روند اجرایی پروژه درس مقدمه‌ای بر بیوانفورماتیک – گروه ۱۸

بررسی، بازتولید و بهبود الگوریتم تشخیص سرطان پستان با استفاده از میکروRNAهای سرمی

(تحلیل مقاله: Novel combination of serum microRNA for detecting breast cancer in the early stage)

علی هاشمیان

محمدمهدی عابدیان

امیرعلی شعیری

علیرضا کاظمی

۱. چکیده (Abstract)

سرطان پستان یکی از شایع‌ترین علل مرگ‌ومیر در زنان است و تشخیص زودهنگام آن نقشی حیاتی در بقای بیمار دارد. روش‌های فعلی نظیر ماموگرافی دارای محدودیت‌هایی مانند درد، تهاجم نسبی و کاهش دقت در بافت‌های متراکم هستند. این گزارش به بررسی مطالعه‌ای می‌پردازد که در آن از پروفایل بیان میکروRNAهای (miRNA) موجود در سرم خون به عنوان یک روش غیرتهاجمی برای تشخیص سرطان استفاده شده است. در مطالعه مرجع (Shimomura et al., 2016)، با بررسی بیش از ۴۰۰۰ نمونه بالینی، ترکیبی از ۵ میکروRNA شناسایی شد که قادر است سرطان پستان را با حساسیت ۹۷,۳٪ تشخیص دهد. در این پروژه، علاوه بر تشریح کامل متدولوژی مقاله، ما با استفاده از داده‌های خام (GSE73002)، نتایج مقاله را بازتولید کرده و سپس با به‌کارگیری الگوریتم‌های هوش مصنوعی پیشرفته (XGBoost)، مدلی با ویژگی (Specificity) بالاتر توسعه داده‌ایم.

۲. مقدمه و بیان مسئله

تشخیص سرطان در مراحل اولیه (Early Stage) چالش اصلی انکولوژی است. ماموگرافی به عنوان استاندارد طلایی، برای برخی زنان دردناک است و نرخ مثبت کاذب قابل توجهی دارد. از سوی دیگر، بیوپسی مایع (Liquid Biopsy) با تمرکز بر میکروRNAهای در گردش خون، افق جدیدی را روشن کرده است. miRNAها مولکول‌های پایداری هستند که تغییر بیان آن‌ها با تومورزایی مرتبط است. هدف این مطالعه، شناسایی یک پنل اختصاصی از miRNAهاست که بتواند با دقت بالا، حتی کارسینوم درجا (Stage 0) را شناسایی کرده و جایگزینی برای غربالگری‌های سنتی باشد.

۳. داده‌ها و جامعه آماری (Data Description)

داده‌های مورد استفاده در این مطالعه یکی از بزرگترین مجموعه‌های داده در حوزه ترانسکریپتومیکس سرم است که تحت شناسه GSE73002 در پایگاه داده NCBI GEO در دسترس می‌باشد.

حجم کل نمونه‌ها: ۴۱۱۶ نمونه سرم.

تفکیک نمونه‌ها:

بیماران مبتلا به سرطان پستان (۱۲۸۰ نفر).

گروه کنترل سالم و بیماران غیرسرطانی (۲۸۳۶ نفر).

سایر انواع سرطان‌ها (پانکراس، معده، روده و ...) و بیماری‌های خوش‌خیم (برای تست اختصاصیت).

دسته‌بندی کوهورت: داده‌ها به دو بخش آموزش (Training) جهت کشف بیومارکر و آزمون (Test) جهت اعتبارسنجی مستقل تقسیم شدند.

۴. روش کار (Methodology) – تحلیل مقاله اصلی

فرآیند کشف بیومارکر در مقاله اصلی طی مراحل دقیق زیر انجام شده است:

۴-۱. جمع‌آوری و نگهداری نمونه‌ها

نمونه‌های سرم قبل از هرگونه درمان جراحی یا شیمی‌درمانی جمع‌آوری شدند. نکته مهم تکنیکال این مطالعه، بررسی پایداری نمونه‌ها در شرایط دمایی مختلف (۸۰- درجه برای گروه آموزش و ۲۰- درجه برای گروه آزمون) بود تا استحکام (Robustness) مدل اثبات شود.

۴-۲. تکنولوژی سنجش بیان (Microarray Analysis)

از پلتفرم بسیار حساس 3D-Gene™ (Toray Industries) استفاده شد که قابلیت شناسایی ۲۵۵۵ نوع miRNA را دارد. این تکنولوژی به دلیل نویز پس‌زمینه کم و حساسیت بالا انتخاب گردید.

۴-۳. پیش‌پردازش داده‌ها و نرمال‌سازی

حذف نویز: داده‌هایی که شدت سیگنال آن‌ها کمتر از میانگین کنترل منفی (به علاوه ۲ انحراف معیار) بود، حذف شدند.

نرمال‌سازی: برای یکسان‌سازی داده‌ها، از ۳ میکروRNA مرجع داخلی (miR-149-3p، miR-2861 و miR-4463) استفاده شد که بیان آن‌ها در تمام نمونه‌ها ثابت بود.

۴-۴. استراتژی انتخاب ویژگی (Feature Selection Pipeline)

برای رسیدن از ۲۵۵۵ ژن به ۵ ژن نهایی، مراحل زیر طی شد:

فیلتر شدت: انتخاب ژن‌هایی که در بیش از ۵۰٪ نمونه‌ها بیان قوی داشتند.

تست آماری تک‌متغیره: استفاده از t-test با اصلاحیه بونفرونی (Bonferroni correction) برای یافتن ژن‌های با اختلاف معنی‌دار ($P < 0.01$).

انتخاب چندمتغیره (Stepwise LDA): استفاده از آنالیز افتراقی خطی فیشر (Fisher's LDA). در این مرحله، الگوریتم به صورت گام‌به‌گام ترکیب‌های مختلف ژن‌ها را بررسی کرد تا تیمی از ژن‌ها را بیابد که "مکمل" یکدیگر باشند و بیشترین تفکیک را ایجاد کنند.

۴-۵. مدل نهایی و فرمول تشخیصی

مدل نهایی روی ۵ میکروRNA همگرا شد:

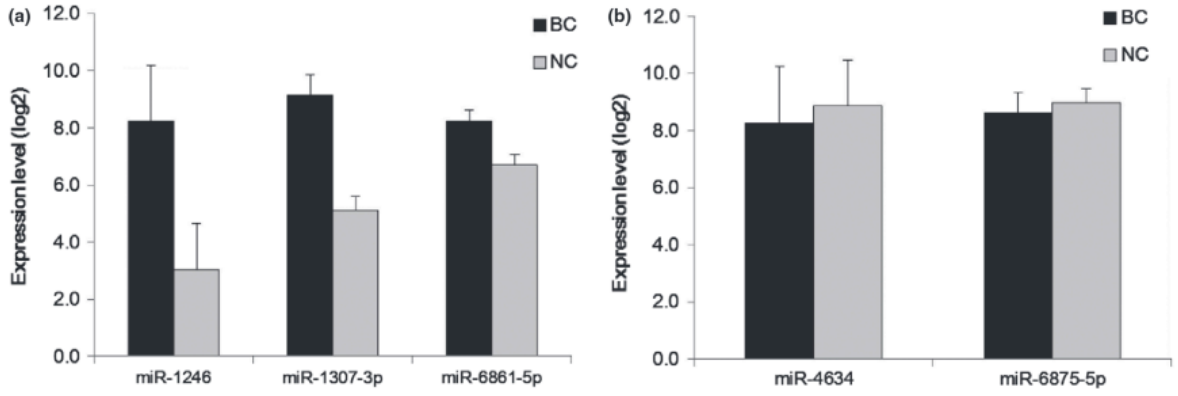
miR-1246 (افزایشی - مهم‌ترین مارکر)

miR-1307-3p (افزایشی)

miR-4634 (کاهشی)

miR-6861-5p (افزایشی)

miR-6875-5p (کاهشی)



فرمول محاسبه شاخص تشخیصی (Index) به صورت زیر استخراج شد:

$$\text{Index} = (0.25\times\text{miR-1246}) + (0.49\times\text{miR-1307-3p}) - (1.06\times\text{miR-4634}) + (1.89\times\text{miR-6875-5p}) + (0.31\times\text{miR-6861-5p}) - 13.94$$

۵. نتایج و یافته‌های کلیدی (Original Results)

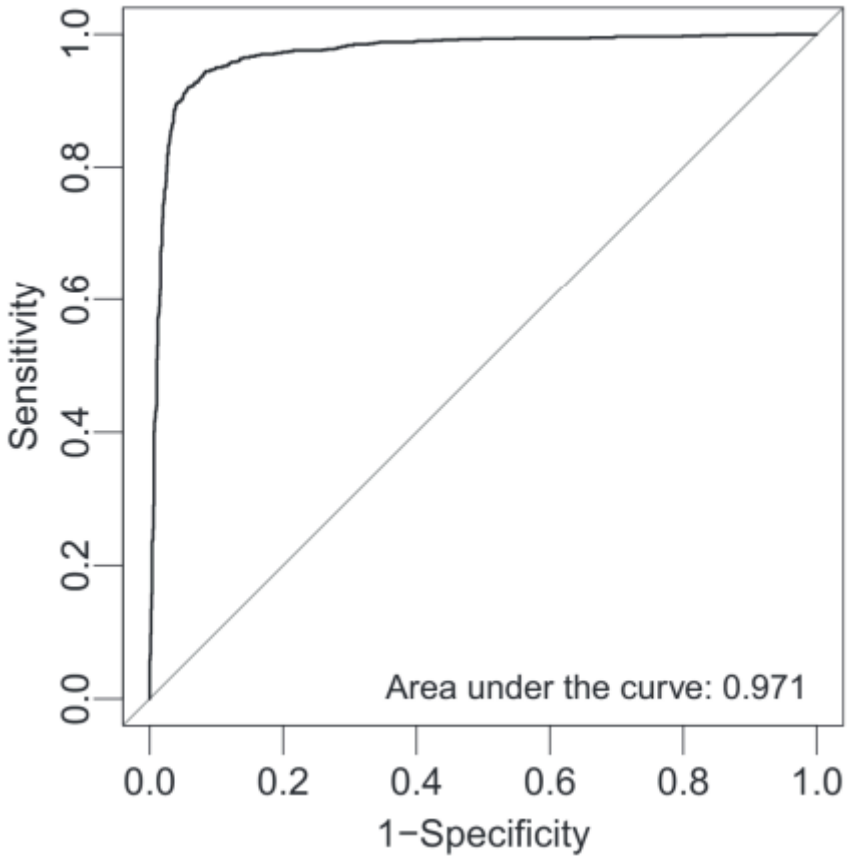
عملکرد مدل در کوهورت مستقل (Test Cohort) بسیار درخشان بود:

حساسیت(Sensitivity): ٪۹۷٫۳

ویژگی(Specificity): ٪۸۲٫۹

دقت کلی(Accuracy): ٪۸۹٫۷

سطح زیر منحنی(AUC): ۰٫۹۷۱



نکته مهم: این مدل توانست ٪۹۸ از بیماران مبتلا به Stage 0 (کارسینوم درجا) را شناسایی کند که نشان‌دهنده قدرت بالای آن در غربالگری زودهنگام است.

۶. پیاده‌سازی عملی و نوآوری گروه (Project Implementation)

در این پروژه، ما کدنویسی و تحلیل داده‌ها را در محیط پایتون انجام دادیم که شامل سه فاز بود:

فاز ۱: بازتولید (Replication): با پیاده‌سازی مجدد مدل LDA روی داده‌های اصلی، به AUC برابر با ۰,۹۶۶ رسیدیم که صحت نتایج مقاله را تایید کرد.

فاز ۲: اعتبارسنجی فرمول (Verification): اعمال مستقیم فرمول مقاله روی داده‌های خام، AUC برابر با ۰,۹۴۷ را نشان داد که قدرت تعمیم‌پذیری بیومارکرها را اثبات می‌کند.

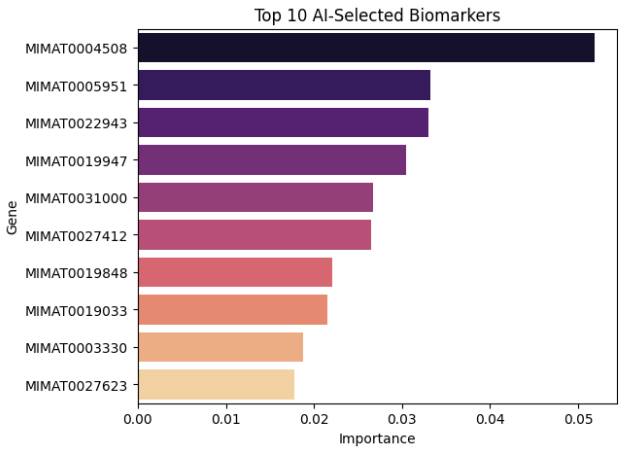
فاز ۳: نوآوری با هوش مصنوعی (AI Innovation):

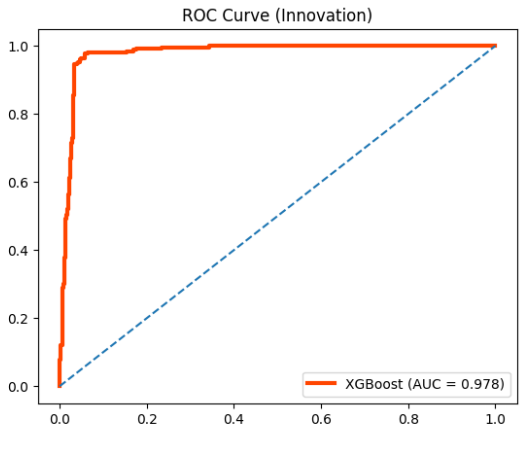
ما از الگوریتم یادگیری ماشین XGBoost (Gradient Boosting) به جای روش خطی LDA استفاده کردیم.

نتیجه: مدل هوشمند ما توانست ویژگی (Specificity) را از ۸۲,۹٪ (در مقاله) به ۹۵,۶٪ ارتقا دهد، در حالی که حساسیت همچنان بالای ۹۵٪ باقی ماند. این بهبود چشمگیر، نرخ مثبت کاذب را در غربالگری کاهش می‌دهد.

شاخص ارزیابی	گزارش مقاله (Original)	مدل بازتولید شده (Replication)	مدل پیشنهادی هوشمند (XGBoost)
AUC (قدرت تفکیک)	0.971	0.966	🏆 **0.978**
Accuracy (دقت کل)	89.7%	93.9%	🏆 **95.5%**
Sensitivity (حساسیت)	97.3%	90.0%	**95.3%**
Specificity (ویژگی)	82.9%	97.4%	🏆 **95.6%**

عملکرد مدل جدید ابتکاری ما:





۷. جمع‌بندی نهایی

این مطالعه نشان داد که پروفایل میکروRNAهای سرمی پتانسیل بالایی برای جایگزینی یا تکمیل روش‌های غربالگری فعلی دارد. ترکیب ۵ تایی معرفی شده، به ویژه برای تشخیص مراحل اولیه بیماری کارآمد است. همچنین، پیاده‌سازی عملی ما نشان داد که استفاده از الگوریتم‌های مدرن هوش مصنوعی می‌تواند نقاط ضعف مدل‌های آماری کلاسیک (مانند ویژگی پایین) را برطرف کرده و راه را برای تست‌های بالینی دقیق‌تر هموار سازد.

۸. پیوست و منابع (Appendix & References)

دسترسی به کدها و مستندات پروژه:

تمامی کدهای مربوط به تحلیل داده‌ها، بازتولید نتایج و مدل‌های هوش مصنوعی پیاده‌سازی شده در این پژوهش، به صورت متن‌باز (Open Source) در مخزن گیت‌هاب زیر در دسترس می‌باشد:

https://github.com/hashemian1382/Breast-Cancer-miRNA-Analysis-2026/tree/master

مقاله مرجع اصلی:

Shimomura, A., Shiino, S., Kawauchi, J., Takizawa, S., Sakamoto, H., Matsuzaki, J., ... & Ochiya, T. (2016). Novel combination of serum microRNA for detecting breast cancer in the early stage. Cancer Science, 107(3), 326-334.