. В табл. 6 отражен набор нейронных сетей наиболее высокого качества для маркера ArmsIndex, построенных по различным наборам предикторов (гемодинамических, клинических, антропометрических и др.).

Все модели, характеристики которых приведены в табл. 5, отличаются по значениям основных метрик качества — чувствительности, специфичности, доли правильных ответов и AUC, однако по причине того, что в качестве наиболее приоритетного критерия медицинскими экспертами был выбран интегральный критерий AUC, наиболее достоверное значение коэффициента асимметрии соответствовало сети с максимальным значением этого критерия.

Для маркера ArmsIndex наиболее информативным коэффициентом асимметрии было выбрано значение 14 (у сети с максимальным значением AUC — 0,87), для LegsIndex — 15. В табл. 7 приведены характеристики наиболее качественных моделей, построенных для каждого из рассматриваемых маркеров.

Максимального качества по кросс-проверке (AUC = 0,93) удалось достигнуть нейронной сети для маркера ЛПИ, построенной по набору гемодинамических параметров. В набор гемодинамических признаков входят признаки давления, ЧСС и скорость каротидно-фemorальной пульсовой волны. Поскольку повышение скорости пульсовой волны является общепринятым фактором развития атеросклероза, его

Таблица 7

Нейронные сети наилучшего качества по метрике AUC ROC

Маркер	Набор предикторов	Качество по кросс-проверке
ЛПИ	САДпр, САДлн, ЧСС, ДАДпр, cfPWV_calc > 10 м/с	0,93 (отличное)
ArmsIndex ($\Delta_1 = 14$)	Пол, Анамнез, АГ, СД, ХСН, Возраст, Рост, Вес	0,8 (очень хорошее)
LegsIndex ($\Delta_2 = 15$)	ЧСС, САДлн, САДпр, САДлр, cfPWV_calc > 10 м/с, ДАДлр, ДАДпр	0,73 (хорошее)
ЛПИ	Возраст, Рост, Вес, АГ, ОНМК, АКШ/ЧКВ, СД, ФП/ТП	0,77 (хорошее)

тесная связь с ЛПИ подтверждает эффективность ЛПИ в качестве маркера атеросклероза конечностей. Также необходимо отметить, что нейронные сети для ЛПИ обладали высоким качеством на наборе социально-демографических, антропометрических и клинических предикторов. Модель по этим параметрам проявила высокую чувствительность (более 81 %) и в целом является моделью хорошего качества по критерию AUC (0,77). Ценность такой модели заключается в том, что ее возможно применять в качестве инструмента быстрой диагностики, без проведения дополнительных лабораторных исследований, измерений артериального давления и инвазивного вмешательства, что позволяет использовать ее в системах первичной дистанционной диагностики. Также

модель хорошего качества удалось построить для ArmsIndex ($\Delta_1 = 14$) и группы предикторов, включающих пол, возраст, рост, вес, отягощенный анамнез, заболевания АГ, ХСН и СД.

Результаты исследования

По результатам данного исследования можно сделать следующие выводы:

1. Подтверждена эффективность ЛПИ как важного маркера атеросклероза магистральных артерий.

2. Разработана нейросетевая модель диагностики атеросклероза по антропометрическим, клиническим и социально-демографическим признакам пациента, позволяющая выявлять риск заболевания с точностью более 77 %, не требуя лабораторной диагностики и инвазивных вмешательств для использования в дистанционных диагностических системах.

3. С помощью карт Кохонена установлены общие закономерности риска атеросклероза, подтверждающие и дополняющие общепринятые признаки развития данного заболевания.

4. Проведен всесторонний анализ взаимосвязи маркеров атеросклероза магистральных артерий, полученных на основе результатов МОС и их предикторов. Задачи, исследованные в процессе анализа, и методы их решения представлены на рис. 4.

5. Вычислены наиболее информативные коэффициенты асимметрии артериального давления на руках. С учетом чувствительности и специфичности построенных моделей можно рекомендовать последовательность использования маркеров атеросклероза (рис. 5).

6. Наряду с представленными программными модулями для выявления маркеров атеросклероза с помощью методов машинного обучения было также разработано веб-приложение для пациентов и врачей, предназначенное для онлайн-диагностики атеросклероза, которое размещено по адресу: <http://cardio-002-site1.dtempurl.com/>. Данный ресурс, во-первых, содержит основную информацию о проведенном анализе и его результатах, а также ряд диагностических тестов для определе-



Рис. 4. Взаимосвязь использованных методов анализа

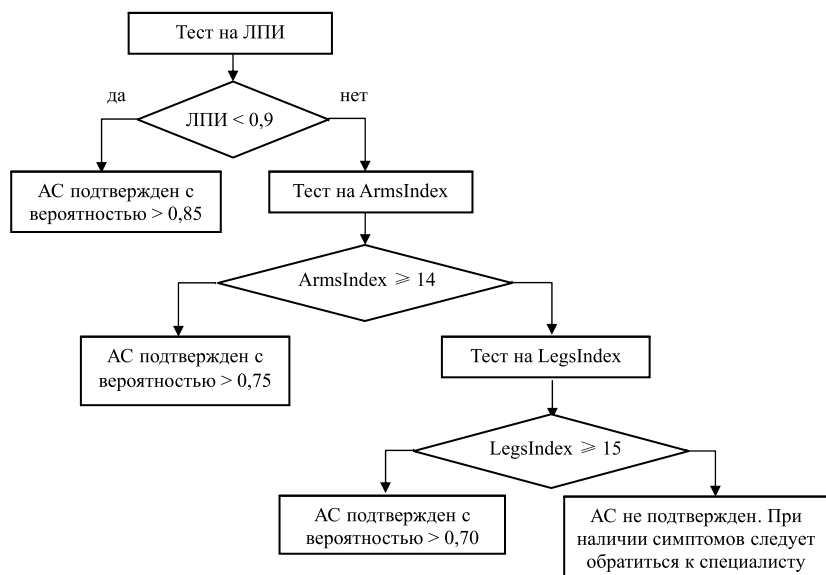


Рис. 5. Схема диагностики атеросклероза магистральных артерий с учетом полученной калибровки пороговых значений маркеров

ния риска атеросклероза, сформированных на основе выявленных закономерностей.

Заключение

Таким образом, в результате данного исследования был проведен всесторонний анализ маркеров атеросклероза и их предикторов, определены значения асимметрии систолического АД на руках и ногах, являющиеся наиболее информативными признаками атеросклероза магистральных артерий, а также разработан комплекс программ для диагностики атеросклероза с помощью нейросетевых моделей, в том числе веб-сайт для пациентов и врачей, позволяющий пройти ряд диагностических тестов в целях дистанционной оценки риска атеросклероза артерий конечностей.

Список литературы

1. Tuzcu E. M., Kapadia S. R., Tutar E. et al. High prevalence of coronary atherosclerosis in asymptomatic teenagers and young adults: evidence from intravascular ultrasound // *Circulation*. 2001;103:2705–2710.
2. Mitchell E. Noninvasive diagnosis of arterial disease. UpToDate. URL: <https://www.uptodate.com/contents/noninvasive-diagnosis-of-arterial-disease>.
3. Weinberg I. et al. The Systolic Blood Pressure Difference Between Arms and Cardiovascular Disease in the Framingham Heart Study // *The American Journal of Medicine*. 2014;127: 209–215.

4. Clark C. E., Taylor R. S., Shore A. C. et al. Association of a difference in systolic blood pressure between arms with vascular disease and mortality: a systematic review and meta-analysis // *Lancet*. 2012;380 (9838):218.

5. Sheng C. S., Liu M., Zeng W. F. et al. Four-Limb Blood Pressure as Predictors of Mortality in Elderly Chinese // *Hypertension*. 2013; 61: 1155–60.

6. Su H.-M., Lin T.-H., Hsu P.-C. et al. Association of Interarm Systolic Blood Pressure Difference with Atherosclerosis and Left Ventricular Hypertrophy // *PLoS ONE*. 2012; 7(8): e41173.

7. Su H.-M., Lin T.-H., Hsu P.-C. et al. Association of Bilateral Brachial-Ankle Pulse Wave Velocity Difference with Peripheral Vascular Disease and Left Ventricular Mass Index // *PLoS ONE*. 2014;9(2): e88331.

8. Демченко М. В., Каширина И. Л. Сравнительный анализ и оценка эффективности маркеров атеросклероза магистральных артерий // Сб. тр. Междунар. науч.-техн. конф. "Актуальные проблемы прикладной математики, информатики и механики" (Воронеж, 18–20 декабря 2017 г.). Воронеж, 2017. С. 636–644.

9. Хохлов Р. А., Остроушко Н. И., Гайдашев А. Э., Кирсанов Д. В., Ахмеджанов Н. М. Предикторы атеросклеротического поражения артерий конечностей по данным кардиоангиологического скрининга взрослого населения // *Рациональная Фармакотерапия в Кардиологии*. 2015. № 11(5). С. 470–476.

10. Хохлов Р. А., Гайдашев А. Э., Ахмеджанов Н. М. Использование многоканальной объемной сфигмографии для кардиоангиологического скрининга взрослого населения // *Рациональная фармакотерапия в кардиологии*. 2015. № 11(4). С. 470–476.

11. Каширина И. Л., Азарнова Т. В. Нейросетевые и гибридные системы: учебно-методическое пособие для вузов. Воронеж: Издательский дом ВГУ, 2014. 80 с.

12. Kukar M., Koponen I. Cost-Sensitive Learning with Neural Networks // *Machine Learning and Data Mining: proceedings of the 13th European Conference on Artificial Intelligence*. 1998. P. 445–449.

13. Демченко М. В., Каширина И. Л. Применение самоорганизующихся карт Кохонена в задаче диагностики атеросклероза // Матер. XVIII Междунар. науч.-метод. конф. "Информатика: проблемы, методология, технологии" (Воронеж, 8–9 февраля 2018 г.). Воронеж, 2018. С. 141–147.

14. 2013 ESH/ESC Guidelines for the management of arterial hypertension // *Journal of Hypertension*. 2013; 31:1281–357.

15. European Guidelines on cardiovascular disease prevention in clinical practice (version 2012) // *EurHeart J*. 2012; 33:1635–701.

16. Criqui M. H., McClelland R. L., McDermott M. M. et al. The Ankle-Brachial Index and Incident Cardiovascular Events in the MESA (Multi-Ethnic Study of Atherosclerosis) // *J Am Coll Cardiol*. 2010;56: 1506–12.

17. ESC Guidelines on the diagnosis and treatment of peripheral artery diseases: Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteries: the Task Force on the Diagnosis and Treatment of Peripheral Artery Diseases of the European Society of Cardiology (ESC) // *European Heart Journal*. 2011; 32, 2851–2906.

18. Jamei M., Nisnevich A., Wetchler E., Sudat S., Liu E., Upadhyaya K. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks // *PLoS ONE*. 2017. 12(7): e0181173. URL: <https://doi.org/10.1371/journal.pone.0181173>.

Y. E. Lvovich, DSc., Professor, e-mail: office@vivt.ru,
Voronezh Institute of High Technologies, Russian Federation, 394043, Voronezh,
I. L. Kashirina, DSc., Professor, e-mail: kash.irina@mail.ru,
M. V. Demchenko, Post-Graduate Student, e-mail: masha-vrn@yandex.ru,
Voronezh State University Russian Federation, 394036, Voronezh

The Use of Machine Learning Methods to Study Markers of Atherosclerosis of the Great Arteries

This study is devoted to the deep investigation of the main atherosclerosis markers and their predictors using the different machine learning methods. The main goal of the investigation is to find the most efficient markers and predictors, which would allow to identify the presence of this disease with the high accuracy. The dataset under consideration is based on the real patients characteristics, including blood pressure values, clinical, anthropometrical and social features. Such methods as classification trees, multilayered neural networks and Kohonen maps are used for solving the considered problem of binary classification. Accuracy, precision, recall (sensitivity) and true negative rate (specificity) are the main metrics, which are used to evaluate the quality of the resulting classification models. Classification trees method allowed not only to generate the most interpretable models, but also to order the main predictors by their importance, and the most important predictors were then passed as inputs to the multilayered neural networks. The base backpropagation algorithm, used by default for learning the network, has been modified to use the misclassification costs, which allowed to increase the quality of the classification for the unbalanced dataset. In order to show the common rules found in the dataset, the Kohonen maps were build for each of the main atherosclerosis markers. All the developed classification models can be considered as instruments of the non-invasive atherosclerosis diagnostics.

Keywords: machine learning, neural networks, multilayered perceptron, classification trees, Kohonen maps, atherosclerosis, binary classification, medical diagnostics, sensitivity, specificity

DOI: 10.17587/it.26.46-55

References

1. **Tuzcu E. M., Kapadia S. R., Tutar E.** et al. High prevalence of coronary atherosclerosis in asymptomatic teenagers and young adults: evidence from intravascular ultrasound, *Circulation*, 2001, vol. 103, pp. 2705–2710.
2. **Mitchell E.** Noninvasive diagnosis of arterial disease. UpToDate. URL: <https://www.uptodate.com/contents/noninvasive-diagnosis-of-arterial-disease>.
3. **Weinberg I.** et al. The Systolic Blood Pressure Difference Between Arms and Cardiovascular Disease in the Framingham Heart Study, *The American Journal of Medicine*, 2014, vol. 127, pp. 209–15.
4. **Clark C. E., Taylor R. S., Shore A. C.** et al. Association of a difference in systolic blood pressure between arms with vascular disease and mortality: a systematic review and meta-analysis, *Lancet*, 2012, vol. 379, pp. 905–14.
5. **Sheng C. S., Liu M., Zeng W. F.** et al. Four-Limb Blood Pressure as Predictors of Mortality in Elderly Chinese, *Hypertension*, 2013, vol. 61, pp. 1155–60.
6. **Su H.-M., Lin T.-H., Hsu P.-C.** et al. Association of Interarm Systolic Blood Pressure Difference with Atherosclerosis and Left Ventricular Hypertrophy, *PLoS ONE*, 2012, vol. 7, pp. e41173.
7. **Su H.-M., Lin T.-H., Hsu P.-C.** et al. Association of Bilateral Brachial-Ankle Pulse Wave Velocity Difference with Peripheral Vascular Disease and Left Ventricular Mass Index, *PLoS ONE*, 2014, vol. 9, pp. e88331.
8. **Demchenko M. V., Kashirina I. L.** The comparative analysis and estimation of the efficiency of the main atherosclerosis markers, *Aktual'nye Problemy Prikladnoj Matematiki, Informatiki i Mehaniki*, 2017, pp. 636–644 (in Russian).
9. **Hohlov R. A., Ostroushko N. I., Gajdashev A. Je., Kirsanov D. V., Ahmedzhanov N. M.** The predictors of the atherosclerosis of the lower extremities based on the data of the adults cardiological screening, *Racional'naja Farmakoterapija v Kardiologii*, 2015, no. 11(5), pp. 470–476 (in Russian).
10. **Hohlov R. A., Gajdashev A. Je., Ahmedzhanov N. M.** The using of the multichannel three-dimensional sphygmography in the adults cardiological screening, *Racional'naja Farmakoterapija v Kardiologii*, 2015, no. 11(4), pp. 470–476 (in Russian).
11. **Kashirina I. L., Azarnova T. V.** Nejrosetevye i gibridnye sistemy: uchebno-metodicheskoe posobie dlja vuzov (Neural networks and hybrid systems: the manual for the universities), Voronezh, Publishing house of VGU, 2014, 80 p. (in Russian).
12. **Kukar M., Kononenko I.** Cost-Sensitive Learning with Neural Networks, *Machine Learning and Data Mining: proceedings of the 13th European Conference on Artificial Intelligence*, 1998, pp. 445–449.
13. **Demchenko M. V., Kashirina I. L.** Using Kohonen maps in the diagnostics of the atherosclerosis, *Materialy XVIII Mezh-dunarodnoj nauchno-metodicheskoy konferencii*, 2018, pp. 141–147 (in Russian).
14. **2013 ESH/ESC Guidelines** for the management of arterial hypertension, *Journal of Hypertension*, 2013, vol. 31, pp. 1281–357.
15. **European Guidelines** on cardiovascular disease prevention in clinical practice (version 2012), *European Heart Journal*, 2012, vol. 33, pp. 1635–701.
16. **Criqui M. H., McClelland R. L., McDermott M. M.** et al. The Ankle-Brachial Index and Incident Cardiovascular Events in the MESA (Multi-Ethnic Study of Atherosclerosis), *J Am Coll Cardiol*, 2010, vol. 56, pp. 1506–12.
17. **ESC Guidelines** on the diagnosis and treatment of peripheral artery diseases: Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteries: the Task Force on the Diagnosis and Treatment of Peripheral Artery Diseases of the European Society of Cardiology (ESC), *European Heart Journal*, 2011, vol. 32, pp. 2851–2906.