УДК 004.852

Кухто Я.С., Голденок Е.Е., Лукьянова Н.А.

МОДЕЛИ БИНАРНОЙ КЛАССИФИКАЦИИ ДЛЯ ДИАГНОСТИКИ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ ПО ПОКАЗАТЕЛЯМ ЭКСПРЕССИИ ГЕНОВ

Красноярский государственный медицинский университет имени профессора В.Ф. Войно-Ясенецкого, г. Красноярск

Представлены результаты сравнения учебных моделей бинарной классификации, развивающих систему поддержки принятия врачебных решений для диагностики рака молочной железы по показателям экспрессии генов. Эксперименты проводились с базой данных ДНК-микрочипирования 590 образцов ткани молочной железы, состоящей из количественных показателей экспрессии 17814 генов по каждому образцу ткани. Для снижения размерности признакового пространства использовались методы случайного леса, рекурсивного удаления признаков и экспертной оценки. Для классификации были использованы методы: k-ближайших соседей (KNN), метод наивного Байеса (NB), метод опорных векторов (SVM), случайный лес (RF). Наибольшая точность классификации рака молочной железы получена при использовании модели случайный лес по показателям экспрессии пяти генов: COL10A1, ING1, ATP2B4, USP40, OXSR1.

Ключевые слова: машинное обучение, рак молочной железы, профили экспрессии генов, точность.

Генетические тесты, такие как ДНК-чипы позволяют проводить мониторинг генной экспрессии, определять функциональное значение генов, изучать патогенетические основы различных заболеваний, в том числе канцерогенез. Опухоль развивается в результате повреждения генетического материала, не приводящего к появлению летальных мутаций [2]. Генами-мишенями для патогенных факторов являются протоонкогены, гены-супрессоры опухолевого роста, гены, регулирующие клеточный апоптоз, гены репарации ДНК [1]. Рак молочной железы является гетерогенным заболеванием и имеет различный ответ на лечение в зависимости от стадии патологического процесса и подтипа опухоли [8]. Раннее лечение рака повышает вероятность излечения и снижает летальность и вероятность рецидива. ДНК чипы используются во многих диагностических тест-системах РМЖ [3,11]. Преимуществом ДНК чипов перед традиционными методами диагностики является возможность одновременного анализа множества образцов ткани с использованием минимальных количеств исследуемого материала и реагентов. Данные экспрессии генов ДНК-микрочипа отражают состояние клетки на молекулярном уровне через стандартизированные показатели интенсивности флюоресценции гибридизованных цепей кДНК и имеют большую перспективу в качестве инструмента для постановки диагноза РМЖ [7], в том числе машинного обучения [4,5,6,9,10,12].

Цель работы

Целью работы является проектирование и сравнение учебных моделей бинарной классификации рака молочной железы по профилю экспрессии генов, развивающих систему поддержки принятия врачебных решений.

Материал и методы

Материалом для исследования является база данных профиля экспрессии генов рака молочной железы из программы Атлас генома рака (TCGA). Данные получены с сайта Mendeley Data (https://data.mendeley.com/datasets/v3cc2p38hb/1), предназначенного для обмена исследовательскими данными с целью повторного их использования для улучшения воспроизводимости исследований. База данных ДНК-микрочипирования состоит из количественных показателей экспрессии 17814 генов по каждому из 590 образцов ткани молочной железы, в числе которых 61 образец нормальной ткани и 529 образцов рака молочной железы.

Методы исследования. Для проверки статистически значимых различий в группах здоровых и с патологией по показателям экспрессии генов использовали критерий Манна-Уитни, уровень статистической значимости был принят: $p < 0,05$. Надо отметить, что база данных имеет не сбалансированные классы, образцов ткани без патологии в 9,67 раз меньше, чем наблюдений с РМЖ, что является типичным для медицинских исследований. Для устранения дисбаланса классов использовался метод передискретизации меньшинства (SMOTE). Общее количество наблюдений после предобработки составило 1058, по 529 в каждом классе. Предобработка данных и проектирование моделей машинного обучения произведено в среде Python (версии 3.10.12). Дизайн исследования представлен на рисунке 1. Для снижения размерности признакового пространства использованы методы случайного леса, рекурсивного удаления признаков и экспертной оценки. Задача классификации решена следующими методами машинного обучения: k-ближайших соседей (kNN), метод наивного Байеса (NB), метод опорных векторов (SVM), случайный лес (RF). Обучающая выборка включала 741 наблюдение, тестовая выборка состояла из 317 наблюдений. Для оценки качества моделей и предотвращения переобучения была проведена 10-кратная перекрестная проверка.

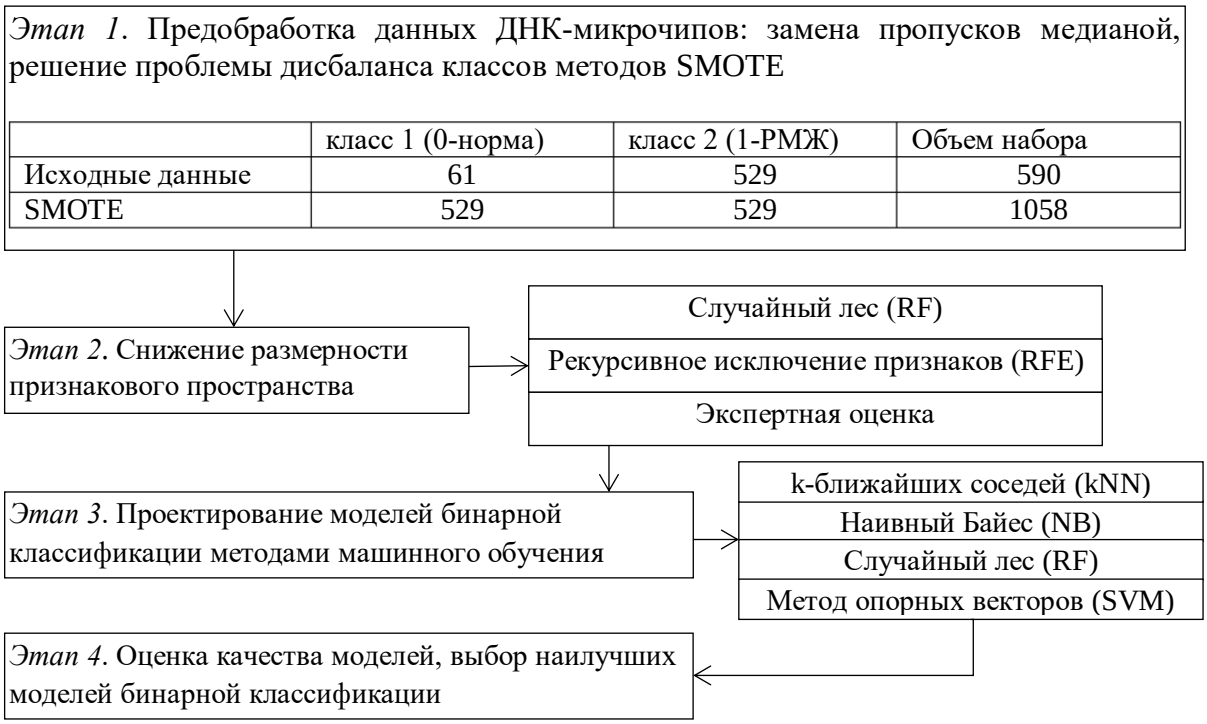

Рис. 1. Дизайн исследования.

Результаты и обсуждение

Результаты. Наилучшие результаты классификации показала модель, построенная по пяти генам (COL10A1, PARP1 WISP1, MMP11, IGSF10), отобранным методом рекурсивного удаления признаков. Увеличение уровня экспрессии в группе пациентов с РМЖ наблюдается для COL10A1, PARP1, WISP1, MMP11 ($p<0,05$, критерий Манна-Уитни). При этом отмечается статистически значимое снижение экспрессии для гена IGSF10 ($p<0,05$). На рисунке 2 представлена диаграмма размаха признаков COL10A1, PARP1 WISP1, MMP11, IGSF10, которая визуализирует явную отделимость значений показателей экспрессии каждого из пяти генов для больных РМЖ и здоровых.

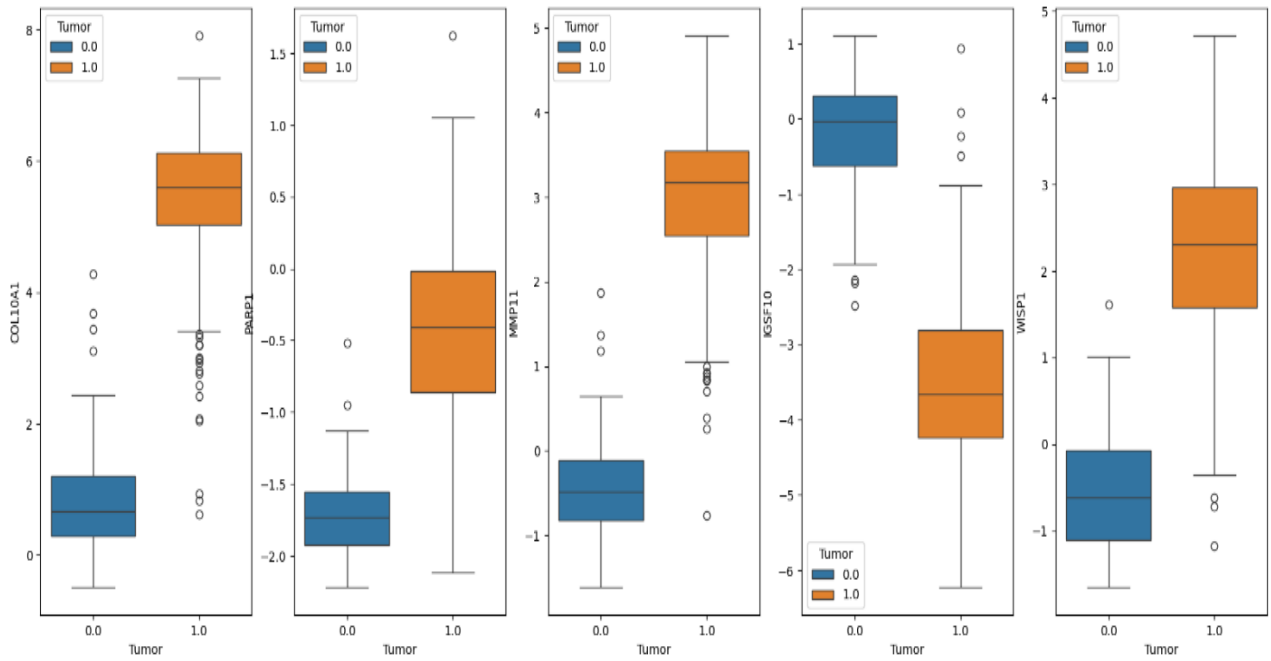

Рис. 2. Диаграмма размаха признаков COL10A1, PARP1, MMP11, IGSF10, WISP1 (0- РМЖ отсутствует, 1- наличие РМЖ)

На основе результатов исследований Genomic Data Commons (GDC) Национального института рака (NCI), принятых в качестве экспертной оценки, отобраны 5 генов из исследуемой базы данных наиболее часто подвергающихся мутациям при раке молочной железы: TP53, TTN, GATA3 ,MAP3K1, HMCN1. Значения экспрессии генов TTN, GATA3, MAP3K1, HMCN1 в группах больных и здоровых имеют статистически значимые различия ($p<0,05$). На рисунке 3 представлены диаграммы размаха признаков TTN, HMCN1, MAP3K1, GATA3, показывающие расхождения их значений в классах больных РМЖ и здоровых.

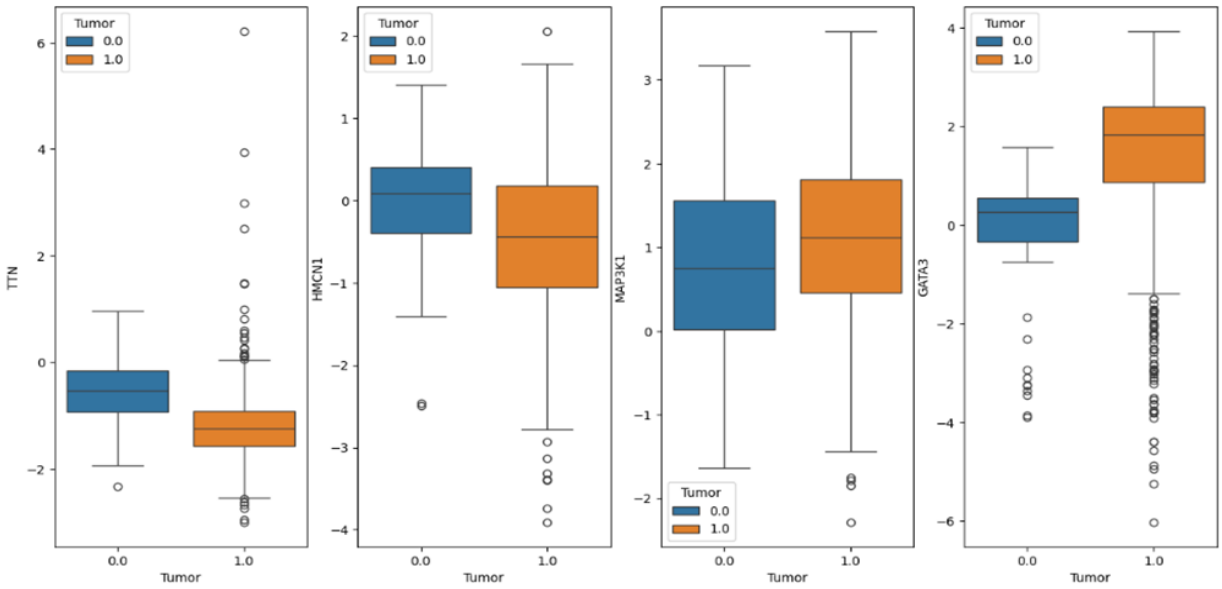

Рис. 3. Диаграмма размаха признаков TTN, HMCN1, MAP3K1, GATA3 (0- РМЖ отсутствует, 1- наличие РМЖ)

Среди отобранных встроенной функцией в методе случайного леса генов (HOOK3, SYP, C6orf57, RABIF, HRASL) статистически значимые различия в группах пациентов с РМЖ и здоровых наблюдаются в показателях экспрессии только одного гена RABIF ($p<0,05$).

Бинарная классификация проведена на тестовой выборке, включающей 317 наблюдений с показателями экспрессии генов, отобранных методами RF, RFE и методом экспертной оценки. В таблице 1 представлены усредненные оценки качества моделей классификации, полученные в результате 10-кратной кросс-валидации. RF, SVM, KNN показали наибольшую точность (accuracy=0,99) и полноту при отборе методом RFE. Худший результат классификации получен при использовании метода NB (accuracy=0,882, recall=0,888) по признакам, отобранным методом экспертной оценки.

Таблица 1

Сравнение классификационных моделей, построенных на основе показателей экспрессии генов, отобранных методами RF, RFE, экспертной оценки

Методы классификации	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1**
RF (HOOK3, SYP, C6orf57, RABIF, HRASL)				
NB	95,4	98,1	96,1	95,4
RF	**97,7**	**99,8**	**96,9**	**97,7**
SVM	**97,7**	**99,7**	**96,9**	**97,7**
KNN	97,7	97,4	96,1	97,6
RFE (COL10A1, PARP1 WISP1, MMP11, IGSF10)				
NB	98,1	99,8	99,0	98,1
RF	**99,8**	**99,9**	**99,0**	**99,3**
SVM	99,0	99,9	99,0	99,0
KNN	99,3	99,3	98,6	99,3
Экспертный метод (TP53, TTN, GATA3, MAP3K1, HMCN1)				
NB	88,2	88,8	89	88,2
RF	**96,1**	**94,2**	**94,6**	**94,3**
SVM	92,5	92,8	93,3	92,9
KNN	94,1	93,3	90,1	93,8

Обсуждение. Сравнительный анализ качества разработанных моделей классификации показал, что модели случайного леса (RF) по всем трем отобранным на первом этапе исследования группам признаков являются наиболее эффективными. Отметим, что с клинической точки зрения отобранные группы признаков являются релевантными для

диагностики РМЖ, что подтверждает потенциал этих генов как биомаркеров для диагностики данного заболевания. В дальнейшем планируется исследовать влияние различных способов устранения дисбаланса классов на качество разработанных моделей.

Заключение и выводы

В результате анализа базы данных TCGA обнаружена повышенная экспрессия генов COL10A1, PARP1, MMP11, WISP1, RABIF и сниженная экспрессия генов IGSF10, TTN, HMCN1, MAP3K1, GATA3 в тканях при раке молочной железы. Ранняя диагностика РМЖ по показателям экспрессии генов методами машинного обучения позволяет сократить время анализа результатов ДНК-чипирования, своевременно начать лечение и снизить летальность от РМЖ. Предложенные в данной работе методы отбора признаков и разработанные учебные модели классификации могут способствовать развитию системы поддержки принятия врачебных решений для диагностики РМЖ по показателям экспрессии генов.

ЛИТЕРАТУРА

1. Аскандирова А. Б. [и др.] Роль эпигенетических исследований в диагностике и лечении рака молочной железы // Онкология и радиология Казахстана. – 2019. – № 2(52). – С. 39–44.
2. Павлова Н. В. [и др.] Современное представление о факторах риска и механизмах развития рака молочной железы // Успехи молекулярной онкологии. – 2023. – Т. 10, № 3. – С. 15-23.
3. Телышева Т. В. Медико-генетическое консультирование, как основа профилактики наследственных болезней / Т. В. Телышева // Вестник научных конференций. – 2019. – № 1-3(41). – С. 106-108.
4. Ahmad F. K. et al. Filter-Based Gene Selection Method for Tissues Classification on Large Scale Gene Expression Data //International Journal of Engineering and Technology. – 2018. – Vol. 7. – №. 2.15. – P. 68-71.
5. Baliarsingh S. K. et al. Jaya optimized extreme learning machine for breast cancer data classification //Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2. – Springer Singapore, 2021. – P. 459-467.
6. Bhargava N. et al. Detection of Malicious Executables Using Rule Based Classification Algorithms //ICITKM. – 2017. – P. 35-38.
7. Cilia N. D. et al. An experimental comparison of feature-selection and classification methods for microarray datasets // Information. 2019. Vol. 10. №. 3. P. 109.
8. Eliyatkın N. et al. Molecular classification of breast carcinoma: from traditional, old-fashioned way to a new age, and a new way // The journal of breast health. 2015. Vol. 11. №. 2. P. 59.
9. Purbolaksono M. D. et al. Implementation of mutual information and bayes theorem for classification microarray data //Journal of Physics: Conference Series. – IOP Publishing, 2018. – Vol. 971. – №. 1. – P. 012011
10. Rajeshwari, J. Dermatology disease prediction based on firefly optimization of ANFIS classifier / J. Rajeshwari, M. Sughasiny // AIMS Electronics and Electrical Engineering. – 2022. – Vol. 6, No. 1. – P. 61-80.

11. Torre L. A. et al. Global cancer statistics, 2012 // CA: a cancer journal for clinicians. 2015. Vol. 65. №. 2. P. 87 –108.
12. Wu Q. et al. A feature selection method based on hybrid improved binary quantum particle swarm optimization //Ieee Access. – 2019. – Vol. 7. – P. 80588-80601.