**Title: Early Detection of Cancer through DNA Methylation Patterns**

**Names:**

* Yousef Abugosh : 315177444, yousef.abu-gosh@mail.huji.ac.il, yousefak
* Areen Mansour : 212900211, areen.mansour@mail.huji.ac.il, areen0507
* Nosiba Othman : 212378111, nosiba.otman@mail.huji.ac.il, nosiba
* Adan Mhameed : 214310666, adan.mhameed2@mail.huji.ac.il, adan.mhameed2

**Background:**

DNA methylation is an essential epigenetic modification that plays a crucial role in gene regulation.  
Aberrant methylation patterns have been linked to various cancers, making them potential **biomarkers for early cancer detection**.  
By leveraging **computational models**, it is possible to analyze DNA methylation data to identify early-stage cancer biomarkers.

**Reference Articles:**

📄 **Methylation Biomarkers for Early Cancer Detection and Diagnosis** – EJ CANCER  
📄 **Early Detection and Diagnosis of Cancer with Interpretable Machine Learning** – ACADEMIC

**Summary:**  
These studies discuss the use of **DNA methylation** as a biomarker for cancer diagnosis, focusing on **machine learning approaches** for identifying neoplastic methylation patterns in tissue samples.

**Research Question and Objectives:**

**Research Question:**

How can **machine learning models** leverage **DNA methylation patterns** to improve early cancer detection?

**Objectives:**

* Develop a **machine learning model** to classify tissue samples based on DNA methylation profiles.
* Utilize **publicly available datasets** to train and validate predictive models.
* Implement **interpretable AI techniques** to ensure model transparency and reliability in biomarker identification.

**Data Description:**

**The dataset includes:**

* **DNA methylation profiles** from multiple cancer types and normal tissue samples.
* Data sources include **TCGA, MethHC 2.0, and MethMarkerDB**.
* Data format: **Raw sequencing data (FASTQ/BAM) and processed methylation matrices (CSV).**

**Data Sources:**

📂 **The Cancer Genome Atlas (TCGA)** 🔗 https://portal.gdc.cancer.gov/  
📂 **MethHC 2.0** – Integrated database for human cancer methylomes.  
📂 **MethMarkerDB** – Repository of known cancer DNA methylation biomarkers.

**Hypothesis:**

DNA methylation patterns contain **distinctive signals** that can be leveraged by machine learning models to **distinguish cancerous from non-cancerous tissues** with **high accuracy**.

**Computational and Statistical Methods:**

* **Model:** Deep neural networks and interpretable AI methods for methylation-based cancer classification.
* **Feature Selection:** Identification of key methylation sites correlated with cancer progression.
* **Machine Learning:** Supervised learning algorithms including **Random Forest and Deep Learning models**.
* **Validation Metrics:** Accuracy, Sensitivity, Specificity, and **AUC-ROC** analysis.
* **Software Tools:** **Python, Scikit-learn, TensorFlow/PyTorch, methCancer-gen (GitHub), and UCSC Xena** for visualization.

**Execution Plan:**

1️⃣ **Data Acquisition:** Collect and preprocess **DNA methylation datasets** from **TCGA and other sources**.  
2️⃣ **Feature Engineering:** Extract key methylation features using statistical and computational methods.  
3️⃣ **Model Training:** Train machine learning models to classify **cancerous vs. non-cancerous** tissues.  
4️⃣ **Model Evaluation:** Assess performance using **sensitivity, specificity, and AUC-ROC metrics**.  
5️⃣ **Visualization:** Use **UCSC Xena** for interactive exploration of results.

**References:**

📄 **EJ Cancer - Methylation Biomarkers for Early Cancer Detection and Diagnosis**  
📄 **Academic - Early Detection and Diagnosis of Cancer with Interpretable Machine Learning**  
📂 **GitHub: methCancer-gen** 🔗 <https://github.com/cbi-bioinfo/methCancer-gen>

**Diversity in the Team:**

Our team consists of researchers from diverse backgrounds, including:

* **Computational Biology**
* **Machine Learning**
* **Computer Science**
* **Bioinformatics**

This **multidisciplinary approach** enhances our ability to analyze **complex biological data** and improve **cancer detection models**.