پروژه درس – سال تحصیلی ۱۴۰۴

طبقه‌بندی انواع سرطان با استفاده از داده‌های بیان ژن RNA-Seq و یادگیری ماشین

Machine Learning Approach to Identify Significant Genes and Classify Cancer Types from RNA-Seq Data

ارائه‌دهنده: مجتبی ریاحی

نام درس: ژنومیک محاسباتی

استاد درس: دکتر قنبری

سال: ۲۰۲۵

RNA-SEQ · ML · BIOINFORMATICS

5 CANCER CLASSES · 20,531 GENES

بر اساس مقاله: Akter, Adesola & Basnet (2025) — Global Medical Genetics

01 مقدمه

چرا طبقه‌بندی سرطان با RNA-Seq؟

سرطان یکی از مهم‌ترین علل مرگ‌ومیر در جهان است و تشخیص سریع و دقیق آن نقش حیاتی در موفقیت درمان دارد. فناوری RNA-Seq امکان اندازه‌گیری میزان بیان هزاران ژن را به‌طور همزمان فراهم کرده است.

- سرطان یکی از مهم‌ترین علل مرگ‌ومیر در جهان است
- تشخیص سریع و دقیق نقش حیاتی در درمان دارد
- فناوری RNA-Seq امکان اندازه‌گیری میزان بیان هزاران ژن را فراهم می‌کند
- الگوی بیان ژن می‌تواند نوع سرطان را مشخص کند

~10 میلیون
مرگ سالانه ناشی از سرطان (2022)

20,531
ژن در هر نمونه RNA-Seq

5 نوع
سرطان قابل طبقه‌بندی

02 بیان مسئله

چالش‌های داده‌های بیان ژن

چرا طبقه‌بندی RNA-Seq دشوار است؟

CH-01 ابعاد بالا

تعداد بسیار زیاد ویژگی‌ها (بیش از ۲۰ هزار ژن) که تحلیل و انتخاب ویژگی را پیچیده می‌کند.

CH-02 نمونه محدود

تعداد نمونه نسبتاً کم (۸۰۱ نمونه) نسبت به تعداد ژن‌ها که خطر Overfitting را افزایش می‌دهد.

CH-03 همبستگی ژن‌ها

همبستگی بالا بین ژن‌ها موجب Multicollinearity و کاهش پایداری مدل می‌شود.

CH-04 خطر Overfitting

نسبت بالای ویژگی به نمونه، احتمال بیش‌برازش مدل‌های یادگیری ماشین را بالا می‌برد.

راه‌حل انتخاب ویژگی (Feature Selection) + اعتبارسنجی متقاطع (Cross-Validation)

03 هدف پژوهش

اهداف مطالعه

1. **طبقه‌بندی ۵ نوع سرطان با استفاده از داده‌های RNA-Seq**
   BRCA, KIRC, COAD, LUAD, PRAD
2. **شناسایی ژن‌های مهم و اثرگذار (بیومارکرها)**
   کاهش از ۲۰۵۳۱ ژن به مجموعه‌ای کوچک و تفسیرپذیر
3. **مقایسه چند الگوریتم یادگیری ماشین**
   هشت مدل طبقه‌بندی شامل SVM، KNN، ANN، RF و غیره
4. **انتخاب بهترین مدل طبقه‌بندی**
   بر اساس دقت و پایداری در Cross-Validation

معیار موفقیت

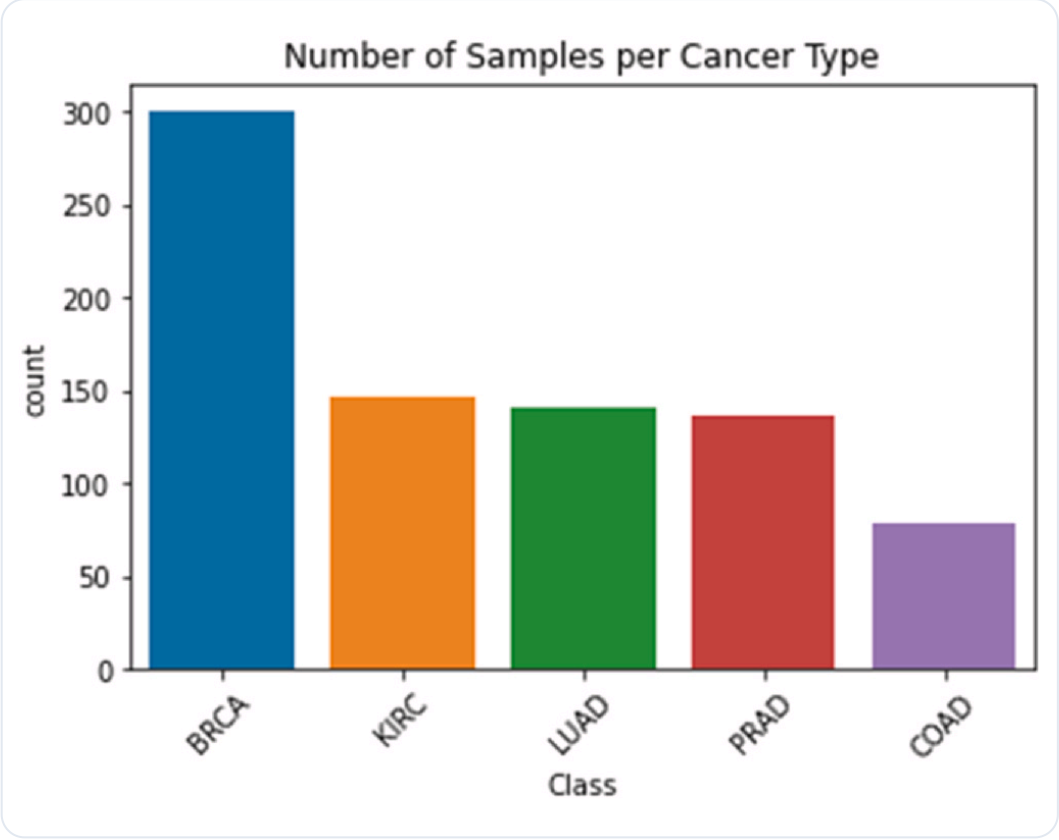
۹۹.۸۷٪

بالاترین دقت به‌دست‌آمده با SVM در 5--Fold Cross Validation

دقت | پایداری | قابلیت تفسیر

04 معرفی داده‌ها

داده‌های RNA-Seq مورد استفاده



شکل ۱ — فراوانی هر نوع سرطان در مجموعه داده PANCAN (منبع: مقاله)

مشخصات کلی

منبع داده:	UCI / TCGA (PANCAN)
پلتفرم:	Illumina HiSeq
تعداد نمونه‌ها:	۸۰۱ نمونه
تعداد ژن‌ها:	20,531 ژن
تعداد کلاس‌ها:	۵ نوع سرطان

انواع سرطان

- BRCA — سینه
- KIRC — کلیه
- COAD — کولون
- LUAD — ریه
- PRAD — پروستات

05 انتخاب ویژگی (Feature Selection)

کاهش ابعاد با روش‌های منظم‌سازی

برای کاهش ابعاد داده‌های RNA-Seq و رفع خطر Overfitting، از سه روش انتخاب ویژگی مکمل هم استفاده شد:

Lasso Regression L1

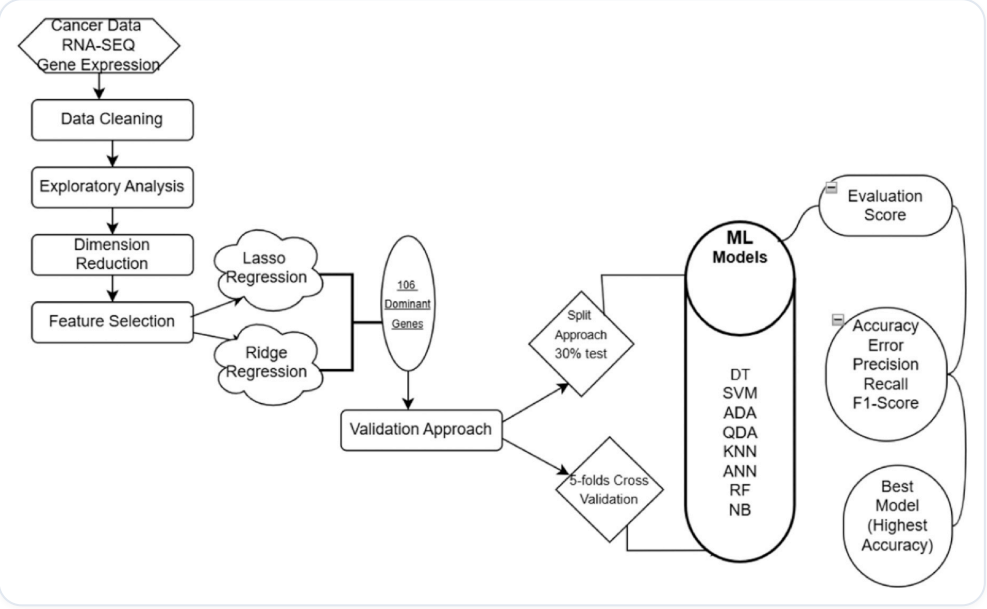
جریمه L1 — برخی ضرایب را دقیقاً به صفر می‌رساند و انتخاب ویژگی خودکار انجام می‌دهد. مناسب داده‌های پربعد ژنومی.

Ridge Regression L2

جریمه L2 — ضرایب بزرگ را تنبیه می‌کند و Multicollinearity را کاهش می‌دهد. پایداری مدل را بالا می‌برد.

Random Forest Importance RF

اهمیت ویژگی بر اساس کاهش ناخالصی Gini در درختان جنگل. شناسایی ژن‌های اثرگذار.



شکل ۲ — پایپ‌لاین تحلیل: از داده خام تا طبقه‌بندی (منبع: مقاله)

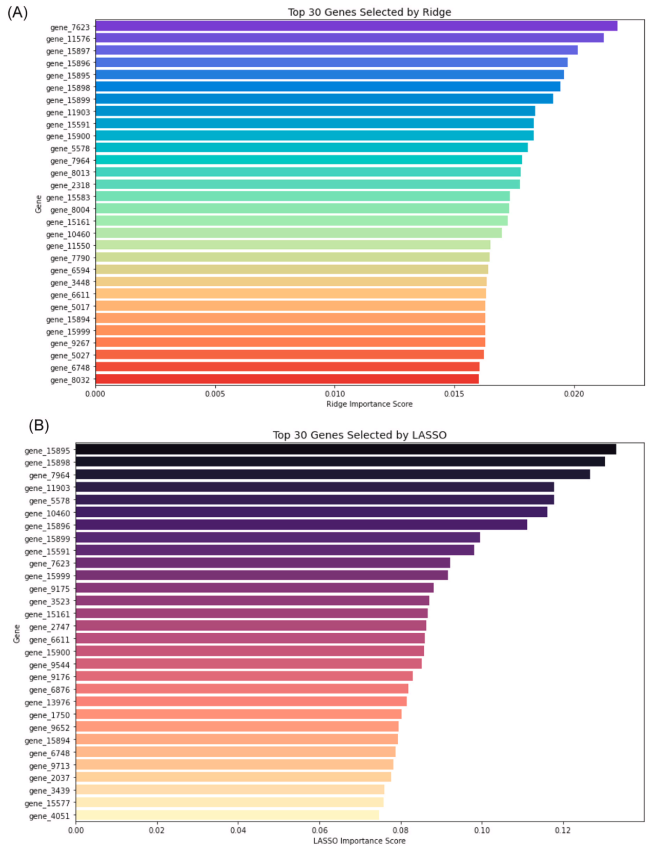
06 فرآیند انتخاب ژن‌ها

از ۲۰,۵۳۱ ژن به ۱۰۶ ژن کلیدی

مراحل کاهش ویژگی‌ها

1. **انتخاب ۲۵۰ ژن برتر با Lasso و Ridge**
   شناسایی ژن‌های با بیشترین اهمیت توسط هر روش
2. **بررسی همپوشانی با نتایج Random Forest**
   اشتراک‌گیری با ۲۵۰ ژن برتر RF
3. **انتخاب ۱۰۶ ژن مشترک**
   تنها ژن‌هایی که حداقل توسط دو روش انتخاب شده‌اند
4. **استفاده از این ۱۰۶ ژن برای آموزش مدل‌ها**
   مجموعه نهایی ویژگی‌ها برای همه الگوریتم‌ها

نتیجه: مجموعه داده نهایی شامل ۸۰۱ نمونه × ۱۰۶ ژن × ۵ کلاس سرطان



شکل ۴ — (A) ۳۰ ژن برتر Ridge (B) ۳۰ ژن برتر Lasso (C) ژن‌های مشترک (منبع: مقاله)

07 الگوریتم‌های طبقه‌بندی

هشت مدل یادگیری ماشین بررسی‌شده

عملکرد هشت طبقه‌بندهای متفاوت از خانواده‌های یادگیری ماشین روی مجموعه ۱۰۶ ژن ارزیابی شد:

SVM
ماشین بردار پشتیبان

KNN
k-نزدیک‌ترین همسایه

ANN
شبکه عصبی مصنوعی

Random Forest
جنگل تصادفی

Decision Tree
درخت تصمیم

AdaBoost
بسته‌بندی تطبیقی

Naïve Bayes
بیز ساده

QDA
تحلیل تفکیک‌کننده درجه دوم

هر مدل با همان مجموعه ۱۰۶ ژن آموزش داده شد تا مقایسه منصفانه‌ای فراهم شود.

08 ارزیابی مدل‌ها

استراتژی ارزیابی و اعتبارسنجی

دو رویکرد اعتبارسنجی

Split Validation

تقسیم 70٪ آموزش / 30٪ آزمون. ارزیابی سریع و اولیه عملکرد مدل.

Fold Cross-Validation-5

داده به ۵ بخش مساوی تقسیم می‌شود؛ هر بار یک بخش به عنوان آزمون و چهار بخش به عنوان آموزش.
کاهش Overfitting و افزایش پایداری.

معیارهای ارزیابی

دقت (Accuracy)

نسبت پیش‌بینی‌های درست

Error Rate

نسبت پیش‌بینی‌های نادرست

Precision / Recall

دقت و فراخوانی به‌علاوه F1 Score

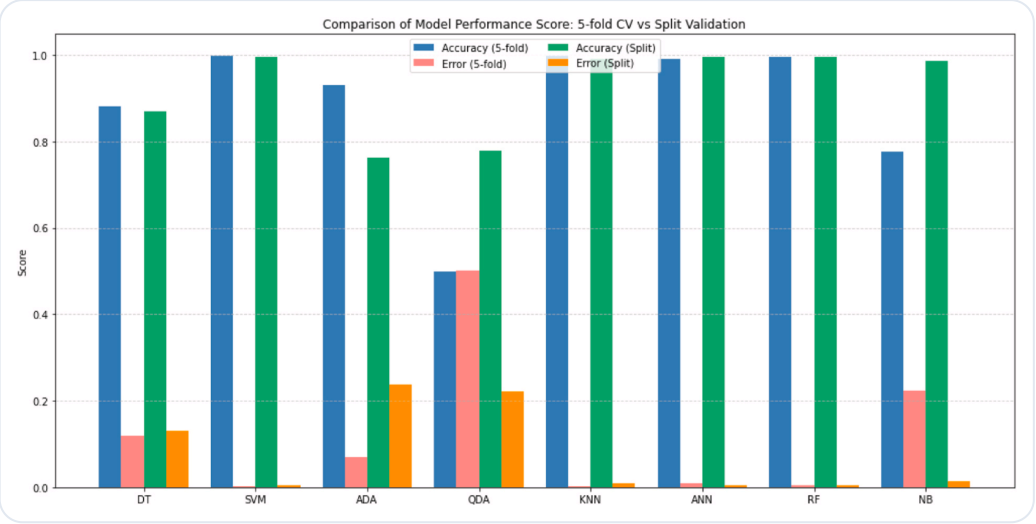
Confusion Matrix

ماتریس درهم‌ریختگی برای تحلیل هر کلاس

معیار اصلی انتخاب: بالاترین Accuracy در Fold CV-5

09 نتایج

SVM — پایدارترین و دقیق‌ترین مدل



شکل ۸ — مقایسه عملکرد مدل‌ها در Split و Fold Cross-Validation-5 (منبع: مقاله)

دقت در Fold Cross-Validation-5

SVM ۹۹.۸۷٪

بهترین عملکرد

KNN ۹۹.۷۵٪

نزدیک به SVM

Random Forest ۹۹.۵۱٪

مدل تولیفی قوی

نتیجه: SVM پایدارترین و دقیق‌ترین مدل بود — Error Rate تنها ۰.۱۳٪

10 جمع‌بندی

دستاوردهای پژوهش

۲۰۵۳۱ → ۱۰۶

کاهش ابعاد

کاهش موفق ابعاد از ۲۰,۵۳۱ ژن به ۱۰۶ ژن کلیدی با Ridge، Lasso و Random Forest.

۵ کلاس

طبقه‌بندی موفق

طبقه‌بندی موفق ۵ نوع سرطان (BRCA, KIRC, COAD, LUAD, PRAD) با دقت ۹۹.۸۷٪.

۱۰۶ ژن

بیومارکر بالقوه

شناسایی ۱۰۶ ژن مهم به‌عنوان بیومارکر بالقوه برای تشخیص مولکولی سرطان.

SVM

کاربرد بالینی

کاربرد SVM در بهبود تشخیص خودکار سرطان و پشتیبانی از تصمیم‌گیری درمانی.

11 محدودیت‌ها

محدودیت‌های پژوهش

با وجود نتایج امیدوارکننده، این مطالعه محدودیت‌هایی دارد که در مطالعات بعدی باید مورد توجه قرار گیرند:

۰۱

کمبود اطلاعات زیستی (Metadata)

مجموعه داده فاقد نام و توصیف ژن‌ها بود. این موضوع شناسایی و درک عملکرد زیستی ژن‌های انتخاب‌شده را دشوار می‌کند و تفسیر بالینی بیومارکرها را محدود می‌سازد.

۰۲

محدود بودن داده به یک منبع

تمام نمونه‌ها از پایگاه TCGA و پلتفرم Illumina HiSeq گردآوری شده‌اند. تنوع کم منابع ممکن است قابلیت تعمیم مدل به داده‌های سایر مراکز یا پلتفرم‌ها را محدود کند.

۰۳

نیاز به اعتبارسنجی روی داده‌های بیشتر

برای تأیید پایداری نتایج، نیاز است مدل‌ها روی مجموعه داده‌های مستقل و بزرگ‌تر اعتبارسنجی شوند. اعتبارسنجی اولیه روی CuMiDa انجام شد اما هنوز کافی نیست.

کار آینده: جمع‌آوری داده‌های چندمرکزی + استخراج Metadata + اعتبارسنجی بیولوژیک بیومارکرها