Table des matières

1. Aperçu général du cancer étudié (TCGA-UCEC)

1.1. Définition clinique

Le cancer de l'endomètre se développe dans les cellules qui tapissent l'intérieur de l'utérus (*endomètre*). C'est l'un des cancers gynécologiques les plus fréquents, avec une majorité de cas diagnostiqués chez des femmes âgées, et une survie à 5 ans élevée si le diagnostic est précoce (~83 %).

1.2. Données disponibles dans TCGA

Richesse et variété des données

- **Cohorte importante** : environ 560 patientes analysées (UCEC).
- **Types de données accessibles :**
  - Données cliniques : âge, stade, survie, histologie.
  - Données génomiques : mutations, CNV (*copy number variations*), expression génique, données d'expression microARN.
  - Données d'imagerie disponibles via *The Cancer Imaging Archive (TCIA)* couplées aux mêmes patientes TCGA, ce qui permet des analyses phénotype–génotype.

Implication pour notre projet bioinformatique : Plusieurs niveaux d'analyse *(transcriptome, mutations, CNV, survie, potentiellement radiomique) avec un jeu de données cohérent.*

1.3. Sous-types moléculaires — base d'analyses avancées

Le projet TCGA a montré que l'endométrial carcinoma n'est pas homogène : il se divise en **quatre sous-types moléculaires distincts** :

1. POLE ultramutated
2. Microsatellite instability (MSI) hypermutated
3. Copy number low
4. Copy number high

Ces sous-types diffèrent par leurs altérations génomiques, leurs profils mutationnels et leurs implications cliniques, ce qui constitue une base solide pour des analyses comparatives et de stratification.

1.4. Découvertes biologiques clefs du projet TCGA

Les analyses du TCGA ont mis en évidence :

- mutations fréquentes de TP53 dans les tumeurs "copy number high",
- importantes altérations du nombre de copies dans certains sous-types,
- mutations plus fréquentes de PTEN et KRAS dans d'autres sous-types plus stables sur le plan génomique.

Ces différences constituent un cadre pertinent pour analyser des signatures moléculaires, explorer des biomarqueurs et les relier à des variables cliniques.

1.5. Pourquoi ce dataset est adapté à notre projet bioinformatique

Les raisons principales sont :

- une cohorte volumineuse et bien caractérisée,
- des données multi-omics accessibles,
- des sous-types distincts permettant la stratification,
- une solide documentation scientifique et des publications comparatives.

Ce choix nous permet donc de réaliser une analyse complète, pertinente et faisable dans un cadre académique.

2. Dataset TCGA-UCEC adapté aux consignes

2.1. Téléchargement + Pré-traitement

Le dataset TCGA-UCEC :

- est disponible publiquement via GDC/Xena
- possède des données standardisées
- inclut plusieurs formats (RNA-seq, mutations, CNV, cliniques)

Ce qui signifie :

- **téléchargement faisable**
- **pré-processing concret mais gérable**
    - normalisation expression
    - sélection des features
    - étiquetage selon sous-types (POLE/MSI/CNL/CNH)
    - filtering des échantillons

Donc le dataset correspond à la 1ère étape.

2.2. Usage de méthodes de Machine Learning

Le dataset permet :

- **classification supervisée :** Prédire les sous-types moléculaires.
- **clustering non supervisé :** Identifier des groupes basés sur l'expression ou mutations.
- **réduction de dimension :** PCA / t-SNE / UMAP.
- **analyse de survie :** éventuellement avec Cox + ML.

Nous pouvons très clairement comparer :

- Random Forest vs SVM vs Logistic Regression pour classification
- K-means vs Spectral Clustering, etc.

Donc les exigences ML sont satisfaites.

2.3. Comparaison de méthodes → Optimal Pipeline

L'existence de sous-types distincts permet :

- définir une tâche ML précise. Exemple : prédire POLE/MSI/CNL/CNH
- comparer plusieurs modèles
- optimiser : normalisation, feature selection, cross-validation

Nous pouvons donc :

- expliquer pourquoi notre pipeline est optimal
- choisir un critère : précision, simplicité, rapidité, interprétabilité

Donc directement aligné avec les attendus du projet.

2.4. Revue de littérature

L'étude TCGA d'origine est :

- bien documentée
- publiée
- abondamment citée

Nous pouvons donc :

- comparer à un baseline scientifique
- justifier l'intérêt du cancer
- utiliser résultats connus comme référence

Donc la littérature est facile à intégrer.

2.5. Présentation orale de 10 minutes

Le dataset :

- n'est pas trop rare ou complexe
- possède des patterns connus et illustrables
- peut montrer visuels impactants

Pour les slides :

- présentation du cancer
- description des sous-types
- pipeline ML
- résultats
- interprétation

Donc présentation clairs et structurable.

2.6. Conclusion

Notre choix du dataset TCGA-UCEC est parfaitement adapté aux consignes du projet.
Il coche toutes les cases : accessible, assez large, multi-omics, sous-types exploitables pour ML, littérature abondante, complexité raisonnable, potentiel de pipeline optimal comparatif
C'est un choix **stratégiquement intelligent, scientifiquement défendable, pédagogiquement viable**
(Et même plus facile qu'un pancréas, un mésothéliome ou un mélanome pour un projet limité à 3 semaines).

3. Objectifs du projet

3.1. Introduction

Développer un pipeline d'analyse permettant de prédire les sous-types moléculaires du carcinome endométrial (POLE / MSI / Copy-Number-Low / Copy-Number-High) à partir des données d'expression génique RNA-seq, en comparant plusieurs méthodes de machine learning et en identifiant l'approche optimalement adaptée.

3.2. Objectifs

3.2.1. Objectif ML principal : classification supervisée

Tâche : prédire le sous-type moléculaire à partir de l'expression génomique.
Méthodes minimum à comparer :

- Logistic Regression
- Random Forest
- SVM
- éventuellement un modèle baseline naïf (KNN ou arbre de décision)

Critères d'évaluation :

- accuracy
- f1-score
- confusion matrix

Pourquoi c'est réaliste ?

- dataset étiqueté
- 4 classes
- RNA-seq bien adapté à ces méthodes
- comparaison simple et légitime

3.2.2. Objectif ML secondaire : réduction de dimension

Tâche : visualiser et explorer les clusters.
Méthodes :

- PCA
- t-SNE ou UMAP (facultatif)

Résultats visuels attendus :

- représentation 2D/3D des clusters
- voir si les sous-types se séparent naturellement

Pourquoi c'est utile ?

- préparation du modèle
- compréhension des données
- élément pédagogique pour présentation

3.2.3. Objectif d'optimisation du pipeline

Proposer un pipeline intégrant :

- étape pre-processing (normalisation, filtering)
- feature selection (par exemple variance threshold / ANOVA / chi-square)
- modèle ML sélectionné
- cross-validation

À optimiser selon un critère :

- performance, ou
- simplicité, ou
- temps de calcul

Ce choix est cohérent avec l'évaluation du projet.

3.2.4. Objectifs optionnels (Pas obligatoires, mais réalisables éventuellement)

- sélection de gènes importants via importance RF
- interprétabilité SHAP
- test avec données mutationnelles en entrée au lieu du transcriptome

3.3. Respect des consignes du projet :

Consigne	**Validé comment**
Exploration	PCA + analyse sous-types
Recherche bibliographique	TCGA marker paper
Méthodes ML	Logistic / SVM / RF

Consigne	**Validé comment**
Comparaison	scores + confusion matrix
Pipeline optimal	feature selection + choix du modèle
Implémentation Python	Sklearn
10 min oral	problématique claire + visuels
Temps 3 semaines	Faisable

4. Planning détaillé – 3 semaines

4.1. Semaine 1 — Exploration et premières expérimentations

4.1.1. Objectifs de la semaine 1 :

- Télécharger les données
- Nettoyer / prétraiter
- Explorer
- Tester des premiers modèles simples

4.1.2. Tâches de la semaine 1 :

1. **Téléchargement et organisation des données**
    - Accès au GDC ou UCSC Xena
    - Sélection RNA-seq + labels sous-types + données cliniques
2. **Prétraitement**
    - normalisation comptages (TPM / log2)
    - filtrage faible expression
    - matching échantillons / labels
3. **Exploration initiale**
    - statistiques descriptives
    - PCA pour observer le clustering
    - étude distribution des sous-types
4. **Lecture bibliographique**
    - article TCGA principal
    - 1 ou 2 publications secondaires
5. **Premiers modèles ML simples (baseline)**
    - logistic regression
    - KNN ou arbre de décision

4.1.3. Livrables internes en fin de semaine 1 :

- dataset prêt et propre
- premiers graphiques (PCA)
- performance baseline
- premières notes bibliographiques

4.2. Semaine 2 — Comparaison et pipeline optimal

4.2.1. Objectifs de la semaine 2 :

- Comparer plusieurs modèles ML
- Implémenter feature selection
- Évaluer et optimiser pipeline

4.2.2. Tâches de la semaine 2 :

1. **Implémentation de plusieurs modèles ML**
    - Random Forest
    - SVM
    - (au moins 3 modèles comparés)
2. **Feature selection**
    - variance threshold
    - ou ANOVA F-test
    - ou RF importance
3. **Évaluation**
    - train/test split ou cross-validation
    - matrices de confusion
    - accuracy / f1-score
4. **Comparaison des performances**
    - tableau comparatif
    - choix du modèle optimal
    - justification (performance / simplicité / temps)
5. **Début de la préparation des slides**

4.2.3. Livrables internes fin semaine 2 :

- tableau comparatif

- choix pipeline optimal
- justification claire

4.3. Semaine 3 — Finalisation et présentation

4.3.1. Objectifs de la semaine 3 :

- Finaliser pipeline et résultats
- Préparer slides
- Préparer oral & Q/A

4.3.2. Tâches de la semaine 3 :

1. **Nettoyage du code**
    - modularisation
    - reproductibilité
2. **Finalisation du pipeline optimal**
    - intégration pre processing + feature selection + modèle
    - visualisation finale PCA/UMAP + confusion matrices
3. **Préparation de la présentation**
    - structure des slides :
        - introduction problème
        - données
        - méthodes testées
        - pipeline optimal
        - résultats
        - comparaison littérature
        - limites & perspectives
4. **Simulation oral 10 min**
    - chronométrage
    - reformulation
    - anticipation des questions

4.3.3. Livrables :

- slides PDF
- modèle final prêt
- oraux répétés

4.4. Résumé des travaux

- Bloc A : data + preprocessing
- Bloc B : ML modèles + comparaison
- Bloc C : slides + bibliographie + interprétation

4.5. Résumé du planning

- Semaine 1 : techniques standard + installation environnement
- Semaine 2 : implémentation ML + comparaison → cœur du projet
- Semaine 3 : polissage + communication