📌 Guidelines for Using the Transcriptomic Subtype Prediction Model

1️⃣ Purpose

This tool predicts cancer subtypes from **RNA-Seq gene-expression data** using a pre-trained model.
It automatically recognizes gene identifiers, checks data compatibility, and shows subtype probabilities interactively.

2️⃣ What You Need

- A **gene-expression file** in .csv or .tsv format.
- Each **row** represents one sample (patient or tissue).
- Each **column** represents one gene.
- The **first column** must be the sample name or ID.

Example:

sample_id,TP53,BRCA1,EGFR,MYC,...

Sample_01,6.23,5.11,7.02,8.43,...

Sample_02,5.91,4.87,6.88,8.12,...

3️⃣ Accepted Formats

- File types: .csv, .tsv, .txt, .gz (compressed)
- Gene identifiers: **Ensembl IDs (ENSG...)** or **HGNC symbols (TP53, BRCA1, etc.)**
- Expression units: Prefer **log2(TPM + 1)** or **FPKM-UQ**

If unsure, the app will detect the format automatically and show a note.

4️⃣ How to Upload and Run

1. Click **“Upload Data File.”**
2. Wait for automatic **gene-ID mapping** to complete.

3. Review the **overlap percentage** between your file and the model's gene set.
    - ≥ 50 % overlap → good quality
    - 20 – 49 % → moderate (use with caution)
    - < 20 % → poor; results may be unreliable
4. Click **"Run Prediction."**
5. View **predicted subtype probabilities** for each sample on the results screen.

5️⃣ Understanding the Output

- The model displays **probability bars** for each possible subtype.
- The **highest probability** indicates the most likely subtype.
- Hover over a sample name to view details.
- You can **download results** as a .csv file or export a PDF summary.

6️⃣ Warnings and Quality Messages

- **Low Gene Overlap:** "Only 12.8 % of genes overlap. Results may be unreliable."
- **Unrecognized IDs:** "Some gene identifiers could not be mapped."
- **Scale Mismatch:** "Your data appears to be raw counts; normalization is recommended."

These warnings help you interpret reliability before using results in analysis.

7️⃣ Tips for Best Results

✅ Use full transcriptome RNA-Seq data (≥ 20 000 genes).
✅ Keep consistent gene identifiers across samples.
✅ Normalize expression values to log2(TPM + 1) when possible.
✅ Avoid datasets with very few genes (< 10 000).
✅ Upload one dataset per run for clearer interpretation.

8️⃣ Example Compatible Sources

- **TCGA / GDC Pan-Cancer Atlas** RNA-Seq matrices
- **UCSC Xena** harmonized TCGA data
- **GTEx** normal tissue expression data

These share the same preprocessing style as the model's training data and give the most reliable predictions.

🆘 Support

If you encounter any issues:

- Check that your file follows the structure above.
- Verify gene IDs using Ensembl or HGNC lookup tools.
- Contact the technical support team with the overlap percentage shown in your report.