АВТОМАТИЗАЦИЯ ПРОЦЕССА КЛАССИФИКАЦИИ РЕЗУЛЬТАТОВ ПРОФИЛЯ ЭКСПРЕССИИ МРНК ДЛЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В ПРАКТИКЕ ДИАГНОСТИКИ ПЛОСКОКЛЕТОЧНЫХ ИНТРАЭПИТЕЛИАЛЬНЫХ ПОРАЖЕНИЙ

Попова Г.М.,[1] Степанов В.Н.,[1] Мельникова Н.В.,[2] Антонова И.Б.,[2] Захаренко М.В.,[2] Боженко В.К.[2]

[1] ФГБУ» Институт проблем управления им. В. А. Трапезникова» РАН, Москва

[2] ФГБУ «Российский Научный Центр Рентгенорадиологии» Минздрава России, Москва

Аннотация: В работе предлагается методика компьютеризации дифференцированной диагностики плоскоклеточных интраэпителиальных поражении на основе нейросетевого классификатора результатов экспрессии мРНК 21 гена в практике молекулярного Пап-теста.

Ключевые слова: патология шейки матки, экспрессия генов, мРНК, амплификация генов, нейросетевая классификация.

Введение

В 2018г. в Российской Федерации рак шейки матки (РШМ) в поздних стадиях (III-IV) диагностирован в 32,6% случаев (2017 г. - 32,4%), что является высоким показателем запущенности [1]. В практике жидкостной цитологии подчеркивается необходимость дополнительной разработки оптимального рефлексного тестирования для пациенток различных возрастных групп [2]. Для изменения тенденции и повышения показателя активного выявления РШМ требуется переосмысление существующей диагностической стратегии и внедрения в клиническую практику достижений молекулярной биологии и компьютеризации диагностики. Поэтому изучение диагностических возможностей молекулярного Пап-теста является актуальной задачей.

В работе на основе мониторинга (сбор, обработка и анализ) данных (жидкостной цитологии, результатов анализа экспрессии мРНК большой группы генов, результатов гистологических исследований операционно-биопсийного материала) строится нейроклассификатор для повышения достоверности диагностики и раннего выявления предраковых процессов и рака шейки матки.

Модуль автоматизированной нейросетевой классификации плоскоклеточных интраэпителиальных поражений с функцией денормализации данных, полученных с детектирующего амплификатора «ДТ Прайм» компании «ДНК Технология»

Ранее в работе [3] показаны возможности построения нейросетевой модели классификации (НМК) плоскоклеточных интраэпителиальных поражений. Обучение НМК проводилось с применением

пакета Statistica Neural Networks. Анализируемые данные представляли собой результаты экспрессии мРНК 21 гена, полученные методом количественной ПЦР с использованием наборов ООО «ДНК-Технология», амплификатор «ДТ Прайм» компании «ДНК Технология» [4]. В качестве экспертных данных для обучения НМК использовались результаты гистологического исследования.

В настоящей работе для компьютеризации процесса дифференцированной диагностики плоскоклеточных интраэпителиальных поражений рассматривается автоматизированный нейросетевой классификатор с денормализованными данными, полученными с детектирующего амплификатора «ДТ Прайм».

Ретроспективно проанализированы результаты обследования 242 пациенток, проходивших лечение в РНЦРР. Во всех случаях выполнен Пап-тест CellPrep (Biodyne Co., LTD, Корея). Категоризация цитологических заключений выполнена согласно классификации Bethesda. В качестве референсного метода было выполнено гистологическое исследование операционного/биопсийного материала. При этом все пациентки были разделены на две клинически значимые группы: LSIL или доброкачественные изменения - (а) и HSIL+ - (б).

Во всех случаях определяли уровень экспрессии мРНК 21 гена (Ki-67, STK-15, CCNB1, CCND1, MYC, MYBL2, P16INK4A, PTEN, BIRC5, BCL2, BAG1, TERT, NDRG1, ESR1, PGR, HER2, GRB7, MGB1, MMP11, CTSL2, CD68) и трех генов-хаускипингов (GUSB, HPRT1, B2M) методом количественной ПЦР с использованием наборов ООО «ДНК-Технология» в материале консервирующей жидкости флакона после Пап-теста CellPrep [5].

Детектирующий амплификатор «ДТ Прайм» проводит нормализацию выходных данных, причем независимо для каждого гена. Это делает невозможным применение нейронных сетей для идентификации и классификации новых данных с этого прибора, так как соотношения значений генов могут меняться. Прибор «ДТ Прайм» накапливает результаты амплификации 21 гена по двум пробам (например, для каждого гена X_1 и X_2). Если полученное значение находится ниже порога детектирования, оно не выводится. Результирующее значение амплификации каждого гена вычисляется следующим образом: сначала вычисляется среднее значение двух проб:

$$X_C = \frac{X_1 + X_2}{2}$$

Затем вычисляется промежуточное значение X_A нормированное по генам хаускипингам коэффициентом N:

$$X_A = \frac{2^{\min(X_C) - X_C}}{N},$$

где коэффициент N представляет собой среднее геометрическое:

$$N = \sqrt[3]{B_1 * B_2 * B_3},$$

где

$$B_n = 2^{\min(X_C) - X_C}$$

Для генов хаускипингов соответственно B2M, HPRT1 и GUSB. Далее, финальное значение амплификации каждого гена вычисляется так:

$$X = \frac{X_A}{\min(X_A)}$$

Поскольку значение генов хаускипингов в наборе данных не представлено, обратно рассчитать исходные значения не представляется возможным.

Амплификатор накапливает данные и каждый раз проводит нормализацию, в том числе и старых данных. Это позволяет вычислить коэффициенты нормализации и провести денормализацию новых данных, для приведения их к форме, пригодной к классификации в ранее обученной нейронной сети. Нормализация выполняется по следующей формуле:

$$X_i^{'} = X_i * k,$$

где X_i – исходное значение гена, $X_i^{'}$ – нормализованное, k -линейный коэффициент, который вычисляется следующим образом:

$$k = \frac{1}{\min(X_1 \dots X_i \dots X_n)}; \quad X_i \neq 0$$

Поскольку значения X_i (до нормализации) не известны, для нахождения коэффициента k можно использовать данные, по которым проводилось обучение сети:

$$k = \frac{X_i^{'}}{X_i}$$

Коэффициенты вычисляются отдельно для каждого гена. После процедуры денормализации, соотношения величин генов приводятся к виду удобному для обучения нейронной сети, и при этом новые данные также становятся пригодными для идентификации и классификации нейросетевым методом.

Таким образом, для автоматизации процесса идентификации и классификации результатов экспрессии мРНК 21гена была сформирована и обучена нейронная сеть из 242 наблюдений с 21 входом для ввода независимых переменных (генов) и одним выходом, классифицирующим состояния пациентов а или б.

242 наблюдения с денормализованными результатами анализов экспрессии мРНК 21-генной панели подлежали анализу как при обучении, так и при контроле и тестировании системы. Критерием достоверности служили результаты сопоставления компьютерных оценок правильности разделения наблюдений на группы а и б с результатами гистологического исследования (т.е. с оценкой эксперта).

Для выбора варианта архитектуры сети использовался анализ автоматических нейронных сетей из пакета Statistica 10, позволивший:

- выбрать из множества разных типов нейронных сетей лучшую сеть и архитектуру модели для классификации наблюдений со 100% производительностью (долей правильно классифицированных наблюдений) на обучающей и контрольной выборках.
- получить графическое изображение схемы, сконструированной НС; матрицы классификации обучающей, контрольной, тестовой выборок и других новых выборок, с дифференцированным счетом правильно, неправильно и неоднозначно определенных наблюдений.

Обучение проводилось в режиме с кросс-проверкой на контрольном множестве наблюдений для оценки ошибки и тестового множества для сравнения альтернативных моделей, т.е. для независимого контроля качества сети со средней ошибкой, равной принятому в медицине значению $p_{теор} = 0{,}05$

В результате анализа 1500 разных нейронных сетей были выбраны 2 лучших модели трёхслойного персептрона с одинаковой архитектурой: 1. MLP 21-6-2 и 3. MLP 21-6-2 с 100% производительностью на обучающей [83**а**; 74**б**] и контрольной [30**а**; 13**б**] выборках; с одинаковой функцией активации скрытого слоя – Логистическая (её график имеет форму S-образной кривой, при этом выходные значения лежат в интервале (0, 1), что очень удобно); с одинаковой функцией активации выходного слоя – Софтмакс. Для обучения этих сетей был использован один и тот же алгоритм: BFGS (который разработали Бройден-Флетчер-Гольдфарб-Шанно) с функцией ошибок сумма квадратов, но оптимальное решение было найдено за разное количество обучающих эпох: 27 и 21 соответственно. Тестовые выборки [30**а**; 12**б**] двух нейронных сетей содержали по одной, причем идентичной ошибке (ошибочная интерпретация категории а как категория б), не совпадающей с оценками эксперта.

Для выбора рабочей модели был проведен анализ на способность этих моделей к обобщению, исходя из результатов жидкостной цитологии (данных мониторинга), выбрано 21 наблюдение, из которых 18 наблюдений совпали с результатами цитологического исследования, 3 результата не совпали. Оставили две модели для дальнейших исследований.

Если результаты нейронной сети не совпадают с результатами жидкостной цитологии и/или гистологии - данный пациент должен оставаться под динамическим наблюдением в мониторинге.

Разработанный модуль автоматизированной нейросетевой классификации плоскоклеточных интраэпителиальных поражений с функцией денормализации данных, получаемых с детектирующего амплификатора «ДТ Прайм», позволяет, для большей диагностической точности, использовать две разные нейронные сети, причем, объемы сетей могут пополняться новыми достоверными наблюдениями.

Заключение

Таким образом, разработанный модуль - автоматизированный нейросетевой классификатор, используемый в практике молекулярно-биологической диагностики предраковых заболеваний шейки матки по материалу Пап-теста CellPrep, позволяет с высокой точностью дифференцировать плоскоклеточные интраэпителиальные поражения, что должно учитываться при разработке национальной программы скрининга рака шейки матки.

Литература

1. *Под ред. А.Д. Каприна, В.В. Старинского, Г.В. Петровой* Состояние онкологической помощи населению России в 2018 году. М.: МНИОИ им. П.А. Герцена филиал ФГБУ «НМИЦ радиологии» Минздрава России, 2019. илл. - 236 с. ISBN 978-5-85502-250-6
2. *Мельникова Н.М., Яровая Н.Ю., Болотина Н.А., Антонова И.Б.* Диагностические возможности Пап-теста у женщин в постменопаузе. // Новости клинической цитологии России.2017. Т.21. №3-4. С. 15-17.
3. *Попова Г.М., Степанов В.Н., Мельникова Н.В., Антонова И.Б., Боженко В.К.* Нейросетевая модель классификации плоскоклеточных интраэпителиальных поражений в практике молекулярного Пап-теста //
4.Системный анализ и управление в биомедицинских системах. 2018. Т. 17 № 4. С. 964-971.
https://www.dna-technology.ru/sites/default/files/dtpraym_v11.pdf
5. *Мельникова Н.В., Боженко В.К., Антонова И.Б., Бабаева Н.А., Яровая Н.Ю., Болотина Н.А., Захаренко М.В., Сенчукова А.Л., Акопова Н.Б., Александрова Н.В., Бурменская О.В., Ашрафян Л.А.* Цервикальные интраэпителиальные неоплазии: анализ профиля мРНК в практике жидкостной цитологии. Акушерство и гинекология, 2017.-N 4.-С.95-100. DOI: 10.18565/aig.2017.4.95-100