Предвидување на молекуларен подтип на рак на дојка со машинско учење врз основа на генска експресија.

Ракот е кога абнормалните клетки почнуваат да се делат и растат на неконтролиран начин. Клетките можат да растат во околните ткива или органи и може да се прошират на други делови од телото. Сите видови на рак започнуваат во клетките. Нашите тела се составени од повеќе од (100.000.000.000.000) клетки. Ракот започнува со промени во една клетка или мала група клетки.

Што е рак на дојка?

Ракот на дојка е болест кај која абнормалните клетки на дојката растат неконтролирано и формираат тумори. Доколку не се контролираат, туморите можат да се шират низ целото тело и да станат фатални. Клетките на ракот на дојка започнуваат во млечните канали и/или лобулите на дојката што произведуваат млеко. Најраната форма (in situ) не е опасна по живот и може да се открие во раните фази. Клетките на ракот можат да се шират во блиското ткиво на дојката (инвазија). Ова создава тумори кои предизвикуваат грутки или задебелување.

Инвазивните карциноми можат да се шират во блиските лимфни јазли или други органи (метастазираат). Метастазите можат да бидат опасни по живот и фатални. Според податоците од *(World Health Organization)*, ракот на дојка е најчесто дијагностицираниот канцер кај жени на светско ниво. Раната детекција и правилната класификација на туморот се клучни за избор на терапија и подобрување на преживувањето.

Хистолошка наспроти молекуларна класификација

Традиционално, класификацијата на карциномот на дојка се потпира на хистолошките карактеристики, кои се фокусираат на морфолошкиот изглед на клетките и нивната архитектонска организација под микроскоп. Овој модел ги категоризира туморите во специфични типови, како што се инвазивен дуктален карцином (IDC) или инвазивен лобуларен карцином (ILC). Иако овој систем е во употреба со децении, се покажа дека хистологијата има ограничена моќ во предвидувањето на индивидуалниот одговор на терапијата кај пациенти со слични патолошки карактеристики (4).

Наспроти ова, молекуларната класификација воведе револуција преку анализа на генската експресија (microarray analysis). Овој пристап откри дека карциномот на дојка не е една болест, туку група од биолошки различни ентитети. Преку молекуларно профилирање, идентификувани се четири главни подтипови: Luminal A, Luminal B, HER2-enriched и Basal-like (4).

Главната разлика лежи во клиничката примена: додека хистолошкиот степен (grade) дава информации за диференцијацијата на клетките, молекуларните подтипови овозможуваат

многу попрецизно разбирање на биолошкото однесување на туморот. На пример, туморите кои изгледаат хистолошки слично, може да имаат драстично различни молекуларни профили, што директно влијае на нивната чувствителност на хемотерапија или ендокрина терапија (4). Според авторите Прат и Пероу, разбирањето на оваа молекуларна хетерогеност е клучно за развој на персонализирана медицина и подобрување на стапката на преживување кај пациентите (5).

Молекуларни подтипови и нивното значење за терапијата

Современиот пристап кон третманот на ракот на дојка се заснова на препознавање на неговата молекуларна разновидност. Истражувањата поддржани од BCRF потврдуваат дека определувањето на подтипот е критично за изборот на најефикасната терапија.

- Luminal A: Овој подтип се карактеризира со присуство на естрогенски (ER) и прогестеронски (PR) рецептори, но е негативен за HER2 протеинот. Овие тумори обично растат побавно и имаат најдобра прогноза.
- Luminal B: Слично како Luminal A, овие тумори се хормон-позитивни, но имаат повисоко ниво на протеинот Ki-67, што укажува на побрза делба на клетките.
- HER2-Enriched: Овие карциноми имаат дополнителни копии на генот HER2, што доведува до прекумерно производство на соодветниот протеин кој поттикнува раст на туморот. Иако се агресивни, тие многу добро реагираат на насочена (таргет) терапија како трастузумаб.
- Triple-Negative (Basal-like): Овој тип не поседува ниту еден од трите рецептори (ER, PR, HER2). Поради ова, хормонската и HER2-таргет терапијата не се ефикасни, па главниот метод на лечење останува хемотерапијата (6).

Генска експресија и биоинформатика

RNA-Seq и microarray технологии

Во ерата на геномиката, прецизното квантифицирање на транскриптомот — збирот на сите RNA молекули во една клетка — е клучно за разбирање на фенотипските промени кај болестите. Биоинформатиката игра централна улога во овој процес, трансформирајќи ги суровите биолошки податоци во значајни клинички информации. Двете доминантни технологии за оваа намена, Microarray и RNA-Seq, се засноваат на фундаментално различни принципи.

1. Microarray: Пристап заснован на хибридизација

Microarrsys беа првата технологија која овозможи паралелно испитување на илјадници гени. Овој метод користи стаклени или силиконски плочки со фиксирани ДНК сонди. Примерокот на RNA се претвора во флуоресцентно обележана cDNA, која потоа се хибридизира со сондите. Иако овој метод е економичен и брз, тој има сериозни ограничувања. Најзначајното е што микрочипот е „затворен систем“ — тој може да го

измери само она за што веќе има сонда. Доколку во примерокот постои нова мутација или непознат ген, микрочипот нема да го регистрира (1).

2. RNA-Seq: Револуцијата на секвенционирањето од следната генерација (NGS)

За разлика од микрочиповите, RNA-Seq користи масивно паралелно секвенционирање (Next-Generation Sequencing) за директно отчитување на нуклеотидната секвенца на секоја RNA молекула. Овој процес започнува со фрагментација на RNA, конверзија во библиотека од cDNA и секвенционирање на милиони кратки фрагменти (reads).

„RNA-seq нуди супериорна резолуција бидејќи овозможува детекција на нови егзони, фузиони гени и алтернативни настани на сплајсинг (alternative splicing), кои се невидливи за традиционалните методи засновани на хибридизација“ (1). Дополнително, RNA-Seq има екстремно широк динамички опсег, што овозможува прецизно мерење и на многу ниски и на многу високи нивоа на експресија во истиот примерок (2).

3. Биоинформатички предизвици и обработка на податоци

Биоинформатичкиот (pipeline) за RNA-Seq е значително покомплексен од оној за microarrays. Додека кај microarrays главниот предизвик е нормализацијата на интензитетот на флуоресценцијата, кај RNA-Seq податоците бараат неколку фази:

- Контрола на квалитет: Отстранување на лоши отчитувања.
- Порамнување (Alignment): Мапирање на милионите фрагменти врз референтниот човечки геном.
- Квантификација: Броење на фрагментите кои паѓаат врз одреден ген.
- Статистичка анализа: Идентификација на диференцијално експресирани гени (DEG) (2).

Современите истражувања покажуваат дека интеграцијата на овие податоци со напредни алгоритми за машинско учење овозможува идентификација на специфични биомаркери за карцином, што е невозможно со класичните хистолошки методи (2).

Примена на машинско учење (ML) и длабоко учење (DL) во онкологијата

Традиционално машинско учење

Традиционалното машинско учење се заснова на користење математички алгоритми за наоѓање патерни во податоците. Во онкологијата, ова најчесто значи анализа на табеларни податоци каде секој ген или клинички параметар претставува една променлива.

- Класификација и регресија: Алгоритмите како Random Forest (Шума на одлучување) и Support Vector Machines (SVM) се користат за да се одреди дали

одреден генетски профил припаѓа на агресивен или на индолентен (помирен) тип на рак. Овие модели се исклучително ефикасни кога располагаме со помал број на пациенти, но со многу прецизни лабораториски мерења.
- Идентификација на биомаркери: Преку процеси како "feature selection", машинското учење помага да се издвојат само 5 или 10 клучни гени од вкупно 20.000, кои се најрелевантни за прогнозата на болеста. Ова драстично ги намалува трошоците за дијагностичките тестови.
- Предикција на токсичност: ML моделите можат да предвидат како пациентот ќе реагира на хемотерапија, односно дали постои висок ризик од тешки несакани ефекти врз основа на неговиот метаболички профил.

Во онколошките студии често се користат следниве методи:

- Support Vector Machines (SVM) и Random Forest: Овие алгоритми често се користат за класификација на туморите врз основа на податоци од генска експресија (на пр. разликување на бениген од малиген тумор).
- Логистичка регресија: Се користи за проценка на веројатноста за преживување на пациентот врз основа на клинички и геномски маркери.
- Ограничување: Традиционалното машинско учење бара претходно процесирање на податоците и избор на најрелевантните гени за да се избегне „преоптоварување“ на моделот (1).

Длабоко учење

Длабокото учење претставува напредна подгрупа на машинското учење која користи повеќеслојни вештачки невронски мрежи. За разлика од традиционалните методи, DL може автоматски да учи сложени обрасци директно од суровите податоци.

- Конволуциски невронски мрежи (CNN): можат да "скенираат" илјадници дигитални микроскопски снимки и да препознаат микро-метастази во лимфните јазли кои патолозите лесно можат да ги превидат поради замор или субјективност.
- Интеграција на "Multi-omics": Длабокото учење е моќно во комбинирање на податоци од различни извори — на пример, спојување на податоци од RNA-Seq со епигенетски промени и протеомика за да се добие попрецизна молекуларна слика на карциномот (1).

Иако моќни, овие технологии се соочуваат со предизвикот на „црната кутија“ (black box). Кај длабокото учење, често е тешко да се разбере *зошто* моделот донел одредена одлука, што е клучен проблем во медицината каде етиката и транспарентноста се приоритет. Затоа, новиот фокус на биоинформатичарите е Explainable AI (XAI) – создавање на модели кои не само што ќе дадат дијагноза, туку и ќе објаснат кои биолошки фактори довеле до неа.

Методолошки предизвици и ограничувања во геномските истражувања

И покрај револуционерниот напредок во технологиите за секвенционирање и машинското учење, постојат фундаментални технички и статистички бариери кои можат да доведат до погрешни заклучоци доколку не се соодветно адресирани.

1. Проблемот со прекумерно прилагодување (Overfitting)

Overfitting е можеби најголемиот непријател на моделите за машинско учење во онкологијата. Овој феномен се случува кога еден алгоритам станува „премногу добар“ во препознавање на специфичностите на една конкретна база на податоци, до тој степен што ги меморира случајните статистички варијации (шум) наместо вистинските биолошки сигнали.

- Зошто се случува? Во геномските студии, често имаме „проблем на димензионалност“ ($p >> n$). Имаме илјадници гени (променливи), а релативно мал број на пациенти (примероци). Моделот лесно наоѓа случајни корелации кои не постојат во реалниот свет.
- Последица: Кога овој модел ќе се примени на нови пациенти во клиниката, неговата точност драстично опаѓа бидејќи тој научил како да го препознае само тој специфичен сет на податоци врз кој бил трениран.

2. Ефекти на серија (Batch Effects)

„Batch effects“ се небиолошки варијации во податоците кои се појавуваат кога примероците се обработуваат во различни лаборатории, во различни денови или од страна на различни техничари.

- Извор на грешка: Температурата во лабораторијата, квалитетот на реагенсите или дури и времето поминато од земањето на биопсијата до замрзнувањето на примерокот може да ја промени генската експресија.
- Проблем: Биоинформатичките алатки може погрешно да ги интерпретираат овие технички разлики како вистински биолошки разлики помеѓу здрави и болни пациенти. Ако сите контролни групи се обработени во јануари, а сите болни групи во март, моделот може да ги анализира разликите во сезоната наместо болеста.

3. Дискорданца помеѓу платформите (RNA-Seq vs Microarray)

Иако и двете технологии ја мерат генската експресија, тие често даваат резултати кои не се целосно компатибилни. Ова го прави комбинирањето на постарите студии (базирани на microarrays) со најновите (RNA-Seq) исклучително тешко.

- Различни мерни единици: Microarray користи интензитет на флуоресценција, додека RNA-Seq користи број на прочитани секвенци (read counts).
- Динамички опсег: RNA-Seq може да детектира екстремно ниски нивоа на експресија кои за микрочипот се невидливи. Поради ова, мета-анализите кои се обидуваат да ги обединат овие податоци често се соочуваат со „статистички шум“ кој ги маскира клучните биомаркери.

4. Недостаток на надворешна валидација

Многу научни трудови објавуваат „висока точност“ на нивните модели, но таа точност е тестирана само во рамките на нивната лабораторија.

- Внатрешна vs Надворешна валидација: Внатрешната валидација (на пр. cross-validation) е добра почетна точка, но не е доволна. За еден биомаркер да биде клинички корисен, тој мора да ги даде истите резултати на сосема независна група пациенти од друг континент, со друга етничка припадност и обработени со друга опрема.
- Пречка за примена: Без робусна надворешна валидација, лекарите не можат да имаат доверба во предвидувањата на вештачката интелигенција, што е главната причина зошто многу ветувачки модели никогаш не влегуваат во болничка примена.

Претходни студии за примена на машинско учење во класификација на молекуларни подтипови на рак на дојка

Во последната деценија, примената на машинско учење и длабоко учење во анализата на генска експресија доведе до значителен напредок во автоматската класификација на молекуларните подтипови на рак на дојка. Повеќето современи студии се базираат на податоци добиени од големи јавно достапни геномски конзорциуми како што се The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO) и METABRIC, кои содржат илјадници примероци со молекуларни и клинички информации.

Раните истражувања во оваа област се фокусираа на традиционални алгоритми за машинско учење како Support Vector Machines (SVM), Random Forest и логистичка регресија. Овие модели се покажаа како особено ефикасни во ситуации каде бројот на карактеристики (гени) е значително поголем од бројот на примероци, што е типично за геномските податоци ($p >> n$ проблем). Со примена на техники за селекција на карактеристики (feature selection), истражувачите успеале да идентификуваат релативно мал сет на биомаркери кои со висока точност ги разликуваат подтиповите Luminal A, Luminal B, HER2-enriched и Basal-like.

Во поновите студии, како онаа на Gupta (1), се применуваат техники на длабоко учење врз microarray податоци со цел автоматска екстракција на сложени нелинеарни обрасци во генската експресија. Авторите покажуваат дека повеќеслојните невронски мрежи можат да

постигнат висока точност во класификацијата, но истовремено нагласуваат дека моделите се подложни на прекумерно прилагодување (overfitting), особено кога се тренираат на ограничен број примероци. Дополнително, тие посочуваат дека иако точноста е висока во рамките на истата база на податоци, перформансите може да се намалат при примена на независен dataset.

Во трудот на Kallah-Dagadu (2), објавен во списанието Scientific Reports, авторите применуваат интерпретабилни алгоритми како XGBoost во комбинација со Explainable AI техники. Нивниот пристап овозможува не само класификација на туморите, туку и идентификација на клучните гени кои придонесуваат за донесената одлука. Овој тренд укажува на зголемен интерес кон транспарентноста на моделите, особено во медицински контекст каде довербата и објаснивоста се од суштинско значење.

Покрај класичната класификација на подтипови, одредени студии се насочени кон интеграција на мулти-омикс податоци (геномика, транскриптомика, протеомика) со цел добивање поцелосна молекуларна слика. Длабокото учење, особено автоенкодери и хибридни невронски архитектури, се користи за намалување на димензионалноста и комбинирање на различни извори на податоци. Иако овие пристапи покажуваат потенцијал, нивната комплексност и потребата од големи количини податоци претставуваат значајна бариера за широка клиничка примена.

Сепак, и покрај охрабрувачките резултати, значителен дел од објавените студии остануваат на ниво на експериментални лабораториски анализи. Во многу случаи, моделите се развиени и тестирани исклучиво во контролирана академска средина, без интеграција во реални клинички информациски системи. Дополнително, ретко се адресира проблемот на динамичко ажурирање на моделите врз основа на нови клинички податоци и повратни информации од лекарите.

Овие ограничувања ја нагласуваат потребата од премин од чисто алгоритамска оптимизација кон развој на интегрирани клинички алатки кои комбинираат машинско учење, објаснивост и интерактивност со медицински персонал.

Идентификуван истражувачки јаз (Research Gap)

И покрај значителниот напредок во примената на машинско и длабоко учење за анализа на генска експресија кај рак на дојка, постојат повеќе структурни и методолошки недостатоци во постојната литература.

Прво, поголемиот дел од студиите се фокусираат исклучиво на оптимизација на перформансите на моделите (точност, AUC, F1-score), без да ја разгледаат нивната практична интеграција во клинички работни процеси. Моделите често се презентираат како изолирани алгоритми, без развиен систем кој овозможува интеракција со лекарите, внесување на нови пациенти или динамичко управување со различни верзии на моделот.

Второ, иако се споменува важноста на надворешната валидација, релативно мал број студии имплементираат механизми за континуирано учење (continual learning). Во реална клиничка пракса, нови податоци се генерираат секојдневно, а статичен модел обучен еднаш може постепено да ја изгуби својата релевантност поради промени во популацијата, дијагностичките протоколи или лабораториските техники. Недостатокот на human-in-the-loop архитектури претставува значајна празнина во постојната литература.

Трето, иако Explainable AI добива сè поголемо внимание, во многу студии објаснивоста се третира како дополнителна анализа, а не како интегрален дел од системот. Во клинички контекст, способноста да се објасни зошто одреден тумор е класифициран како Basal-like или Luminal B не е луксуз, туку неопходност за донесување терапевтска одлука.

Четврто, постои недостаток на апликативни платформи кои овозможуваат избор помеѓу повеќе модели и споредба на нивните предвидувања во реално време. Повеќето трудови анализираат еден алгоритам во изолација, без да обезбедат инфраструктура за споредба и верзионирање на различни модели во продукциска средина.

Врз основа на овие согледувања, може да се идентификува јасен истражувачки јаз: недостиг на интегриран, интерактивен и адаптивен систем за класификација на молекуларни подтипови на рак на дојка, кој комбинира машинско учење, објаснивост и механизам за повратна информација од лекарите.

Предложената работа има за цел да го адресира овој јаз преку развој на AI-базиран клинички систем за поддршка на одлуки кој овозможува:

- избор на различни обучени модели за предикција,
- внес на индивидуален пациент или серија пациенти,
- добивање автоматска класификација на молекуларниот подтип,
- евидентирање на одлуката на лекарот (прифаќање или корекција),
- акумулација на нови податоци за последователно дообучување на моделот,
- и можност за објаснување на предвидувањата преку интерпретабилни техники.

Со ова, истражувањето не се ограничува само на алгоритамска оптимизација, туку претставува чекор кон практична имплементација на машинско учење во клиничката онкологија, со фокус на транспарентност, адаптивност и интерактивност.

Користена литература

1. *Gupta S, Gupta MK, Shabaz M, Sharma A. Deep learning techniques for cancer classification using microarray gene expression data. Front Physiol. 2022 Sep 30;13:952709. doi: 10.3389/fphys.2022.952709. PMID: 36246115; PMCID: PMC9563992.*
2. *Kallah-Dagadu, G., Mohammed, M., Nasejje, J.B. et al. Breast cancer prediction based on gene expression data using interpretable machine learning techniques. Sci Rep **15**, 7594 (2025). https://doi.org/10.1038/s41598-025-85323-5*
3. *World Health Organization (WHO). Breast cancer fact sheets.*
4. *Malhotra GK, Zhao X, Band H, Band V. Histological, molecular and functional subtypes of breast cancers. Cancer Biol Ther. 2010 Nov 15;10(10):955-60. doi: 10.4161/cbt.10.10.13879. Epub 2010 Nov 15. PMID: 21057215; PMCID: PMC3047091.*
5. *Prat A, Perou CM. Deconstructing the molecular heterogeneity of breast cancer. Mol Oncol. 2011;5(1):5-23.*
6. *Breast Cancer Research Foundation. Molecular Subtypes of Breast Cancer [Internet]. New York: BCRF; [cited 2024 May]. Available from: https://www.bcrf.org/about-breast-cancer/molecular-subtypes-breast-cancer/*