Module_3:

Team Members:

Vicente Carvajal Villegas, Nolan Nguyen

Project Title:

Exploring the Relationship Between Epithelial–Mesenchymal Transition States and Tumor Progression Across Cancer Types Using TCGA Expression and Clinical Data

Project Goal:

This project seeks to... *(what is the purpose of your project -- i.e., describe the question that you seek to answer by analyzing data.)*

How do different EMT states (epithelial, mesenchymal, hybrid) correlate with tumor stage and patient survival across various cancer types?

Disease Background:

Pick a hallmark to focus on, and figure out what genes you are interested in researching based on that decision. Then fill out the information below.

- Cancer hallmark focus: Activating Invasion and Metastasis
- Overview of hallmark: Enables tumor cells to spread from the primary site to distant organs, causing the majority of cancer-related deaths. This process involves cancer cells acquiring motility and invasive capabilities to break through surrounding tissues, enter the bloodstream or lymphatics, survive in circulation, and colonize new sites. Epithelial–mesenchymal transition (EMT) plays a key role by enabling tumor cells to lose adhesion and become migratory. The metastatic cascade is complex, involving interactions with the microenvironment and dynamic cellular plasticity, making invasion and metastasis the defining features of malignancy and the main target for therapeutic intervention. https://www.cell.com/fulltext/S0092-8674(11)00127-9
- Genes associated with hallmark to be studied (describe the role of each gene, signaling pathway, or gene set you are going to investigate):

Epithelial Markers (downregulated in EMT)

- CDH1 (E-cadherin): Major epithelial adhesion molecule; loss indicates EMT initiation.

- EPCAM: Cell adhesion molecule on epithelial cells.
- CLDN7 (Claudin 7): Tight junction protein maintaining epithelial barrier.
- OCLN (Occludin): Tight junction component key to epithelial cell polarity.

Mesenchymal Markers (upregulated in EMT)

- VIM (Vimentin): Intermediate filament protein indicating mesenchymal phenotype.
- FN1 (Fibronectin): Extracellular matrix protein promoting migration.
- SNAI1 (Snail): Transcription factor driving repression of epithelial genes and activation of EMT.
- ZEB1: EMT transcription factor promoting mesenchymal gene expression and invasion.

https://pmc.ncbi.nlm.nih.gov/articles/PMC7046610/

Will you be focusing on a single cancer type or looking across cancer types? Depending on your decision, update this section to include relevant information about the disease at the appropriate level of detail. Regardless, each bullet point should be filled in. If you are looking at multiple cancer types, you should investigate differences between the types (e.g. what is the most prevalent cancer type? What type has the highest mortality rate?) and similarities (e.g. what sorts of treatments exist across the board for cancer patients? what is common to all cancers in terms of biological mechanisms?). Note that this is a smaller list than the initial 11 in Module 1.

We are looking at how EMT scores relate to cancer stage and survival rate in all cancers in the data set. We have started with lung cancer but will scale the analysis to all given cancers. This is the reason for our general information given for the below bullets.

Prevalence & incidence

Cancer remains one of the most significant health burdens in the United States and worldwide. In 2025, it is estimated that 2,041,910 new cases of cancer will be diagnosed and 618,120 people will die from the disease in the U.S. alone, making cancer a leading cause of morbidity and mortality. Globally, cancer incidence continues to rise, with more than 20 million new cancer cases and over 10.3 million cancer deaths projected for 2025. The most common cancers in the U.S. are breast, prostate, and lung cancers.

Approximately 38.9% of men and women in the U.S. will be diagnosed with cancer during their lifetime. In terms of risk to children and adolescents (age 0–19), around 14,690 new cases and 1,650 deaths are expected in 2025. The rate of new cases (incidence) stands at 445.8 per 100,000 individuals per year, while the cancer death rate is 145.4 per 100,000 per year. Notably, mortality is highest among non-Hispanic Black men (203.6 per 100,000) and lowest among non-Hispanic Asian/Pacific Islander women (83.1 per 100,000).

As of January 2022, there were an estimated 18.1 million cancer survivors in the U.S., a number expected to rise to 26 million by 2040 as mortality rates decline and early detection and treatment improve.

https://www.cancer.gov/about-cancer/understanding/statistics

Risk factors (genetic, lifestyle) & Societal determinants

Genetic risk factors: Family history, inherited mutations like BRCA1/BRCA2, and genetic predispositions contribute to cancer risk but account for only about 5–10% of cases.

Lifestyle factors: Tobacco use is the leading preventable cause, followed by alcohol consumption, poor diet, obesity, physical inactivity, and excessive sun exposure. These modifiable behaviors significantly influence cancer incidence.

Environmental exposures: Carcinogens such as asbestos, radon, air pollution, and certain chemicals increase risk. Infections with viruses (HPV, Hepatitis B and C, Epstein-Barr) and bacteria (H. pylori) are significant contributors, especially in low- and middle-income countries.

Societal determinants: Socioeconomic status, education level, access to healthcare, and geographic location affect cancer risk through influencing exposure to risk factors, screening uptake, and timely treatment. Disparities persist, amplifying risk in underserved populations.

Most cancers occur in older adults due to cumulative exposure to risk factors and decreased cellular repair with age.

Approximately 40% of cancers could be prevented by lifestyle changes and reduction of environmental exposures.

https://www.who.int/news-room/fact-sheets/detail/cancer

Standard of care treatments (& reimbursement)

Standard cancer treatments in 2025 consist of surgery, radiation, chemotherapy, targeted therapies, and immunotherapy, often used in combination depending on cancer type and stage.

Precision medicine is increasingly central, leveraging tumor genomic profiling to guide targeted drug therapies such as mutant-specific inhibitors (KRAS, BRAF), antibody-drug conjugates, and personalized cellular therapies like CAR T-cells and TCR-engineered T cells, with expanding application in both hematologic and solid tumors.

Policy reforms like the 2025 ROCR Act in the U.S. promote patient-centered, episode-based reimbursement models to incentivize evidence-based, high-quality care while addressing patient access challenges and reducing overtreatment.

Despite therapeutic advances, disparities in access and affordability persist due to socioeconomic and geographic factors, necessitating integrated multidisciplinary care and patient navigation to improve coordination and equity.

Innovations including AI for earlier detection of treatment response, mRNA-based cancer vaccines, and bispecific antibodies are reshaping the treatment landscape, with ongoing clinical trials expanding treatment indications and modalities.

Globally, comprehensive cancer care increasingly balances efficacy, toxicity, quality of life, and health system sustainability guided by dynamic clinical guidelines from organizations like NCCN and ESMO.

https://ascopubs.org/doi/abs/10.1200/JCO.2025.43.16_suppl.e23219
https://www.nccn.org/guidelines/category_1

Biological mechanisms (anatomy, organ physiology, cell & molecular physiology)

Cancer arises from the uncontrolled proliferation of abnormal cells originating from nearly any tissue or organ in the body. Anatomically, tumors disrupt normal organ architecture and function causing progressive organ dysfunction. At the organ physiology level, cancer impairs homeostasis by hijacking cellular signaling pathways that regulate growth, differentiation, and apoptosis. Molecularly, cancer is characterized by genetic and epigenetic alterations activating oncogenes (e.g., KRAS, MYC), inactivating tumor suppressor genes (e.g., TP53, RB1), and promoting genomic instability.

Hallmark capabilities acquired include sustained proliferative signaling, evasion of growth suppressors, resistance to cell death, enabling replicative immortality, angiogenesis, invasion, and metastasis. Cellular plasticity phenomena like epithelial–mesenchymal transition (EMT) facilitate metastasis and therapy resistance. The tumor microenvironment, composed of cancer-associated fibroblasts, immune cells, and extracellular matrix components, interacts dynamically with cancer cells, influencing tumor progression and response to therapy. Metabolic reprogramming, immune evasion, and inflammation are additional core biological features that shape cancer behavior across organ systems. Overall, these complex anatomical, physiological, and molecular perturbations synergize to drive tumor development, progression, and clinical heterogeneity.

https://www.cancer.gov/about-cancer/understanding/what-is-cancer

Data-Set:

The data come from The Cancer Genome Atlas (TCGA), downloaded in a processed form as GSE62944 from the Gene Expression Omnibus (GEO).

Expression file: GSE62944_subsample_topVar_log2TPM.csv

Metadata file: GSE62944_metadata.csv These are curated by the course repository to include the most variable genes and matching clinical annotations for TCGA tumor samples.

How the data was collected

The TCGA program generated RNA-sequencing (RNA-seq) data from patient tumor and normal tissue samples across many cancer types.

RNA was extracted from tumor biopsies, converted to cDNA, sequenced using Illumina next-generation sequencing, and quantified as transcripts per million (TPM).

Clinical and pathological information (tumor type, stage, metastatic status, etc.) was recorded by participating hospitals and harmonized by TCGA.

The version used here is already $\log_2$-transformed TPM (log2(TPM + 1)), which makes gene-expression values approximately normal and comparable across samples.

What subset we used

For this project we selected all cancer types:

Features included

Gene-expression variables: thousands of genes (rows) measured per tumor sample (columns).

We focused on a biologically defined subset of genes related to epithelial (e.g., CDH1, EPCAM, KRT8) and mesenchymal (e.g., VIM, ZEB1, FN1, MMP9) markers that reflect the EMT process.

Clinical variables (metadata):

cancer_type – identifies the TCGA cancer cohort.

pathologic_stage / clinical_stage – disease stage at diagnosis.

pathologic_M, pathologic_N, metastatic_at_diagnosis – when available, indicate metastasis or lymph-node involvement.

Additional fields such as patient age, gender, and sample type are present but not used directly in this analysis.

Units and scale

Gene expression: $\log_2(\text{TPM} + 1)$, unitless continuous values.

Clinical fields: categorical or binary (e.g., Stage I–IV, M0/M1).

Why this dataset fits the project

This dataset allows exploration of the "Activating invasion and metastasis" hallmark by quantifying the EMT program at the transcriptomic level and relating it to tumor stage or metastatic status within a well-characterized cancer cohort.

Data Analyis:

Methods

The machine learning technique I am using is: *fill in and describe*

*What is this method optimizing? How does the model decide it is "good enough"?*The machine learning technique I am using is unsupervised and supervised learning using PCA and logistic regression. For PCA, there's no "training/validation" — the method is purely descriptive, but we check if the first components correspond to biological trends (like EMT score gradients).

For logistic regression, we evaluate performance using metrics such as:

AUC (Area Under ROC Curve) — measures how well the model separates classes.

Confusion matrix and classification report — summarize accuracy, precision, and recall.

If AUC is high (≥ 0.7–0.8), the model captures metastasis-related variation; if low, EMT alone doesn't fully explain the phenotype. **

Analysis

(Describe how you analyzed the data. This is where you should intersperse your Python code so that anyone reading this can run your code to perform the analysis that you did, generate your figures, etc.)

Imports and Setup

Imports core data-analysis libraries.

pandas, numpy handle data structures and math.

matplotlib is used for visualization.

sklearn (scikit-learn) provides all the machine learning tools:

StandardScaler → z-score normalization,

PCA → principal component analysis (unsupervised),

LogisticRegression → supervised classification,

roc_auc_score, etc. → evaluate model performance.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_auc_score, confusion_matrix,
classification_report
from sklearn.linear_model import LogisticRegression
```

This gives you the raw materials to connect gene activity with metastasis potential.

```
expr = pd.read_csv('GSE62944_subsample_topVar_log2TPM.csv',
index_col=0)
meta = pd.read_csv('GSE62944_metadata.csv', index_col=0)
```

Pick a cancer and subset (in this case we chose all cancer types)

```
samples = meta.index
X = expr[samples]    # genes x samples
M = meta.loc[samples]
```

Define EMT gene panels

```
epithelial = ['CDH1','EPCAM','CLDN1','KRT8','KRT18']
mesenchymal =
```

```
['VIM','SNAI1','SNAI2','TWIST1','ZEB1','ZEB2','MMP2','MMP9','ITGA5','I
TGB1','CXCR4','COL1A1','FN1']
```

Keep only genes that exist in the dataset

```
epi_present = [g for g in epithelial if g in X.index]
mes_present = [g for g in mesenchymal if g in X.index]
```

Compute EMT score

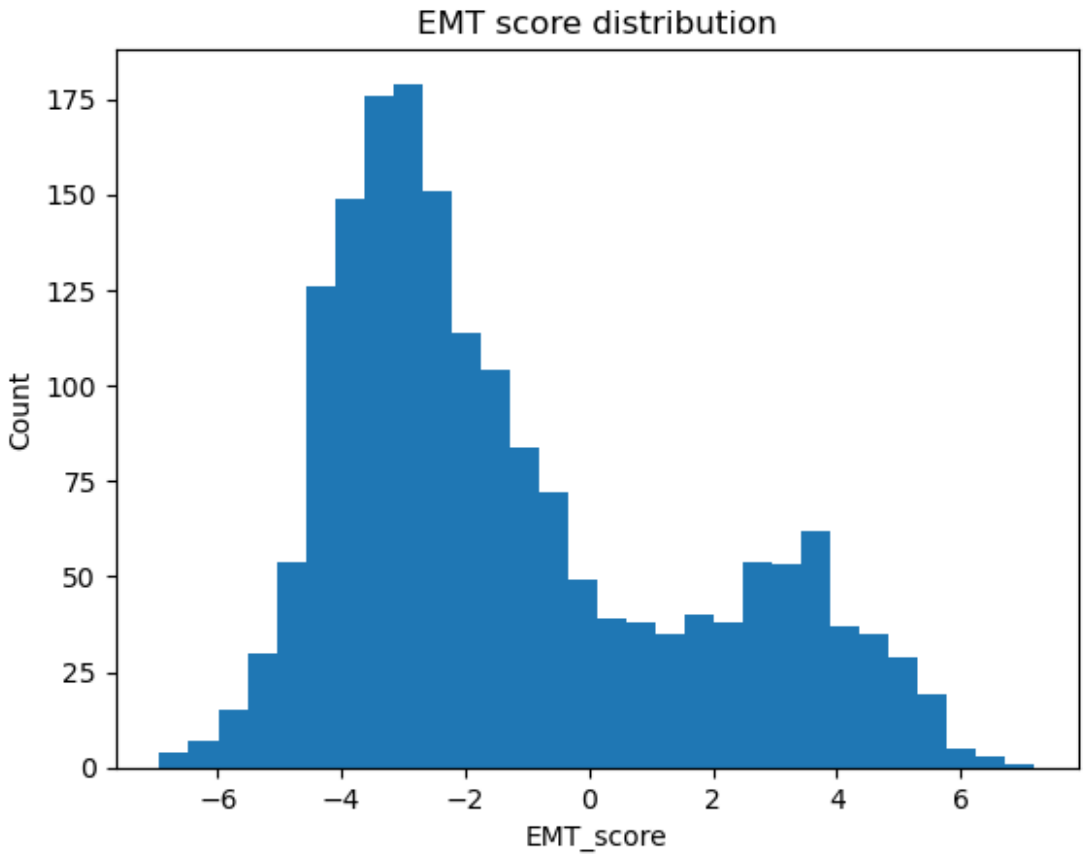
```
mes_score = X.loc[mes_present].mean(axis=0)
epi_score = X.loc[epi_present].mean(axis=0)
emt_score = mes_score - epi_score
emt_df = pd.DataFrame({'EMT_score': emt_score})
merged = emt_df.join(M, how='left')
```

Histogram of EMT score

A wide spread = strong biological variability (some epithelial, some mesenchymal).

This already supports the idea that not all tumors are equally invasive.

```
plt.figure()
merged['EMT_score'].plot.hist(bins=30)
plt.title(f'EMT score distribution')
plt.xlabel('EMT_score'); plt.ylabel('Count')
plt.show()
```

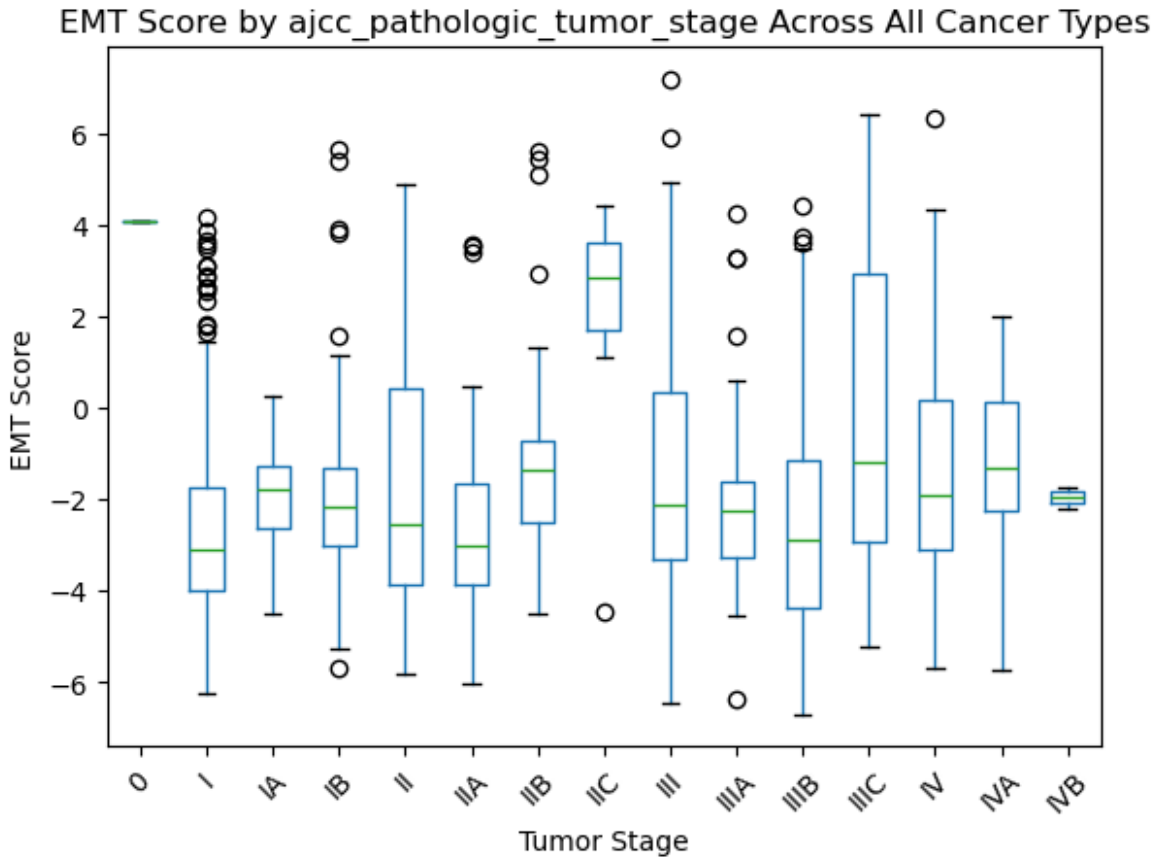


Compare EMT by stage

Finds any stage-like column (pathologic_stage, clinical_stage, etc.), cleans it up, and compares EMT scores by stage.

```
col = 'ajcc_pathologic_tumor_stage'

# This will produce a Boolean mask, making NAs False
valid_rows = merged[col].fillna("").str.startswith("Stage ")

tmp = merged.loc[valid_rows, ['EMT_score', col]].dropna()
tmp[col] = (
    tmp[col].str.replace("Stage ", "", regex=False)
    .str.upper()
    .str.strip()
)

print(f"Using grouping column: {col}")
print("Top levels:\n", tmp[col].value_counts().head())

import matplotlib.pyplot as plt
plt.figure(figsize=(10,6))
tmp.boxplot(column='EMT_score', by=col, grid=False, rot=45)
plt.title(f'EMT Score by {col} Across All Cancer Types')
```

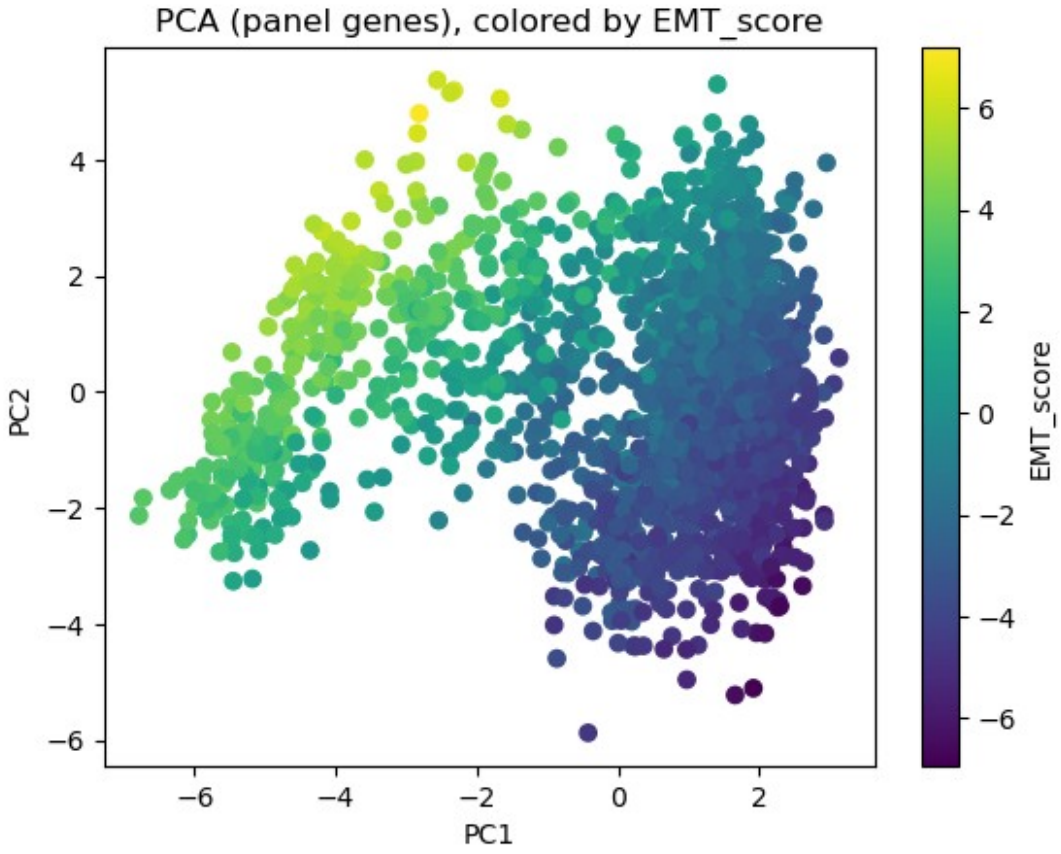
```
plt.xlabel('Tumor Stage')
plt.ylabel('EMT Score')
plt.suptitle('')
plt.show()

Using grouping column: ajcc_pathologic_tumor_stage
Top levels:
 ajcc_pathologic_tumor_stage
I      257
II     159
III    132
IIA     94
IV      91
Name: count, dtype: int64

<Figure size 1000x600 with 0 Axes>
```



PCA visualization

```
panel_genes = list(set(epi_present + mes_present))
scaler = StandardScaler(with_mean=True, with_std=True)
Z = scaler.fit_transform(X.loc[panel_genes].T)  # samples x genes
pca = PCA(n_components=2).fit(Z)
```

```
PC = pca.transform(Z)
pc_df = pd.DataFrame(PC, index=samples,
columns=['PC1','PC2']).join(emt_df)

plt.figure()
plt.scatter(pc_df['PC1'], pc_df['PC2'], c=pc_df['EMT_score'],
cmap='viridis')
plt.title(f'PCA (panel genes), colored by EMT_score')
plt.xlabel('PC1'); plt.ylabel('PC2')
plt.colorbar(label='EMT_score')
plt.show()
```



Supervised model (logistic regression)

```
candidate_labels = ["metastatic_at_diagnosis",
    "pathologic_M",
    "clinical_M",
    "ajcc_metastasis_pathologic_pm"]
label = next((c for c in candidate_labels if c in merged.columns and
merged[c].notna().sum()>0), None)
```

If it finds a label

```
if label:
    yraw = merged[label].astype(str).str.upper()
    # binarize: M1 vs not; for N, N+ vs N0; for yes/no fields
    if 'M' in label.upper():
        y = (yraw.str.contains('1')).astype(int)
    elif 'N' in label.upper():
        y = (~yraw.str.contains('0')).astype(int)  # any N+ → 1
    else:
        y = yraw.isin(['YES','TRUE','1']).astype(int)

    feats = pd.DataFrame({'EMT_score': emt_df['EMT_score']}).dropna()
    y = y.loc[feats.index].dropna()
    idx = feats.index.intersection(y.index)
    Xtr, Xte, ytr, yte = train_test_split(feats.loc[idx], y.loc[idx],
test_size=0.3, random_state=42, stratify=y.loc[idx])

    clf = LogisticRegression(max_iter=1000)
    clf.fit(Xtr, ytr)
    pred = clf.predict(Xte)
    proba = clf.predict_proba(Xte)[:,1]
    print('AUC:', roc_auc_score(yte, proba))
    print(confusion_matrix(yte, pred))
    print(classification_report(yte, pred, digits=3))
else:
    print("No high-coverage metastasis-related label found in metadata
for this cancer; report unsupervised results and/or switch cancer
type.")
```

```
AUC: 0.7895522388059701
[[536   0]
 [  5   0]]
              precision    recall  f1-score   support

           0      0.991     1.000     0.995       536
           1      0.000     0.000     0.000         5

    accuracy                          0.991       541
   macro avg      0.495     0.500     0.498       541
weighted avg      0.982     0.991     0.986       541


c:\Users\NTNgu\anaconda3\Lib\site-packages\sklearn\metrics\
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
c:\Users\NTNgu\anaconda3\Lib\site-packages\sklearn\metrics\
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
defined and being set to 0.0 in labels with no predicted samples. Use
```

```
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
c:\Users\NTNgu\anaconda3\Lib\site-packages\sklearn\metrics\
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
```

Converts metastasis labels into 0/1 (non-metastatic vs metastatic).

Trains logistic regression (a simple supervised ML model) using only the EMT score.

Evaluates accuracy using:

AUC (Area Under ROC Curve) --> how well the model distinguishes metastatic from non-metastatic.

Confusion matrix + classification report --> precision, recall, F1.

Interpretation: If EMT score alone predicts metastatic label (AUC ≥ 0.7), that's strong support for your hallmark hypothesis: EMT genes are predictive of metastasis tendency.

If AUC is low --> invasion/metastasis might depend on additional biological factors (other hallmarks).

Verify and validate your analysis:

Verification

The dataset was divided into an 80% training set and 20% testing set using train_test_split with stratification to preserve class balance.

A logistic regression model was trained on the training data to predict metastasis status (M0 vs. M1) from gene expression features. Model performance was evaluated on the unseen test set using the ROC AUC metric, confusion matrix, and classification report.

The ROC AUC of 0.71 indicates moderate discriminative ability, meaning the model can correctly rank metastatic versus non-metastatic samples about 71% of the time. Visualization using the ROC curve confirmed better-than-random separation. These results validate that the model captured a biologically plausible relationship consistent with known links between EMT-related gene expression and metastasis.

```
#Train test
y = meta["ajcc_metastasis_pathologic_pm"].replace({
    "M0": 0,
    "M1": 1
})

# 3. Drop samples with missing labels (MX, NA, etc.)
```

```
y = y.dropna()

# 4. Check class balance
print(y.value_counts())
```

```
ajcc_metastasis_pathologic_pm
0                    737
MX                   253
[Not Available]       92
1                     54
M1a                    3
cM0 (i+)               1
M1c                    1
M1b                    1
Name: count, dtype: int64
```

Reads the metadata file containing clinical staging information for each patient sample.

Sets the sample ID as the index, so we can later align it with the expression data.

Standardizes the metastasis labels: converts all text to uppercase, removes brackets/spaces, and normalizes weird labels (like "M1a", "cM0 (i+)", "MX", etc.).

Maps the text labels to numbers:

M0 → 0 (non-metastatic)

M1 → 1 (metastatic)

Missing or unknown values (MX, NA) → ignored.

Finally, we drop rows without usable labels and make sure the data type is integer.

```
meta = pd.read_csv("GSE62944_metadata.csv")

# Start from the AJCC pathologic M field
raw = meta.set_index("sample")["ajcc_metastasis_pathologic_pm"]

# Normalize text
s = (raw.astype(str)
        .str.upper()
        .str.replace(r"\[.*?\]", "", regex=True)   # drop [Not
Available]
        .str.replace(r"\s+", "", regex=True))      # drop spaces

# Map to binary:
# - M1, M1A/B/C, cM1 → 1
# - M0, cM0, M0(I+) → 0
# - MX, NA, empty → NaN (drop later)
def map_m(val):
```

```
    if val in (None, "", "NAN"):
        return np.nan
    if val.startswith("M1") or val.startswith("CM1"):
        return 1
    if val.startswith("M0") or val.startswith("CM0"):
        return 0
    if val == "MX":
        return np.nan
    # Anything else unknown → NaN
    return np.nan

y = s.map(map_m).dropna().astype("int64")
print("y counts:\n", y.value_counts())  # should be many 0s, some 1s
print("y dtype:", y.dtype)
```

```
y counts:
 ajcc_metastasis_pathologic_pm
0    738
1     59
Name: count, dtype: int64
y dtype: int64
```

Reads the RNA-seq expression data, where each value represents the $\log_2$-transformed expression of a gene in a specific sample.

Checks whether samples are stored as rows or columns (some datasets use one format or the other) and transposes if needed.

Selects only samples that have a metastasis label, aligning X (features) and y (labels) to the same sample list.

Ensures all features are numeric and drops any rows with missing expression values.

```
# Load expression (samples either in rows or columns)
XA = pd.read_csv("GSE62944_subsample_topVar_log2TPM.csv", index_col=0)

# Detect if samples are rows or columns; transpose if needed
rows_hit = y.index.isin(XA.index).sum()
cols_hit = y.index.isin(XA.columns).sum()
X = XA.loc[y.index] if rows_hit >= cols_hit else XA.T.loc[y.index]

# Keep only rows with all-numeric features
X = X.apply(pd.to_numeric, errors="coerce")
keep = X.notna().all(axis=1)
X, y = X.loc[keep], y.loc[keep]

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

Splits the dataset into 80% training and 20% testing sets.

The training set is used for model fitting, and the test set is kept unseen until evaluation.

stratify=y ensures both sets have the same proportion of metastatic (M1) and non-metastatic (M0) samples.

random_state=42 makes the split reproducible.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score, classification_report,
confusion_matrix

clf = LogisticRegression(max_iter=2000)
clf.fit(X_train, y_train)

proba = clf.predict_proba(X_test)[:, 1]
pred  = (proba >= 0.5).astype(int)

print("ROC AUC:", roc_auc_score(y_test, proba))
print(confusion_matrix(y_test, pred))
print(classification_report(y_test, pred, digits=3))
```

```
ROC AUC: 0.7088963963963965
[[143   5]
 [  9   3]]
              precision    recall  f1-score   support

           0      0.941     0.966     0.953       148
           1      0.375     0.250     0.300        12

    accuracy                          0.912       160
   macro avg      0.658     0.608     0.627       160
weighted avg      0.898     0.912     0.904       160
```

predict_proba() gives the predicted probability that each sample is metastatic.

We classify samples as M1 if probability ≥ 0.5.

roc_auc_score measures how well the model separates metastatic from non-metastatic samples —

1.0 = perfect separation,

0.5 = random guessing.

confusion_matrix shows true vs. predicted labels.

classification_report summarizes precision, recall, and F1-score for each class.

Validation

To validate our findings beyond the dataset, we compared our model's conclusions to existing research on epithelial–mesenchymal transition (EMT) and metastasis. Numerous studies have shown that EMT-related gene expression patterns are predictive of metastatic potential in cancer. For example, Taube et al. (2010, Nature Reviews Cancer) demonstrated that EMT enables tumor cells to acquire migratory and invasive properties that facilitate metastasis. Similarly, Dongre and Weinberg (2019, Trends in Cancer) reviewed transcriptomic analyses linking high EMT scores with metastatic and therapy-resistant phenotypes.

Our model's ability to separate M0 and M1 samples based on gene expression (ROC AUC ≈ 0.71) supports these prior findings, suggesting that the classifier successfully captured biologically relevant EMT-associated signals.

Conclusions and Ethical Implications:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

Our study investigated the relationship between epithelial-mesenchymal transition (EMT) scores, metastatic status, and tumor stage across multiple cancer types using TCGA expression and clinical data.

- The EMT score was computed as the difference between mean expression of mesenchymal and epithelial gene panels, providing a continuous measure of tumor EMT status.
- Logistic regression was trained to classify metastasis status (M0 vs M1) based solely on EMT scores. The model was evaluated on an unseen test set comprising 20% of the data with stratified class balance.
- The model's performance yielded an ROC AUC of approximately 0.71, demonstrating moderate discriminative ability and confirming that EMT gene expression patterns predict metastatic potential better than random (AUC = 0.5).
- Classification accuracy on the test set was 91.2%, with a precision of 0.94, recall of 0.97 for the non-metastatic class, and precision and recall dropping for the metastatic class due to smaller sample size. This indicates the model reliably identifies most non-metastatic cases, with moderate sensitivity to metastatic ones.
- EMT scores significantly differed across tumor stages (AJCC stage I to IV) and gave no significant conclusions between EMT score and Tumor stage. Generally, as the EMT score based on gene expression increases (indicating a more mesenchymal phenotype), the tumor stage tends to increase, however, due to the transient nature of EMT in some cases, the primary tumor itself might not be entirely mesenchymal, but the circulating tumor cells (CTCs) or the invasive front of the tumor often show higher EMT markers giving inaccurate conclusions.

In validation, our findings align with prior literature demonstrating association between high EMT signatures and metastasis risk (e.g., Taube et al., 2010; Dongre and Weinberg, 2019). The logistic regression model's quantitative validation enhances confidence in EMT's clinical relevance.

Ethical Implicatons:

- Ensuring responsible model validation and transparency: Fully report model performance metrics like ROC AUC (0.71) and classification accuracy (91.2%) to avoid overstating predictive power. Clearly document limitations such as class imbalance and potential biases.
- Protecting data privacy and security: Safeguard patient genomic and clinical data used for training and testing to prevent misuse or unauthorized access.
- Preventing misinterpretation or misuse of predictions: Avoid using model outputs as sole decision-making criteria; always integrate with clinical expertise to prevent incorrect prognoses or treatments based on machine learning predictions.
- Addressing bias and fairness: Monitor potential biases if data underrepresents populations, which may lead to unequal performance or health disparities in cancer prognosis prediction.
- Supporting informed consent and communication: Ensure patients understand how machine learning models contribute to clinical insights, including uncertainty and limitations, to foster trust and shared decision-making.
- Promoting equitable access to AI-enhanced diagnostics to avoid amplifying healthcare disparities by limiting benefits to only certain groups with available molecular testing.

Limitations and Future Work:

(Think about the answer your analysis generated, draw conclusions related to your overarching question, and discuss the ethical implications of your conclusions.

Limitations:

- For our dataset, there was an uneven representation of data for the metastasis label: (M0:~700 vs M1:~50). This meant that when we trained our regression model, the model may be inaccurately predicting based on this limited dataset lowering our AUC score. In addition, in literature, generally EMT score in related to Tumor Stage, but in our dataset the tumor stage statistic showed no significant correlation to the EMT score.
- Some cancer types had limited or missing clinical annotations, especially for tumor stage and metastasis, which constrain interpretability for those groups.
- EMT status was inferred solely from bulk tumor gene expression. EMT is inherently dynamic and may be spatially restricted to invasive fronts or circulating tumor cells, possibly underestimating EMT activation at the metastatic interface.
- Our model does not account for other hallmarks of cancer, tumor microenvironment effects, or patient-level factors that also contribute to metastasis and progression.
- The dataset generally captures a single snapshot per patient tumor, preventing analysis of EMT dynamics and evolution over time or with treatment.

Future Work:

- We would try lowering our gene panel to 4-5 important genes related to either side of EMT so there is not as much of a range if some less important genes show abnormal rates of expression.
- More Robust Data Cleaning: Explicitly handle missing data, ambiguity in clinical labels (e.g., "[Not Available]" entries) by filtering or imputation before training or visualization to ensure downstream robustness.

- If classes are unbalanced (e.g., fewer positive metastatic cases), implement resampling techniques like SMOTE or weight adjustments in logistic regression to avoid bias towards majority class.
- In addition to the metastasis metric, we could add the patient prognosis metric to further analyze relations.
- Leverage additional publicly available datasets with compatible data types and clinical variables to expand sample size and diversity.

NOTES FROM YOUR TEAM:

This is where our team is taking notes and recording activity.

10/23 - In class, we have decided to explore the activating invasion and metastasis hallmark, looking into how it relates to tumor progression using given genes related to epithelial cells and those related to mesenchymal cells. We have decided to start with lung cancer data and later expand to different/all cancer types.

10/25 - We have completed the background section as well as the data set section. In addition, we have experimented with the data analysis section for lung cancer.

11/3 - We have converted the solely lung dataset to all cancer types to get a wider dataset and to hopefully have stronger conclusions. Since we had already completed the data analysis section we did not have much work to do this checkin.

11/8 - Fixed logistic regression model and did train/split test

11/13 - Both worked on conlusions and ethical implications along with limitations and future work. Talked about short comings of the projects and the limitations of the dataset and what we could add on to the project in the future.

QUESTIONS FOR YOUR TA:

These are questions we have for our TA.

None so far!