Data Analytics for Life Science: Final Capstone Project Requirements

Team Size: 2 Students per Team

1. Project Tracks (Choose One)

Your team must select one of the following three tracks. These tracks correspond to the data types and architectures mastered in the course.

Track A: Medical Imaging (Classification, Segmentation, or Counting)

- **Focus:** 2D Medical Images (X-Ray, MRI, Histopathology, Microscopy).
- **Architectures:** CNN (for Classification) or U-Net (for Segmentation & Counting).
- **Sample Problems:**
    - **Segmentation & Counting:** Segmenting cell nuclei in microscopy images and counting them (e.g., Data Science Bowl).
    - **Counting:** Automated counting of bacterial colonies on Petri dishes.
    - **Segmentation:** Identifying tumor regions in brain MRI slices.
    - **Classification/Detection:** Malaria parasite detection in thin blood smear images.
    - **Classification:** Detecting pneumonia or COVID-19 from Chest X-Rays (using CNNs).
    - **Classification:** Classifying tissue subtypes in histopathology patches.

Track B: Biological Sequence Analysis

- **Focus:** DNA, RNA, or Protein Sequences.
- **Architectures:** 1D-CNN, Bi-LSTM, or Hybrid CNN-LSTM.
- **Sample Problems:**
    - **Protein Property Prediction:** Predicting if a protein sequence is soluble, stable, or an enzyme.
    - **Motif Discovery:** Identifying Transcription Factor Binding Sites (TFBS) in DNA.
    - **Secondary Structure:** Improving our lab model to predict 8-state (Q8) structure with higher accuracy.

Track C: Computational Drug Discovery (Multi-Stage Pipeline)

- **Focus:** Small Molecules (Chemicals/Drugs).
- **Requirement:** You must implement a **Multi-Task Pipeline** that optimizes for both efficacy and safety. Your project must include **at least two** distinct prediction tasks:
    1. **Bioactivity (Primary):** Predict if a drug binds to a specific target or inhibits a pathogen (e.g., Binding Affinity, IC50, or Active/Inactive).
    2. **ADMET (Secondary):** Predict properties like **Toxicity**, **Solubility**, or **BBBP** (Blood-Brain Barrier Penetration) to filter out bad candidates.
- **Architectures:** SVM/Random Forest/Dense Networks (for Fingerprints) or CNN-LSTM (for SMILES).
- **Sample Workflow:**
    - "We trained Model A to predict binding to the EGFR receptor (Bioactivity). Then, we trained Model B to predict Toxicity. We filtered a database to find molecules that are **Predicted Active AND Predicted Non-Toxic**."

3. Technical Requirements (Mandatory)

Every project must include the following five components:

- **Data Curation:**
    - You must find a public dataset (e.g., Kaggle, MoleculeNet, Broad Bioimage Benchmark).
    - **Preprocessing:** You must demonstrate data cleaning. For images: augmentation/resizing. For sequences: tokenization/padding/k-mers. For drugs: calculation of descriptors/fingerprints.
- **The "Baseline" Model and Reference Model:**
    - Before running your Deep Learning model, you must train a simple baseline (e.g., Logistic Regression, Random Forest, or a basic 1-layer CNN) to establish a "floor" for performance.
    - You also need to use transfer learning to establish Reference Model to compare performance.
- **The Advanced Architecture:**

- You must implement a sophisticated model relevant to your track (e.g., a **U-Net** with skip connections, a **Hybrid CNN-LSTM**, or a **ResNet**-style CNN).
- Your model must differently compare to your model in weekly assignments.

- **Rigorous Evaluation:**
  - **Split:** Train/Validation/Test split (preventing data leakage).
  - **Metrics:** You must use domain-appropriate metrics:
    - **Classification:** AUPRC, F1-Score (for imbalance).
    - **Segmentation:** IoU (Intersection over Union) or Dice Score.
    - **Counting:** MAE (Mean Absolute Error) or RMSE.
- **Explainable AI (XAI):**
  - **Mandatory:** You must visualize *why* the model made a decision.
  - **For Images:** Generate **Grad-CAM** or Saliency Maps to show which part of the X-ray/Cell the model looked at.
  - **For Sequences:** Use Saliency Maps to highlight key motifs/amino acids.
  - **For Drug Discovery:** XAI is **mandatory only for the Bioactivity model**. You must show which chemical substructures (pharmacophores) contributed to the binding prediction. XAI for ADMET (Toxicity/Solubility) is optional.
- Other requirements:
  - Use early stopping
  - Tuning hyperparameter to have good performance model.

4. Deliverables

Component	Description
1. Code Repository	A GitHub link. The repo must include a README.md explaining how to run the code and a requirements.txt file.
2. Final Notebook	A clean Jupyter Notebook (.ipynb) that runs from start to finish. It should tell a story: Data Loading -> EDA -> Model -> XAI.
3. Project Report	A 7-10-page PDF summary.

4. Presentation	A 15-minute slide presentation.
5. Streamlit App	A web app (app.py) for demonstration. Users should be able to upload a sample (Image/Sequence/SMILES) and see the Prediction + XAI visualization in real-time.

5. Suggested Datasets

- **Imaging (U-Net/CNN):**
  - *Data Science Bowl 2018 (Kaggle):* Nuclei segmentation and counting.
  - *Malaria Cell Images (Kaggle/NIH):* Parasite detection.
  - *AGAR (Annotated Germs for Automated Recognition):* Bacteria colony counting.
  - *Chest X-Ray Images (Pneumonia):* Classification.
  - *ISIC Archive:* Skin lesion classification.
- **Sequences (CNN-LSTM):**
  - *PDB Secondary Structure:* (The dataset used in class, extended).
  - *UniProt:* Protein function classification (e.g., Enzyme vs Non-Enzyme).
  - *AMPScanner:* Antimicrobial Peptide classification.
- **Drug Discovery (Must combine Bioactivity + ADMET):**
  - **Bioactivity Sources:** *BindingDB*, *ChEMBL*, or *ExCAPE-DB* (for Target-Specific Binding).
  - **ADMET Sources:** *MoleculeNet (DeepChem)*:
    - *Tox21 / ClinTox* (Toxicity).
    - *ESOL* (Solubility).
    - *BBBP* (Brain Penetration).