**Tên dự án:** AI-driven Cancer Risk Stratification from Multi-Omics Data

**Tác giả:** Huỳnh Trung Hiếu

1. **Dữ liệu và Yêu cầu nghiệp vụ | Data and Business Requirements**

**1.1 Bộ dữ liệu TCGA - The Cancer Genome Atlas | TCGA Dataset - The Cancer Genome Atlas**

**Nguồn:** <https://portal.gdc.cancer.gov> / **Source:** <https://portal.gdc.cancer.gov>  
**Định dạng dữ liệu:** RNA-Seq: HTSeq-FPKM (.tsv) / Clinical Metadata: .json hoặc .tsv / **Data Format:** RNA-Seq: HTSeq-FPKM (.tsv) / Clinical Metadata: .json or .tsv  
**Loại ung thư:** BRCA: Ung thư vú, dữ liệu bao gồm RNA-Seq từ hơn 1000 bệnh nhân kèm theo thông tin lâm sàng. / **Cancer Type:** BRCA: Breast cancer, includes RNA-Seq data from over 1000 patients with accompanying clinical information.  
**Dung lượng dữ liệu:** RNA-Seq BRCA: Khoảng 2–3 GB (được nén) / Clinical Metadata: Khoảng 50 MB / **Data Size:** RNA-Seq BRCA: Around 2-3 GB (compressed) / Clinical Metadata: Around 50 MB  
**Mục tiêu sử dụng bộ dữ liệu TCGA:** Xử lý dữ liệu RNA-Seq và clinical metadata để xây dựng mô hình học máy / **Purpose of TCGA Dataset:** Download and process RNA-Seq data and clinical metadata to build machine learning models.

**1.2 Cơ sở dữ liệu GENCODE / ENSEMBL (Annotation Gene) | GENCODE / ENSEMBL Database (Gene Annotation)**

**Nguồn:** <https://www.gencodegenes.org> / **Source:** <https://www.gencodegenes.org>  
**Định dạng dữ liệu:** GTF (Gene Transfer Format) / **Data Format:** GTF (Gene Transfer Format)  
**Mục đích sử dụng:** Ánh xạ gene, loại bỏ nhiễu trong dữ liệu / **Purpose:** Gene Mapping, noise removal in the data.

**Mục tiêu nghiệp vụ:**

* Xây dựng pipeline ETL để xử lý dữ liệu RNA-Seq và metadata lâm sàng.
* Chuẩn hóa và ánh xạ gene với dữ liệu mô tả lâm sàng (như tình trạng sống, đáp ứng thuốc).
* Phân tích mối liên hệ giữa biểu hiện gene và tiên lượng điều trị.
* Xây dựng mô hình ML/DL dự đoán kết quả điều trị theo đặc trưng phân tử.
* Triển khai dashboard giúp bác sĩ hỗ trợ ra quyết định.

**Business Goals:**

* Build an ETL pipeline to process RNA-Seq data and clinical metadata.
* Normalize and map genes with clinical descriptive data (such as survival status, drug response).
* Analyze the relationship between gene expression and treatment prognosis.
* Develop ML/DL models to predict treatment outcomes based on molecular features.
* Deploy a dashboard to assist doctors in decision-making.

1. **Hiểu về Dữ liệu | Understanding the Data**

**2.1 Dữ liệu RNA-Seq | RNA-Seq Data**

Các trường dữ liệu quan trọng trong bộ RNA-Seq / Key Data Fields in RNA-Seq Dataset:

* **Ensembl\_ID**
* **Tên tiếng Anh:** Ensembl\_ID
* **Mô tả (VN):** Mã định danh gene theo hệ Ensembl
* **Description (EN):** Gene identifier according to the Ensembl system
* **Gene\_Symbol**
* **Tên tiếng Anh:** Gene\_Symbol
* **Mô tả (VN):** Tên gene phổ biến
* **Description (EN):** Common gene name
* **FPKM\_Value**
* **Tên tiếng Anh:** FPKM\_Value
* **Mô tả (VN):** Mức độ biểu hiện gene (Fragments per kilobase of exon per million mapped reads)
* **Description (EN):** Gene expression level (FPKM)
* **gene\_id**
* **Tên tiếng Anh:** gene\_id
* **Mô tả (VN):** Mã định danh gene duy nhất (có thể trùng với Ensembl\_ID)
* **Description (EN):** Unique gene identifier (may match Ensembl\_ID)
* **gene\_name**
* **Tên tiếng Anh:** gene\_name
* **Mô tả (VN):** Tên của gene
* **Description (EN):** Name of the gene
* **gene\_type**
* **Tên tiếng Anh:** gene\_type
* **Mô tả (VN):** Loại gene (protein\_coding, lincRNA, pseudogene, v.v.)
* **Description (EN):** Type of gene (protein\_coding, lincRNA, pseudogene, etc.)
* **unstranded**
* **Tên tiếng Anh:** unstranded
* **Mô tả (VN):** Số lượng đọc (reads) không phân biệt chiều
* **Description (EN):** Read counts without strand specificity
* **stranded\_first**
* **Tên tiếng Anh:** stranded\_first
* **Mô tả (VN):** Số lượng đọc theo chiều đầu tiên
* **Description (EN):** Read counts for the first strand
* **stranded\_second**
* **Tên tiếng Anh:** stranded\_second
* **Mô tả (VN):** Số lượng đọc theo chiều thứ hai
* **Description (EN):** Read counts for the second strand
* **tpm\_unstranded**
* **Tên tiếng Anh:** tpm\_unstranded
* **Mô tả (VN):** TPM (Transcripts Per Million) không phân biệt chiều
* **Description (EN):** TPM (Transcripts Per Million) unstranded
* **fpkm\_unstranded**
* **Tên tiếng Anh:** fpkm\_unstranded
* **Mô tả (VN):** FPKM không phân biệt chiều
* **Description (EN):** FPKM unstranded
* **fpkm\_uq\_unstranded**
* **Tên tiếng Anh:** fpkm\_uq\_unstranded
* **Mô tả (VN):** FPKM được chuẩn hóa theo upper-quartile (không phân biệt chiều)
* **Description (EN):** Upper-quartile normalized FPKM (unstranded)
* **sample\_id**
* **Tên tiếng Anh:** sample\_id
* **Mô tả (VN):** Mã định danh của mẫu sinh học
* **Description (EN):** Identifier for the biological sample
* **case\_id**
* **Tên tiếng Anh:** case\_id
* **Mô tả (VN):** Mã định danh của trường hợp bệnh nhân hoặc thực thể nghiên cứu
* **Description (EN):** Identifier for the patient case or research entity

**2.2 Metadata Lâm Sàng | Clinical Metadata**

Các trường dữ liệu trong metadata lâm sàng / Key Fields in Clinical Metadata:

* **case\_id**
* **Tên tiếng Anh:** case\_id
* **Mô tả (VN):** Mã định danh của bệnh nhân
* **Description (EN):** Patient ID
* **patient\_id**
* **Tên tiếng Anh:** patient\_id
* **Mô tả (VN):** Mã định danh bệnh nhân trong hệ thống
* **Description (EN):** Unique patient identifier in the system
* **age\_at\_diagnosis**
* **Tên tiếng Anh:** age\_at\_diagnosis
* **Mô tả (VN):** Tuổi khi được chẩn đoán
* **Description (EN):** Age at diagnosis
* **gender**
* **Tên tiếng Anh:** gender
* **Mô tả (VN):** Giới tính của bệnh nhân
* **Description (EN):** Patient's gender
* **treatment\_outcome**
* **Tên tiếng Anh:** treatment\_outcome
* **Mô tả (VN):** Kết quả điều trị
* **Description (EN):** Treatment outcome (good/bad response or ongoing treatment)
* **survival\_time\_months**
* **Tên tiếng Anh:** survival\_time\_months
* **Mô tả (VN):** Số tháng sống sót
* **Description (EN):** Survival time in months post-diagnosis

**2.3 Cơ sở dữ liệu GENCODE / ENSEMBL (annotation gene)**

**Các trường dữ liệu trong annotation gene từ GENCODE / ENSEMBL / Key Fields in GENCODE / ENSEMBL Gene Annotation:**

**− gene\_id  
• Tên tiếng Anh:** gene\_id **• Mô tả (VN):** Mã định danh duy nhất của mỗi gene theo hệ GENCODE/ENSEMBL **• Description** (EN): Unique identifier for each gene in the GENCODE/ENSEMBL system

**− gene\_name  
• Tên tiếng Anh:** gene\_name **• Mô tả (VN):** Tên chính thức hoặc phổ biến của gene  
**• Description (EN):** Official or commonly used gene name

**− gene\_type  
• Tên tiếng Anh:** gene\_type **• Mô tả (VN):** Loại gene (ví dụ: protein-coding, lncRNA, miRNA, pseudogene,...) **• Description (EN):** Type of gene (e.g., protein-coding, lncRNA, miRNA, pseudogene, etc.)

**− transcript\_id  
• Tên tiếng Anh:** transcript\_id **• Mô tả (VN):** Mã định danh của các bản sao RNA (transcript) tương ứng với gene **• Description (EN):** Identifiers of RNA transcripts associated with the gene

**− exon\_coordinates  
• Tên tiếng Anh:** exon\_coordinates **• Mô tả (VN):** Tọa độ các exon trên bộ gen (genome) **• Description (EN):** Genomic coordinates of the exons

**− strand  
• Tên tiếng Anh:** strand  
**• Mô tả (VN):** Thông tin về chiều của gene trên DNA (+ hoặc −) **• Description (EN):** Strand orientation of the gene on the genome (+ or −)

1. **Thiết kế Cơ sở Dữ liệu OLTP (Chuẩn hóa 3NF) | OLTP Database Design (Normalized to 3NF)**

**OLTP DATABASE DESIGN (3rd NORMAL FORM)**

**Mục tiêu / Objective:**  
Lưu trữ dữ liệu biểu hiện gene, thông tin bệnh nhân và chú giải gene theo cách có cấu trúc, hỗ trợ truy vấn hiệu quả và phân tích chuyên sâu.  
*Store gene expression, patient information, and gene annotations in a well-structured format to support querying and analysis.*

**3.1 Các bảng chính | Main Tables**

**1. Table: genes**

| **Field** | **Description** | **Data Type** | **Constraints** |
| --- | --- | --- | --- |
| gene\_id | Unique gene identifier | VARCHAR | Primary Key |
| ensembl\_id | Ensembl gene ID | VARCHAR |  |
| gene\_symbol | Gene symbol | VARCHAR |  |
| gene\_description | Gene description | TEXT |  |
| gene\_type | Type of gene (e.g. protein\_coding) | VARCHAR |  |

**2. Table: patients**

| **Field** | **Description** | **Data Type** | **Constraints** |
| --- | --- | --- | --- |
| patient\_id | Patient unique ID | VARCHAR | Primary Key |
| case\_id | Case identifier | VARCHAR |  |
| age\_at\_diagnosis | Age at diagnosis | INTEGER |  |
| gender | Gender | VARCHAR |  |
| treatment\_outcome | Treatment outcome | VARCHAR |  |
| survival\_time\_months | Survival time in months | INTEGER |  |

**3. Table: samples**

| **Field** | **Description** | **Data Type** | **Constraints** |
| --- | --- | --- | --- |
| sample\_id | Sample unique identifier | VARCHAR | Primary Key |
| patient\_id | Foreign key to patients | VARCHAR | Foreign Key → patients(patient\_id) |
| case\_id | Case ID from metadata | VARCHAR |  |

**4. Table: gene\_expression**

| **Field** | **Description** | **Data Type** | **Constraints** |
| --- | --- | --- | --- |
| expression\_id | Unique ID for expression record | SERIAL | Primary Key |
| sample\_id | Foreign key to sample | VARCHAR | Foreign Key → samples(sample\_id) |
| gene\_id | Foreign key to gene | VARCHAR | Foreign Key → genes(gene\_id) |
| fpkm\_value | FPKM expression value | FLOAT |  |
| tpm\_unstranded | TPM unstranded | FLOAT |  |
| fpkm\_unstranded | FPKM unstranded | FLOAT |  |
| fpkm\_uq\_unstranded | Upper quartile FPKM | FLOAT |  |
| unstranded | Read count (unstranded) | INTEGER |  |
| stranded\_first | Read count first strand | INTEGER |  |
| stranded\_second | Read count second strand | INTEGER |  |

**5. Table: transcripts**

| **Field** | **Description** | **Data Type** | **Constraints** |
| --- | --- | --- | --- |
| transcript\_id | RNA transcript ID | VARCHAR | Primary Key |
| gene\_id | Linked gene ID | VARCHAR | Foreign Key → genes(gene\_id) |
| exon\_coordinates | Coordinates of exons | TEXT |  |
| strand | DNA strand direction | CHAR(1) |  |

**3.2 Mối quan hệ giữa các bảng | Table Relationships**

* patients.patient\_id ←→ samples.patient\_id
* samples.sample\_id ←→ gene\_expression.sample\_id
* genes.gene\_id ←→ gene\_expression.gene\_id
* genes.gene\_id ←→ transcripts.gene\_id

A diagram of a computer code

AI-generated content may be incorrect.

1. **Thiết kế Kho Dữ liệu (Data Warehouse) - Star Schema | Data Warehouse Design (Star Schema)**

**Mục tiêu | Objective:** Hỗ trợ phân tích đa chiều giữa biểu hiện gene, bệnh nhân và kết quả điều trị. / Support multidimensional analysis of gene expression, patient data, and treatment outcomes.

**Dimension Tables:**

**1.1. DimPatient**

* patient\_id: Mã định danh bệnh nhân.
* age\_group: Nhóm tuổi của bệnh nhân.
* gender: Giới tính của bệnh nhân.
* treatment\_outcome: Kết quả điều trị của bệnh nhân.

**1.2. DimGene**

* gene\_id: Mã định danh gene, liên kết đến bảng gene gốc.
* gene\_symbol: Tên viết tắt của gene.
* pathway: Đường dẫn sinh học mà gene tham gia.
* cancer\_related: Cho biết gene có liên quan đến ung thư hay không.

**1.3. DimDate**

* date\_id: Mã định danh ngày tháng.
* year: Năm.
* quarter: Quý trong năm.
* day: Ngày cụ thể.

**Fact Table: FactGeneExpression**

**FactGeneExpression**

* expression\_id: Mã định danh biểu hiện gene.
* patient\_id: Mã bệnh nhân, khóa ngoại liên kết đến bảng DimPatient.
* gene\_id: Mã gene, khóa ngoại liên kết đến bảng DimGene.
* date\_id: Mã ngày, khóa ngoại liên kết đến bảng DimDate.
* fpkm\_value: Giá trị mức độ biểu hiện của gene (FPKM).
* survival\_time\_months: Thời gian sống sót (tính bằng tháng) sau chẩn đoán.

A diagram of a data flow

AI-generated content may be incorrect.

1. **Thiết kế Hệ thống ELT (Extract - Load - Transform) | ELT System Design (Extract - Load - Transform)**

**Công cụ sử dụng | Tools Used:** Apache Airflow: Điều phối pipeline ETL / Pandas, scikit-learn, XGBoost: Tiền xử lý và xây dựng mô hình / PostgreSQL: Lưu trữ OLTP và kho dữ liệu / Docker: Container hóa hệ thống / Metabase hoặc Streamlit: Trực quan hóa dữ liệu.

Cấu trúc dự án dự kiến

/project-root

│

├── config

│ ├── snakemake\_config.py # Cấu hình Snakemake pipeline

│ ├── database\_config.py # Cấu hình kết nối DB

│ ├── logging\_config.py # Cấu hình logging

│ └── other\_configs.py # Các cấu hình khác (nếu có)

│

├── data

│ ├── raw # Dữ liệu gốc (FASTQ, VCF, BAM, ...)

│ └── processed # Dữ liệu đã xử lý (sau khi ETL)

│

├── database

│ ├── postgres\_connect.py # Kết nối PostgreSQL

│ ├── mongo\_connect.py # Kết nối MongoDB

│ └── migrations # Các migration file

│

├── scripts

│ ├── snakemake\_workflow.py # Main Snakemake workflow

│ ├── preprocess.py # Tiền xử lý dữ liệu

│ ├── model.py # Định nghĩa mô hình

│ └── utils.py # Các hàm tiện ích

│

├── notebooks

│ ├── eda.ipynb # Phân tích khám phá dữ liệu (EDA)

│ ├── model\_experiments.ipynb # Thử nghiệm với các mô hình

│ └── visualization.ipynb # Trực quan hóa dữ liệu

│

├── requirements.txt # Các thư viện Python cần thiết

├── Dockerfile # Cấu hình Docker

├── docker-compose.yml # Cấu hình Docker Compose

├── Makefile # Các lệnh Makefile

└── README.md # Tài liệu hướng dẫn sử dụng

RAW RNA-Seq/Metadata (TSV, JSON, GTF)

↓ (Dockerized Script: Preprocess)

Spark (PySpark or Spark SQL) — xử lý song song gene expression

↓

MySQL (OLTP): Clinical Metadata | PostgreSQL (OLAP): BI ready schema

↓

ETL orchestration (Airflow)

↓

dbt: Modeling/Transform data warehouse

↓

ML Model (Scikit-learn/XGBoost) → Model Registry (joblib/pickle)

↓

Power BI Dashboard (Doctors/Researchers)

**ETL Diagram**

A diagram of data orchestration

AI-generated content may be incorrect.

| **Thành phần** | **Vai trò** |
| --- | --- |
| **Data Source** | TCGA RNA-Seq, metadata lâm sàng, GENCODE annotation |
| **Extract** | Tách dữ liệu thô từ nguồn, validate format |
| **Data Lake** | Lưu trữ dữ liệu thô (.tsv, .json) để truy xuất nhiều lần |
| **Transform** | Làm sạch, chuẩn hóa, join dữ liệu, trích xuất đặc trưng |
| **Data Warehouse** | Lưu trữ dữ liệu đã chuẩn hóa phục vụ ML và BI |
| **Machine Learning** | Huấn luyện mô hình phân tầng rủi ro ung thư |
| **BI Analytics** | Dashboard báo cáo và phân tích kết quả |
| **Data Orchestration** | Tự động hóa toàn bộ pipeline theo lịch hoặc event |
| **Docker** | Đóng gói ứng dụng |

A diagram of a model

AI-generated content may be incorrect.

Chi tiết từng task:

* **1. Extract Data**
* **Task 1.1**: Download RNA-Seq (.tsv) từ TCGA BRCA
* **Task 1.2**: Download clinical metadata (.json/.tsv)
* **Task 1.3**: Download GENCODE annotation file (.gtf)
* **2. Transform Data**
* **Task 2.1**: Parse RNA-Seq files → dataframe với sample\_id, FPKM, gene\_id, gene\_symbol
* **Task 2.2**: Chuẩn hóa clinical metadata → age, gender, survival\_time\_months,...
* **Task 2.3**: Parse GTF file → gene\_id, gene\_name, gene\_type, strand, exon\_coordinates
* **Task 2.4**: Join RNA-Seq với annotation bằng gene\_id
* **Task 2.5**: Join clinical metadata với RNA-Seq qua case\_id hoặc sample\_id
* **Task 2.6**: Clean & xử lý giá trị thiếu
* **3. Load (ETL vào OLTP Database chuẩn 3NF)**
* **Task 3.1**: Tạo schema:
  + genes(gene\_id, ensembl\_id, gene\_symbol, gene\_type)
  + patients(case\_id, patient\_id, age, gender, survival\_time\_months)
  + expression(sample\_id, gene\_id, fpkm\_value)
* **Task 3.2**: Load dữ liệu chuẩn hóa vào PostgreSQL/MySQL
* **4. Analyze & Feature Engineering**
* **Task 4.1**: Tính các đặc trưng từ RNA-Seq (PCA, gene signature)
* **Task 4.2**: One-hot encode categorical clinical data
* **Task 4.3**: Merge gene expression với clinical label: treatment\_outcome
* **5. Train ML/DL Model**
* **Task 5.1**: Chia tập train/test (cross-validation)
* **Task 5.2**: Train các model:
  + Random Forest / SVM / XGBoost
  + Deep Learning (DNN, Autoencoder)
* **Task 5.3**: Lưu model (.pkl hoặc .h5)
* **6. Evaluate & Export**
* **Task 6.1**: Đánh giá AUC, ROC, F1-score, confusion matrix
* **Task 6.2**: Lưu performance report
* **Task 6.3**: Export mô hình phục vụ inference
* **7. Dashboard & Deployment**
* **Task 7.1**: Thiết kế dashboard trên Streamlit hoặc Metabase
* **Task 7.2**: Tích hợp model inference real-time
* **Task 7.3**: Triển khai cho bác sĩ sử dụng: lọc theo bệnh nhân, gene, tình trạng

**ETL Tool**

A diagram of a diagram of a software

AI-generated content may be incorrect.

[Raw Data - data/raw]

↓

[ETL Scripts (Airflow DAG + Python)] -- xử lý → [Processed Data - data/processed]

↓

[MySQL OLTP (mysql\_oltp/schema.sql)] ← đẩy dữ liệu sạch vào

↓

[dbt trên PostgreSQL (dwh\_postgres)] -- chuyển đổi từ OLTP sang DWH (Star Schema)

↓

[PostgreSQL Data Warehouse] ← Dữ liệu sẵn sàng phân tích & báo cáo

↓

[ML Pipeline - ml\_pipeline] -- huấn luyện, dự đoán

↓

[Power BI] -- Kết nối trực tiếp DWH, tạo dashboard hỗ trợ bác sĩ

Mọi thành phần chạy trên Docker container (docker/)

breast\_cancer\_project/

│

├── docker/ # Docker Compose & Dockerfile

│ ├── docker-compose.yml

│ ├── airflow/

│ │ ├── Dockerfile

│ │ └── requirements.txt

│ ├── spark/

│ │ └── Dockerfile

│ ├── mysql/

│ │ └── Dockerfile (nếu cần tùy chỉnh)

│ ├── postgres/

│ │ └── Dockerfile (nếu cần tùy chỉnh)

│ └── mlflow/

│ └── Dockerfile

│

├── data/ # Raw và processed data

│ ├── raw/

│ │ ├── rna\_seq.tsv

│ │ ├── clinical\_metadata.json

│ │ └── gencode\_annotation.gtf

│ └── processed/

│ ├── gene\_expression.csv

│ ├── clinical\_info.csv

│ └── gene\_annotation.csv

│

├── etl/

│ ├── dags/ # Airflow DAGs

│ │ ├── etl\_rna\_seq.py

│ │ ├── etl\_clinical\_metadata.py

│ │ └── etl\_gene\_annotation.py

│ ├── scripts/ # Python ETL scripts

│ │ ├── process\_rna\_seq.py

│ │ ├── process\_clinical.py

│ │ └── process\_annotation.py

│ └── logs/

│

├── mysql\_oltp/ # MySQL OLTP schema + init scripts

│ ├── schema.sql

│ ├── init\_data.sql

│ └── config/

│

├── dwh\_postgres/ # dbt project cho Data Warehouse PostgreSQL

│ ├── dbt\_project.yml

│ ├── models/

│ │ ├── staging/

│ │ ├── marts/

│ │ │ ├── fact\_expression.sql

│ │ │ └── dim\_patient.sql

│ │ └── schemas.yml

│ └── seeds/

│

├── notebooks/ # Khám phá, tạo đặc trưng, huấn luyện mô hình

│ ├── 1\_eda.ipynb

│ ├── 2\_feature\_engineering.ipynb

│ ├── 3\_train\_model.ipynb

│ └── 4\_eval\_kaplan\_meier.ipynb

│

├── ml\_pipeline/

│ ├── train.py # Huấn luyện mô hình, dùng MLflow

│ ├── predict.py # Dự đoán cho profile mới

│ ├── mlflow\_tracking.py

│ └── requirements.txt

│

├── tests/

│ ├── test\_etl.py

│ ├── test\_dbt.py

│ └── test\_model.py

│

├── .env # Biến môi trường DB, Airflow...

├── .gitignore

├── README.md

└── Makefile # Lệnh tiện ích setup, chạy