Classification of Histopathological Textures of Colon Cancer

- Presentation by Ilay Sabah and Maria Alexeenko
- Lecturers - Miriam Carmon & Yakir Menachem
- Digital Medical Technologies

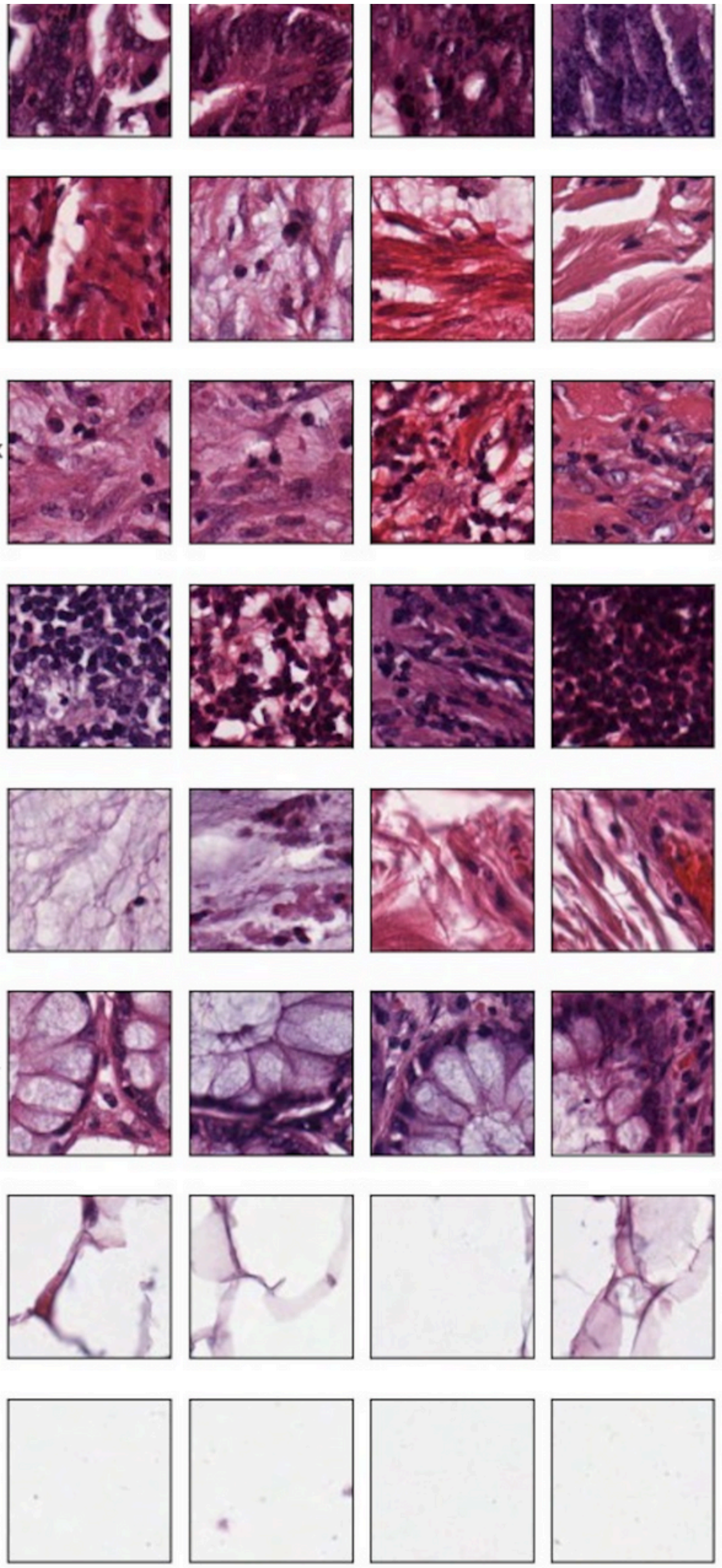
2025 - תשפ"ה
2nd Semester

Final Project
MACHINE LEARNING

מכון טכנולוגי חולון
Holon Institute of Technology

DATA SET

80% - train



Tumour epithelium

Simple stroma

Complex stroma

Immune cells

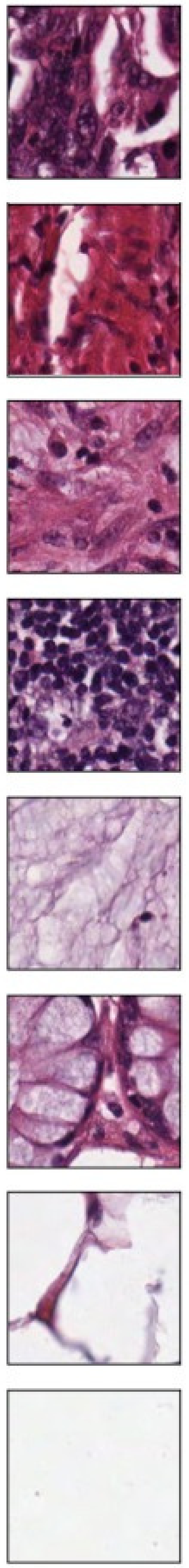
Debris

Normal mucosal glands

Adipose tissue

Empty

20% - test



GOAL

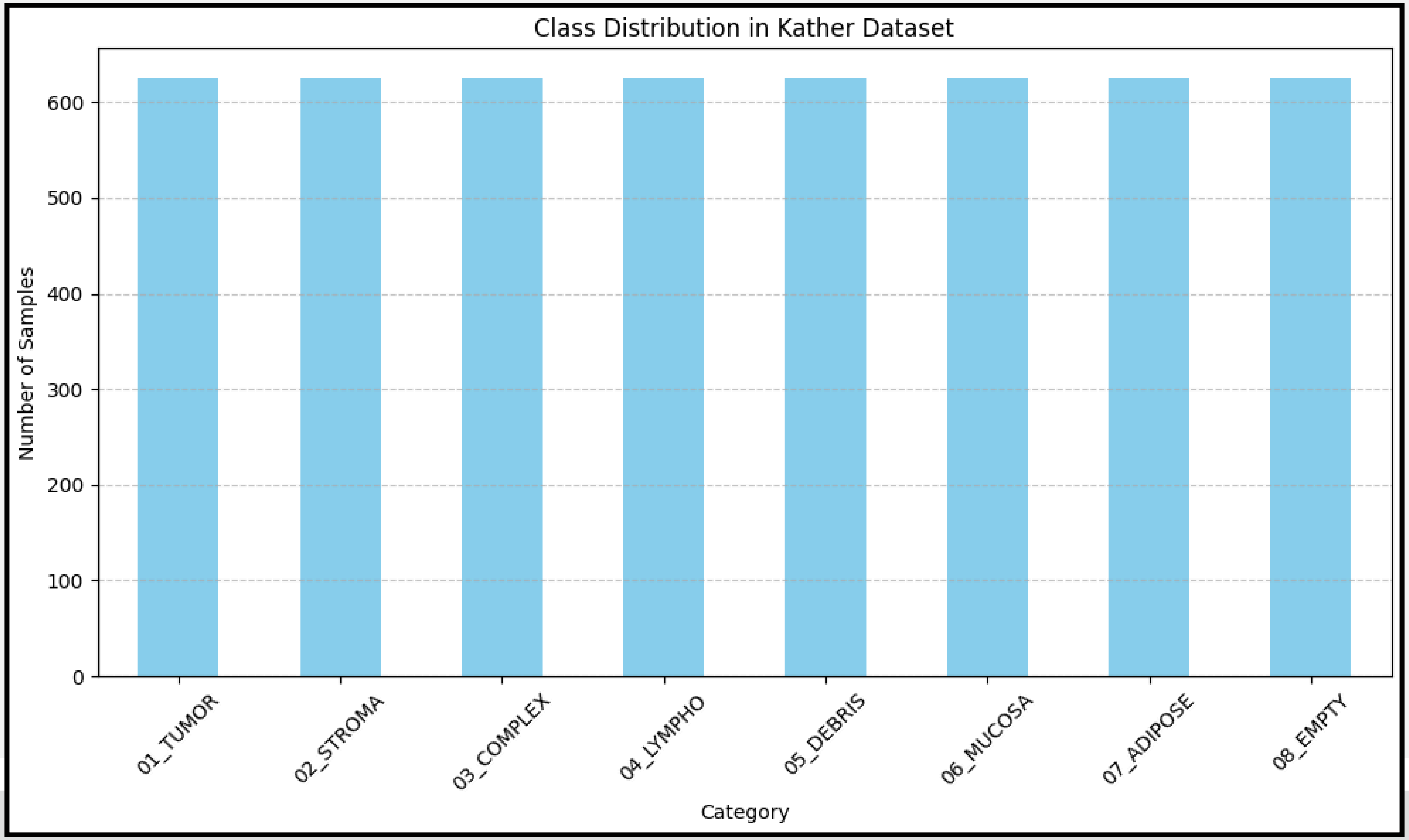
Method Comparison
Evaluate which feature extraction methods and models provide the highest accuracy

Tissue Classification
Automatically distinguish between 8 tissue types using RGB image patches

Clinical Application
Develop a model simulating the use of digital tools in tissue diagnostics

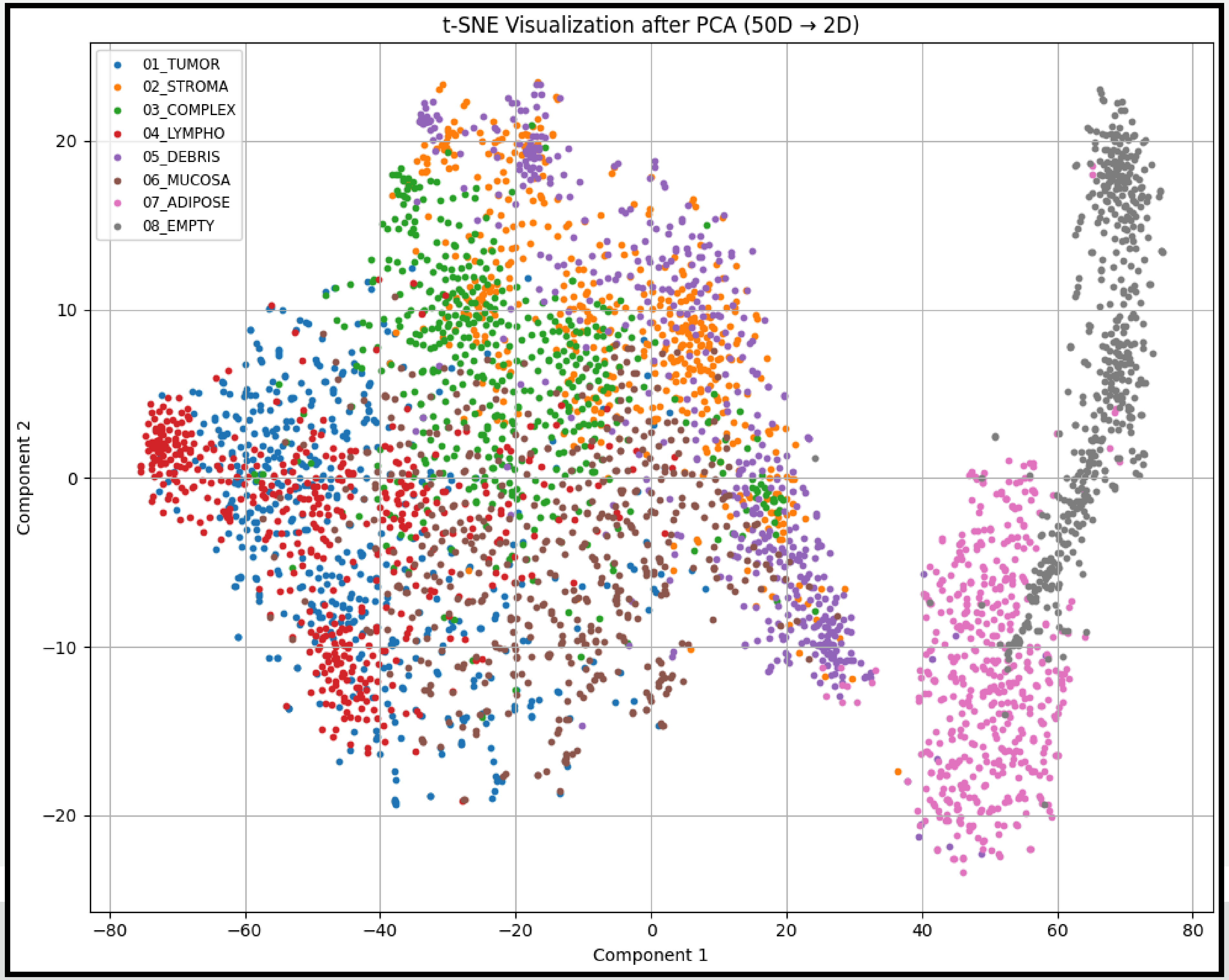
Model Comparison
Assess the effectiveness of deep learning vs. classical machine learning in texture classification tasks

INITIAL EDA



Class Distribution in Kather Dataset
Number of Samples
0
100
200
300
400
500
600
01_TUMOR
02_STROMA
03_COMPLEX
04_LYMPHO
05_DEBRIS
06_MUCOSA
07_ADIPOSE
08_EMPTY
Category

INITIAL EDA



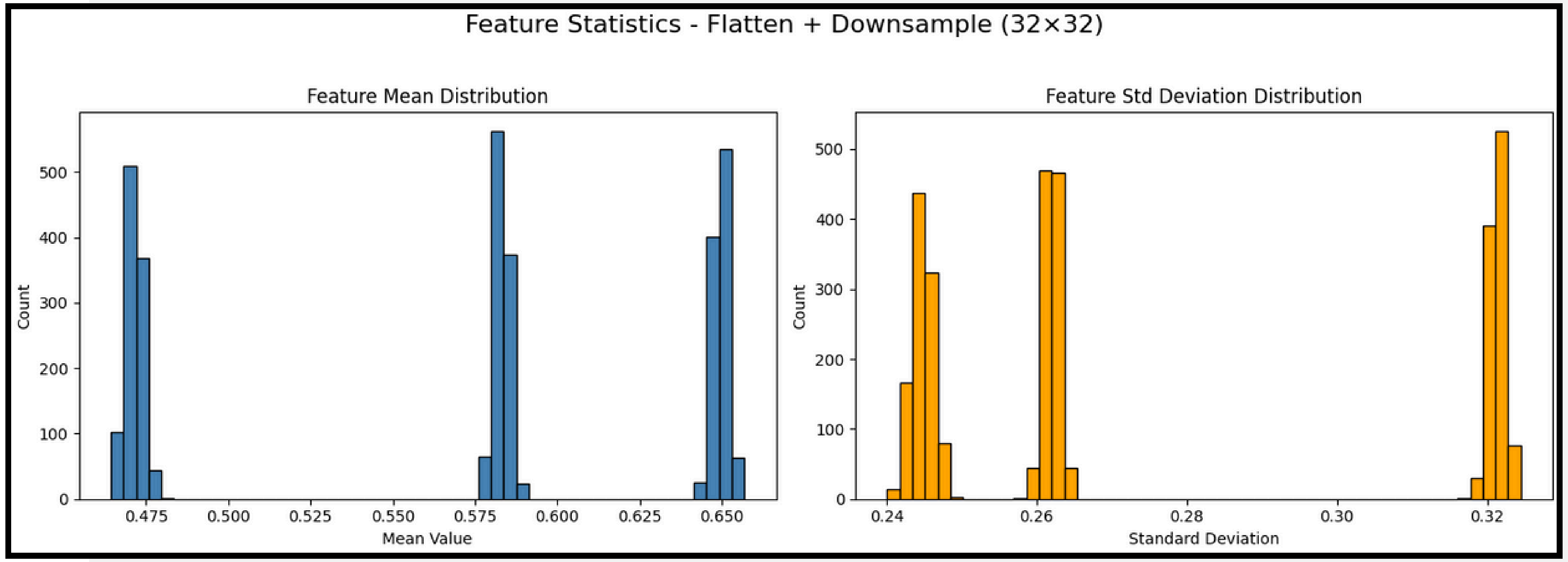
t-SNE Visualization after PCA (50D → 2D)
01_TUMOR
02_STROMA
03_COMPLEX
04_LYMPHO
05_DEBRIS
06_MUCOSA
07_ADIPOSE
08_EMPTY
Component 2
Component 1
20
10
0
−10
−20
−80
−60
−40
−20
0
20
40
60
80

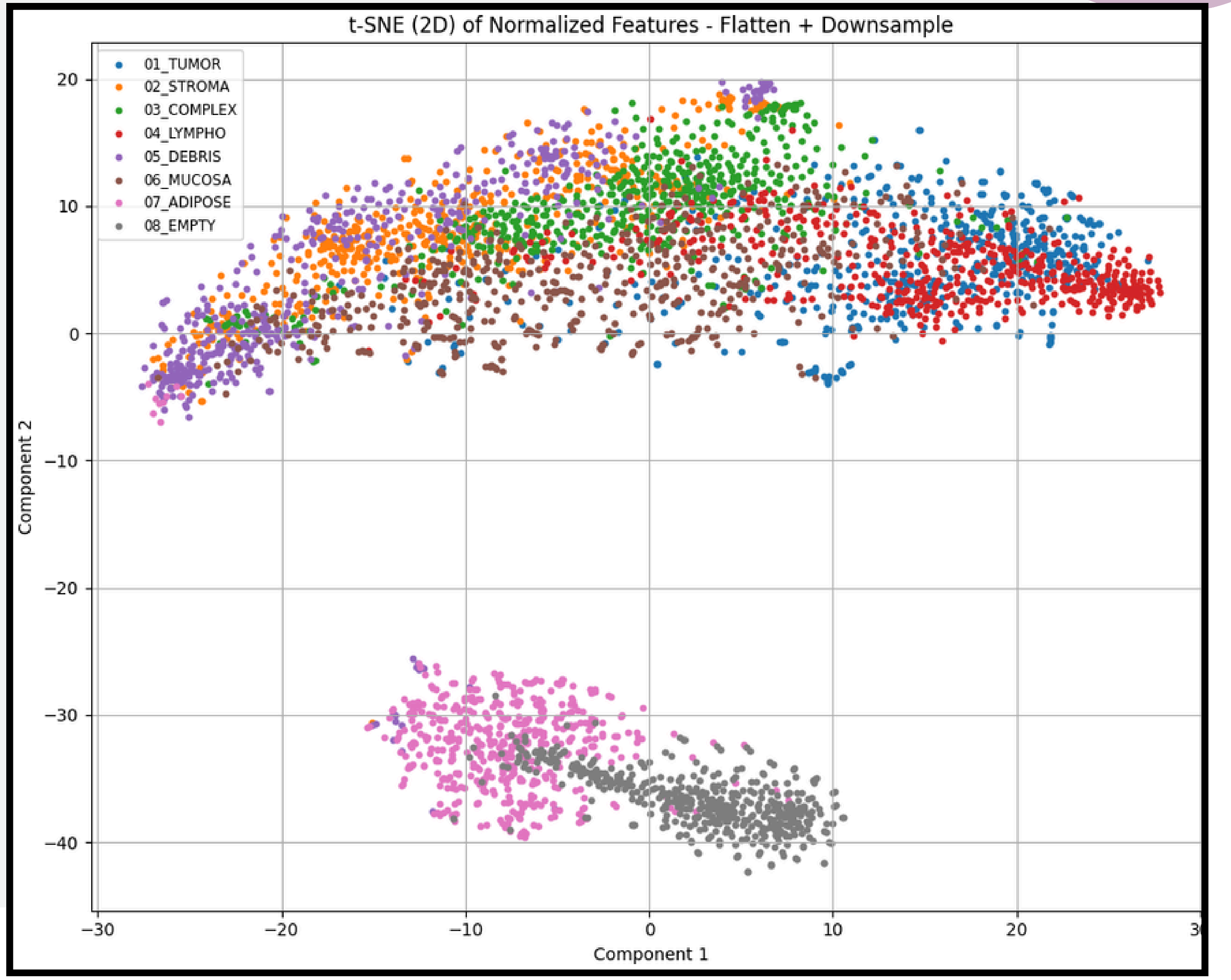
NORMALIZATION

Is It Necessary?

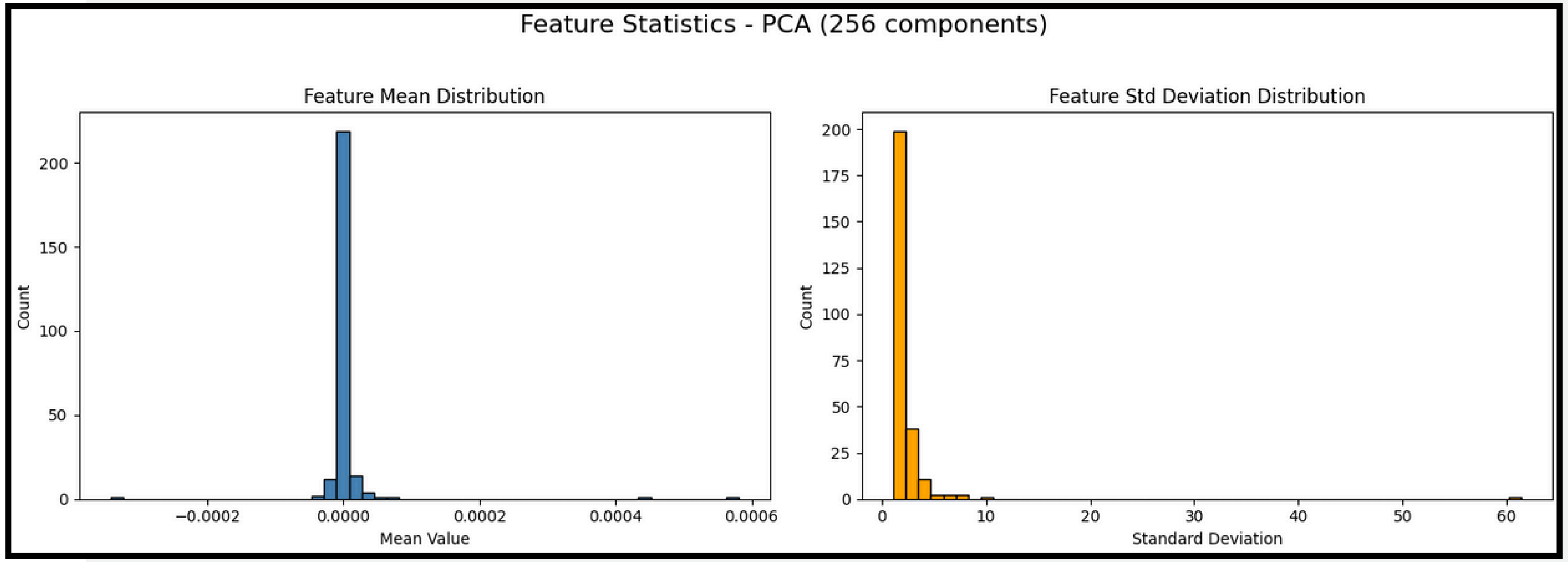
Yes, feature normalization is necessary after the feature extraction stage. This is because the standard deviation varies significantly across features, especially in methods like PCA and VGG16. Normalization is important to ensure that all features contribute equally to the learning process and to prevent features with larger scales from dominating the model's training.

FLATTEN + DOWNSAMPLE

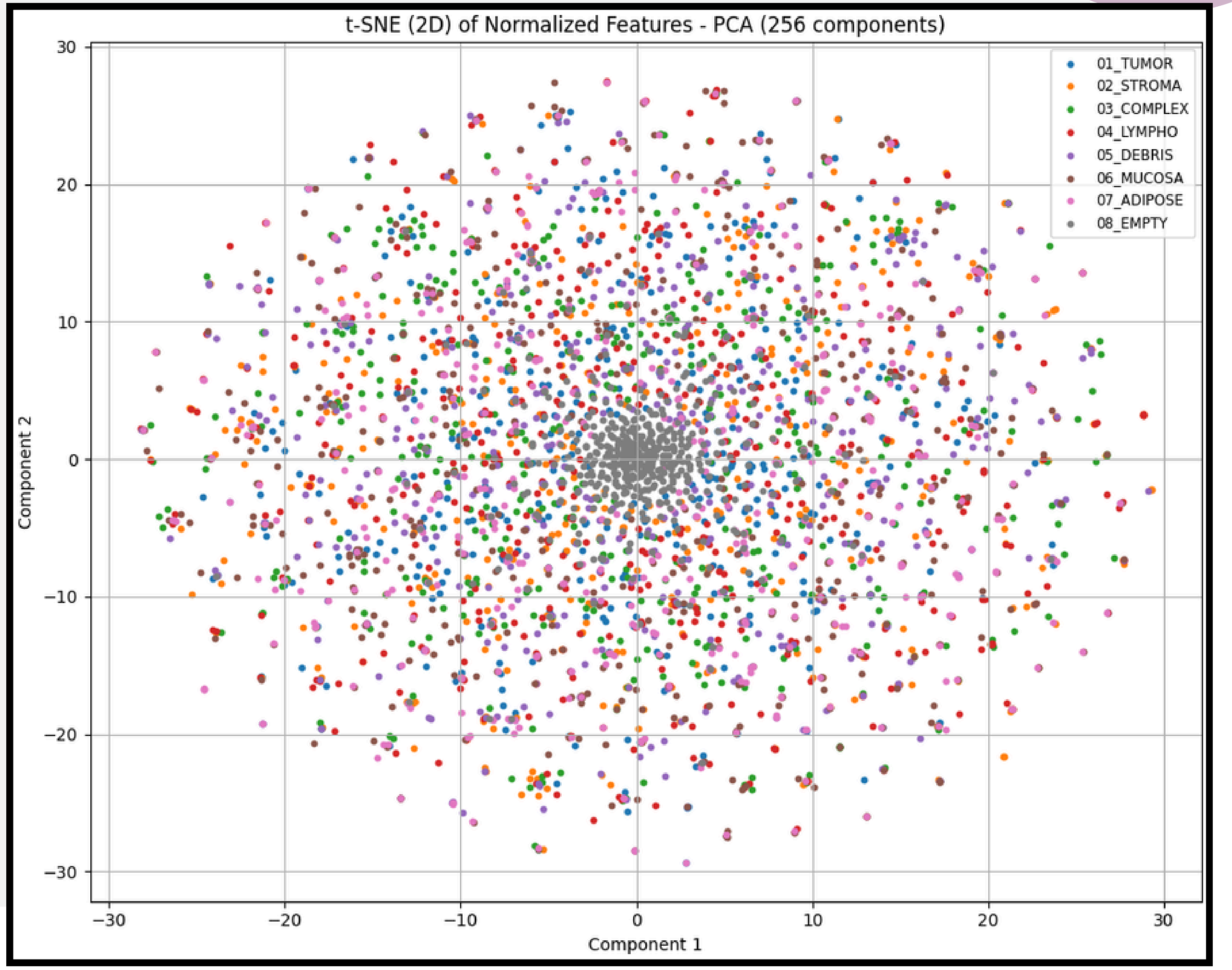




PCA

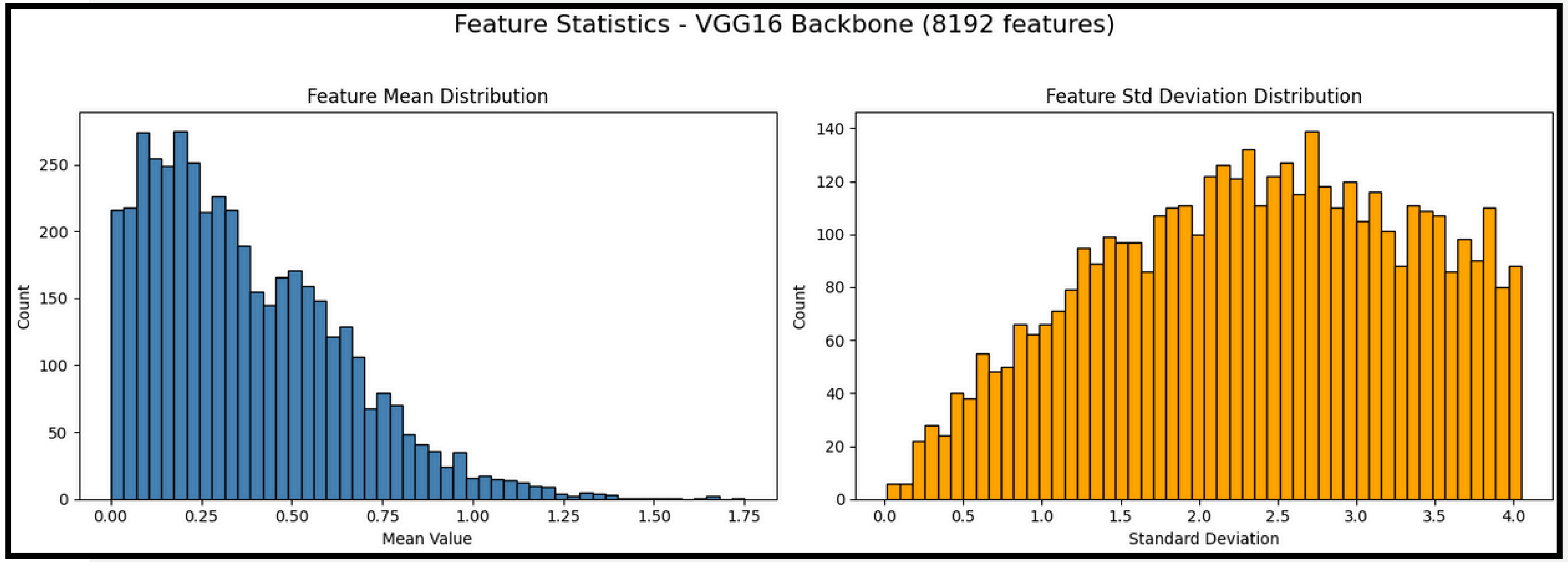


Feature Statistics - PCA (256 components)
Feature Mean Distribution
Count
Mean Value
Feature Std Deviation Distribution
Standard Deviation

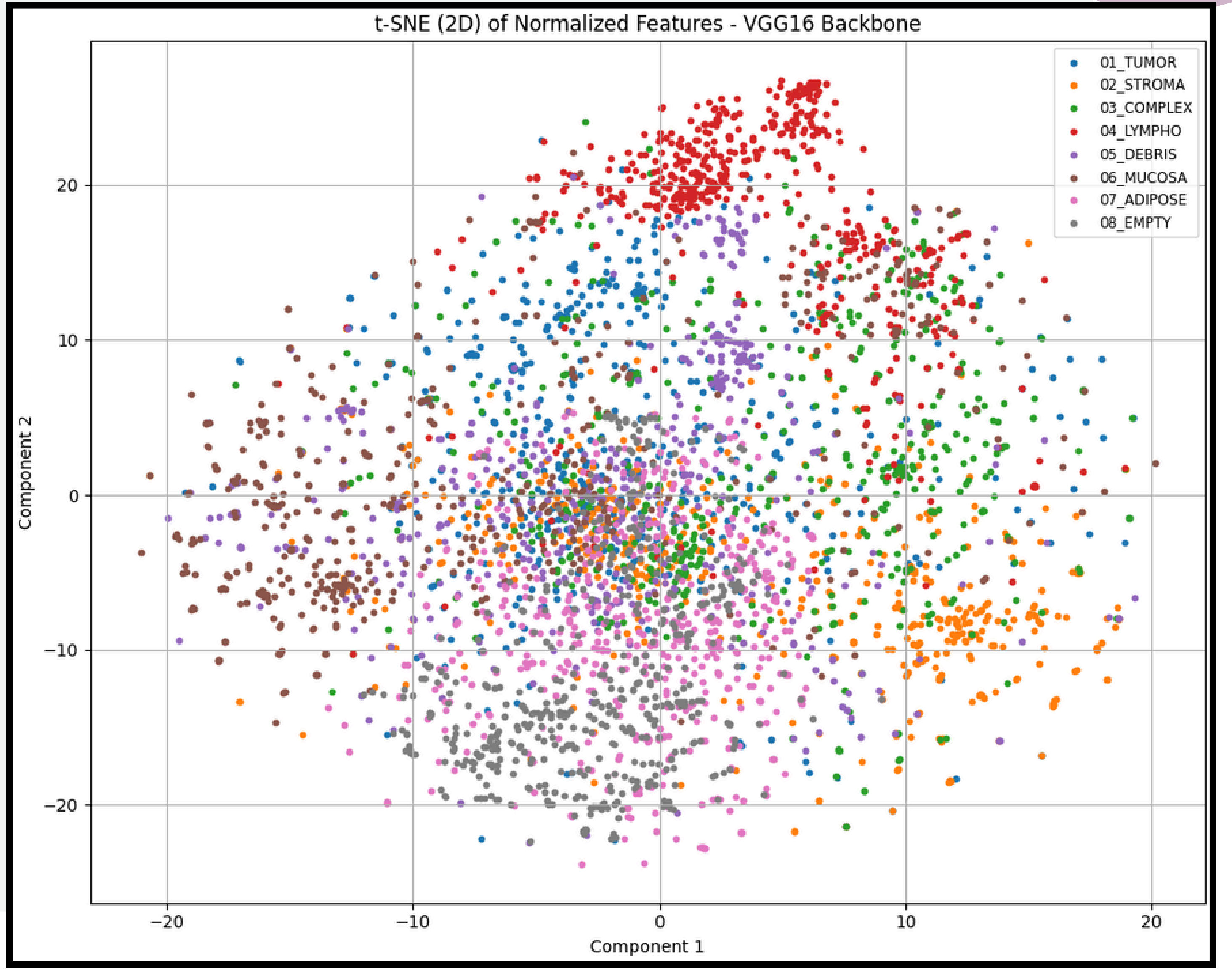


t-SNE (2D) of Normalized Features - PCA (256 components)
01_TUMOR
02_STROMA
03_COMPLEX
04_LYMPHO
05_DEBRIS
06_MUCOSA
07_ADIPOSE
08_EMPTY
Component 1
Component 2

VGG16



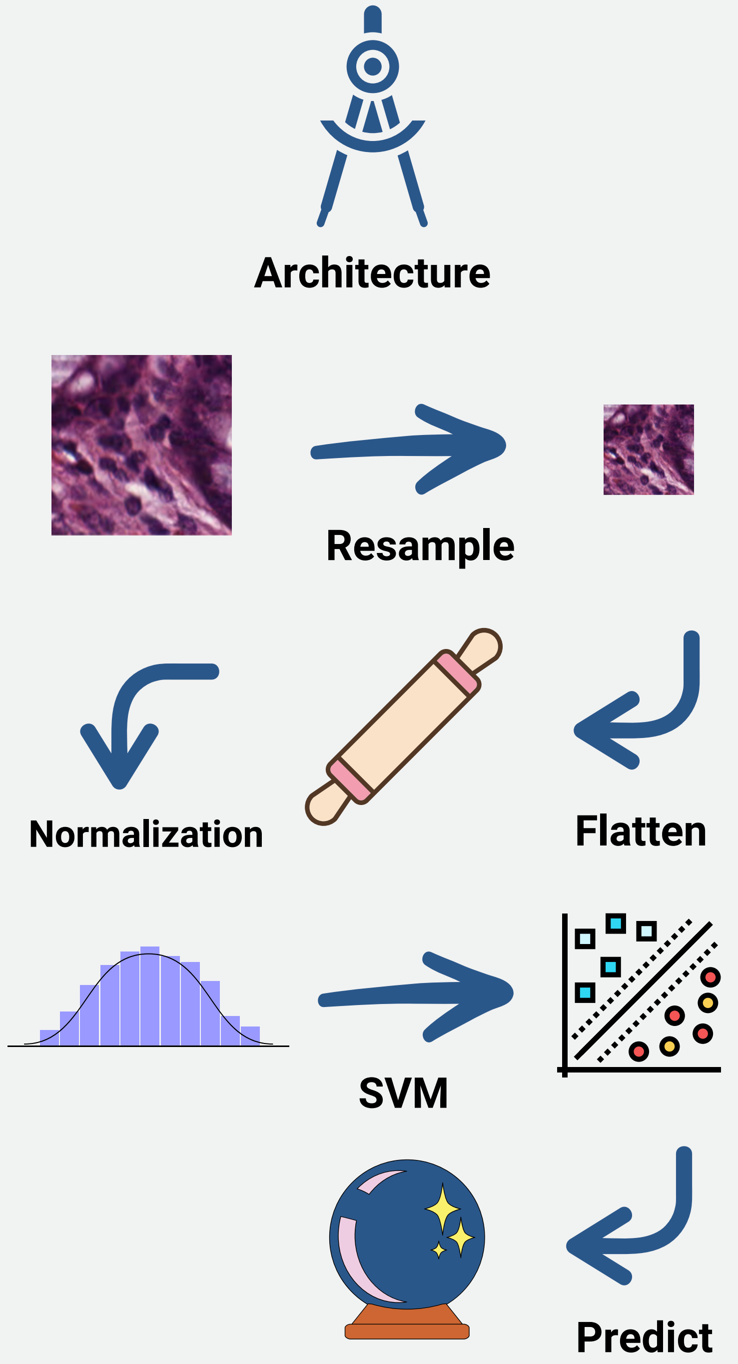
Feature Statistics - VGG16 Backbone (8192 features)
Feature Mean Distribution
Count
Mean Value
0.00 0.25 0.50 0.75 1.00 1.25 1.50 1.75
0 50 100 150 200 250
Feature Std Deviation Distribution
Count
Standard Deviation
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0
0 20 40 60 80 100 120 140

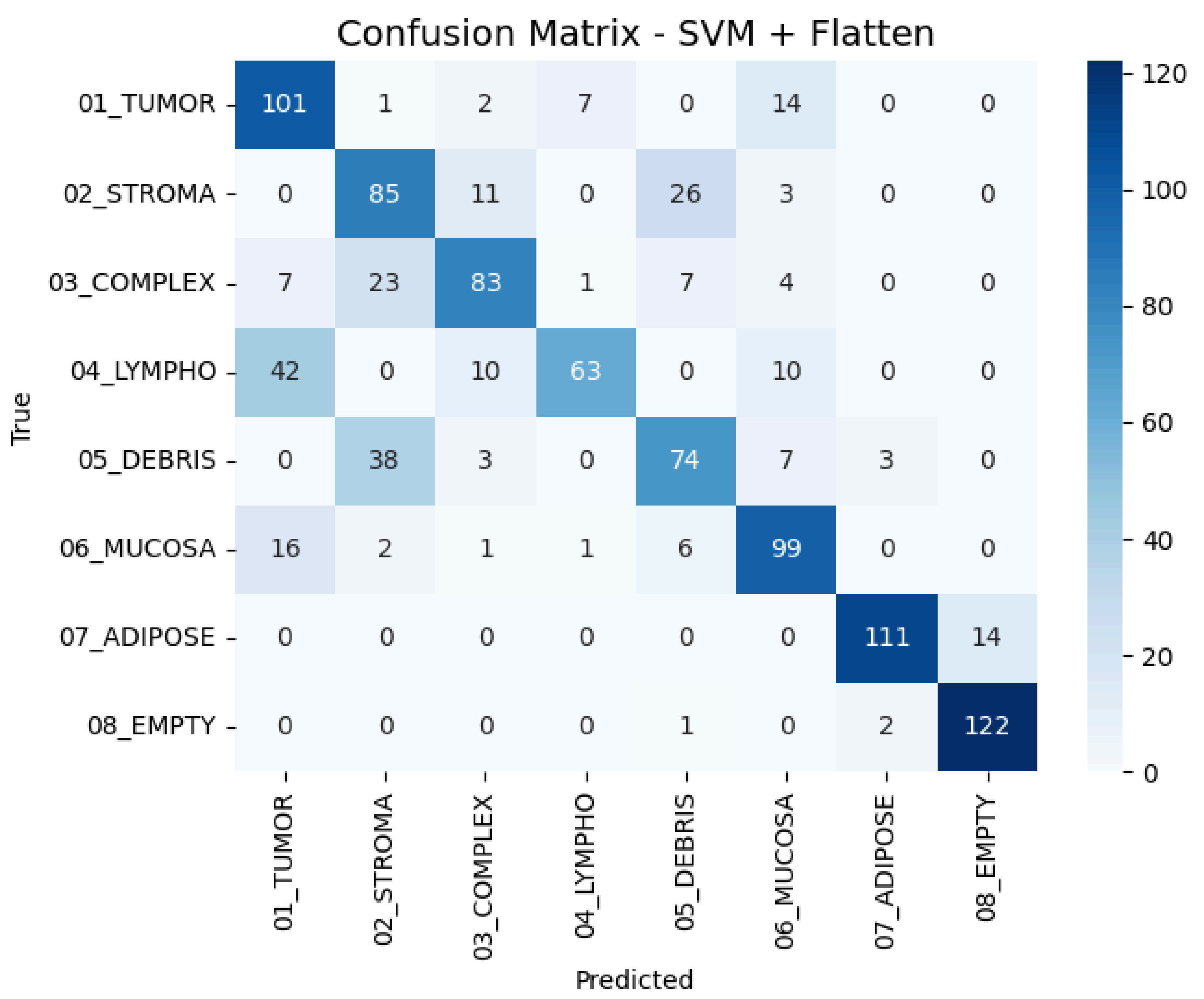


t-SNE (2D) of Normalized Features - VGG16 Backbone
01_TUMOR
02_STROMA
03_COMPLEX
04_LYMPHO
05_DEBRIS
06_MUCOSA
07_ADIPOSE
08_EMPTY
Component 1
Component 2
−20 −10 0 10 20

FIRST EXPERIMENT

SVM + FLATTEN





True \ Predicted	01_TUMOR	02_STROMA	03_COMPLEX	04_LYMPHO	05_DEBRIS	06_MUCOSA	07_ADIPOSE	08_EMPTY
01_TUMOR	101	1	2	7	0	14	0	0
02_STROMA	0	85	11	0	26	3	0	0
03_COMPLEX	7	23	83	1	7	4	0	0
04_LYMPHO	42	0	10	63	0	10	0	0
05_DEBRIS	0	38	3	0	74	7	3	0
06_MUCOSA	16	2	1	1	6	99	0	0
07_ADIPOSE	0	0	0	0	0	0	111	14
08_EMPTY	0	0	0	0	1	0	2	122

Accuracy → 0.84625

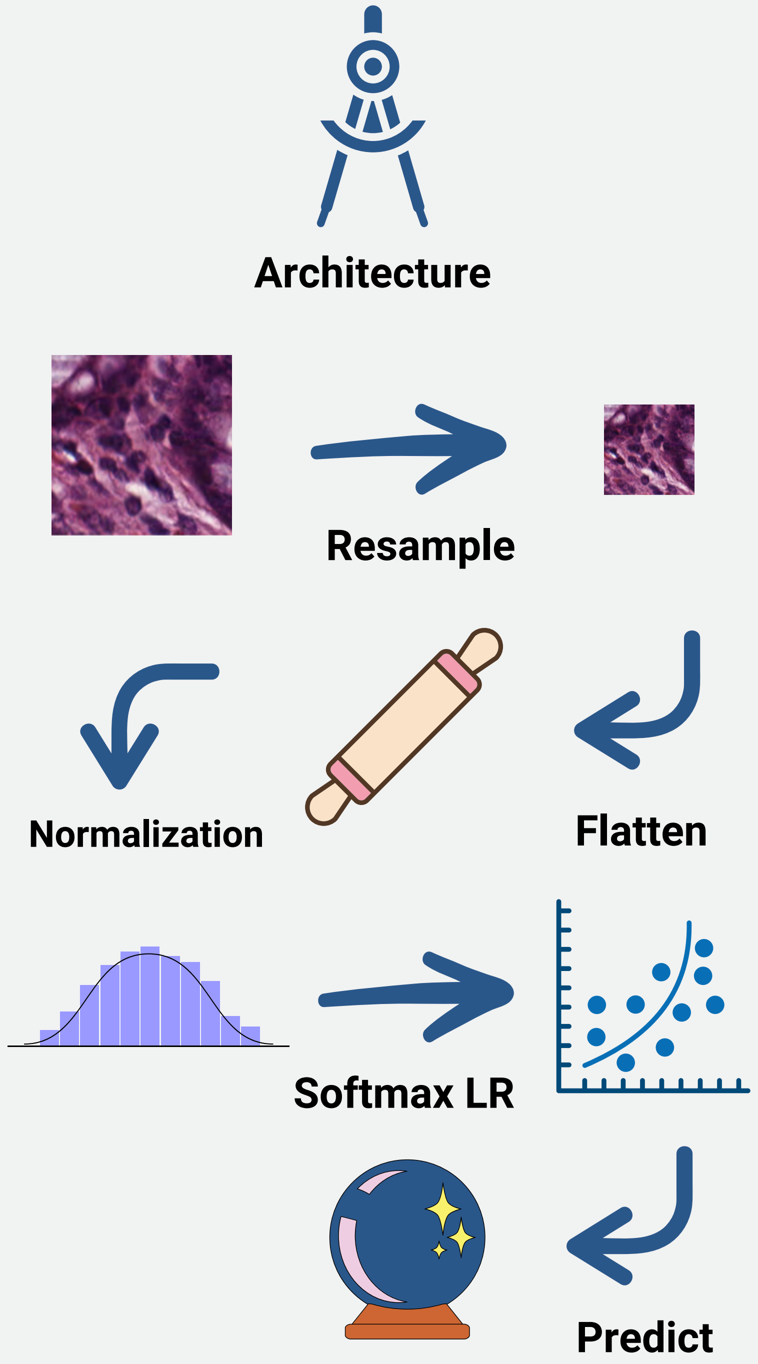
Test Accuracy → 0.738

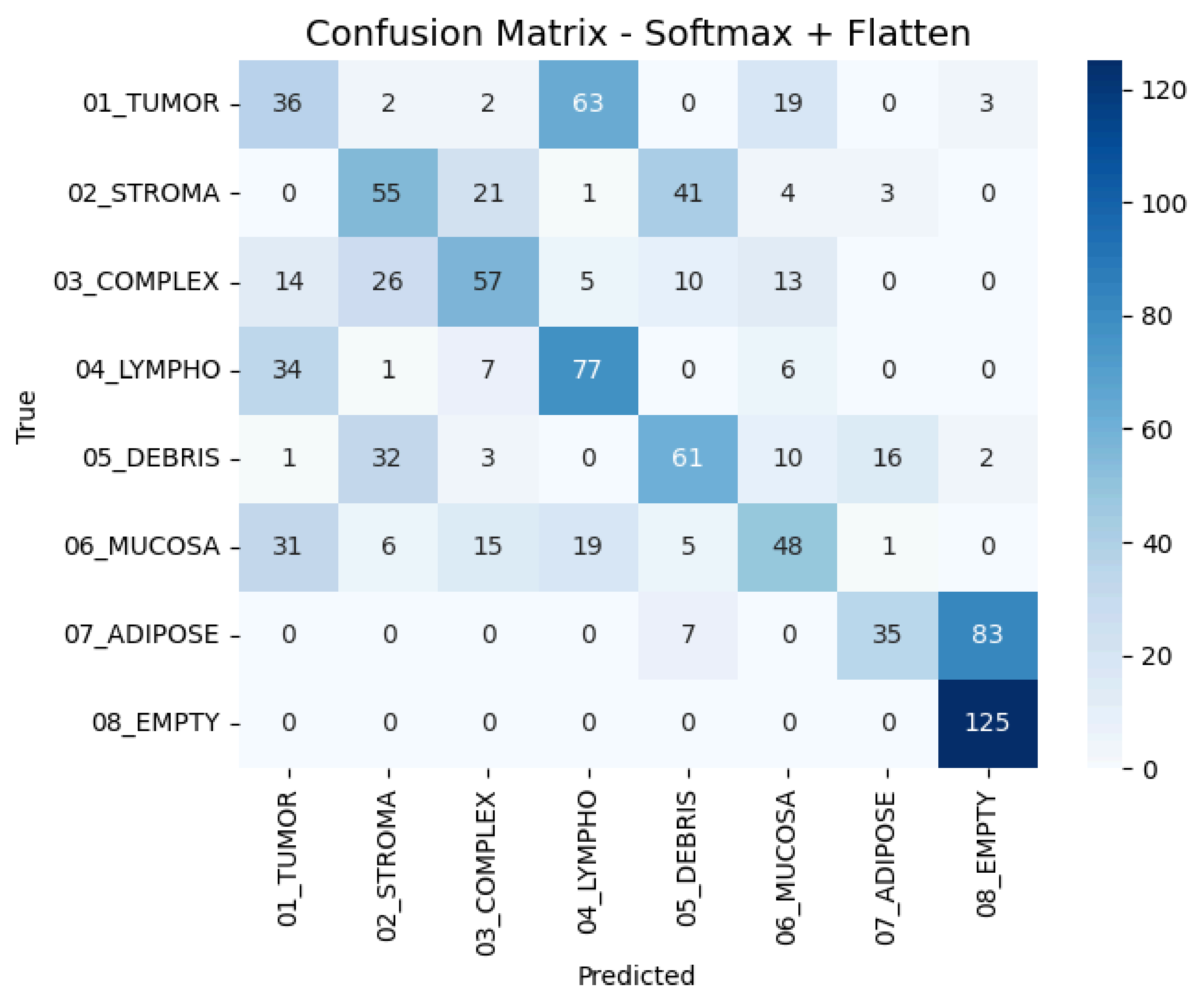
F1 Score → 0.736447

Overfitting Gap → 0.10825

SECOND EXPERIMENT

SOFTMAX + FLATTEN





Accuracy → 0.76850

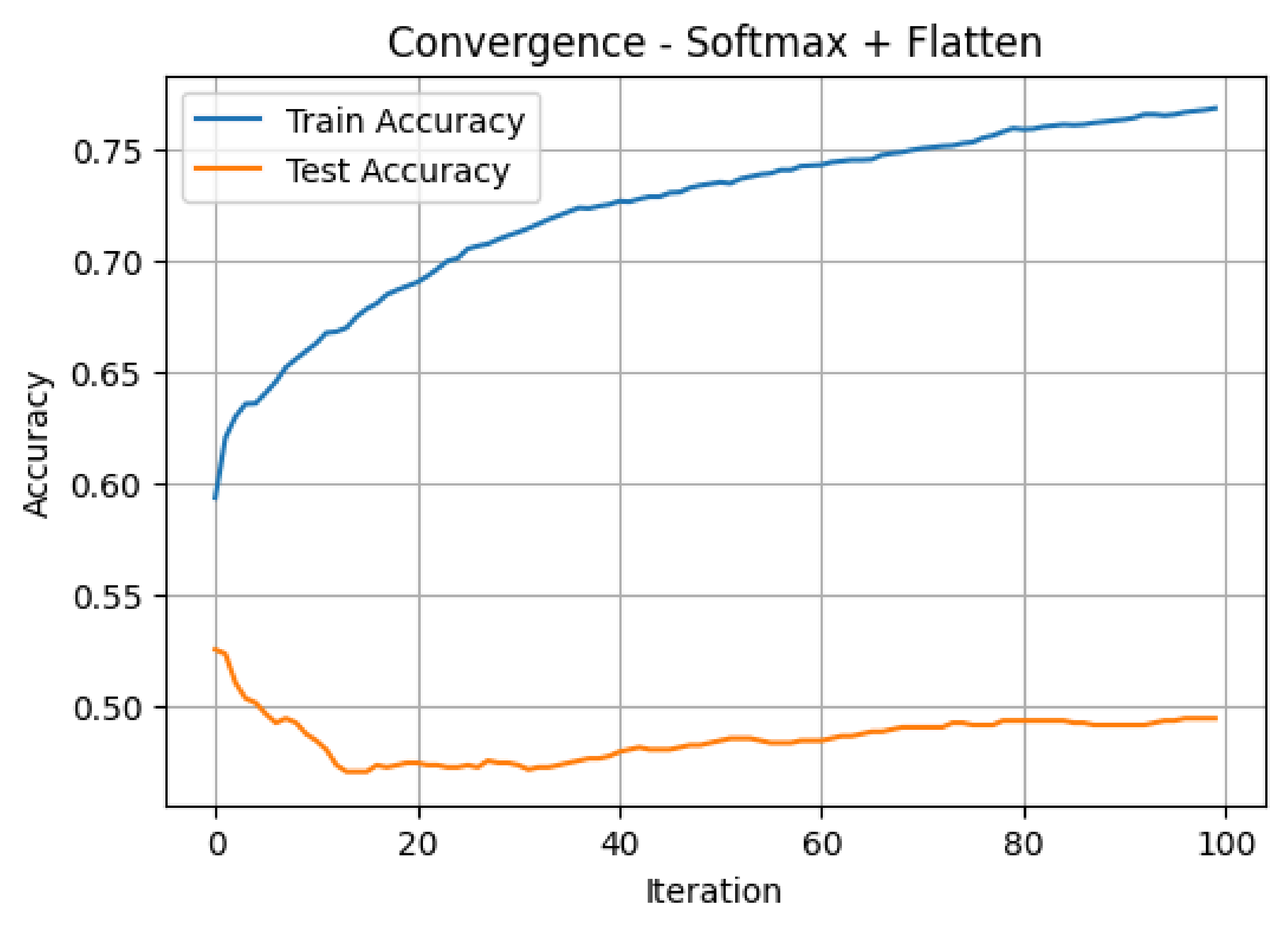
Test Accuracy → 0.494

F1 Score → 0.476993

Overfitting Gap → 0.27450

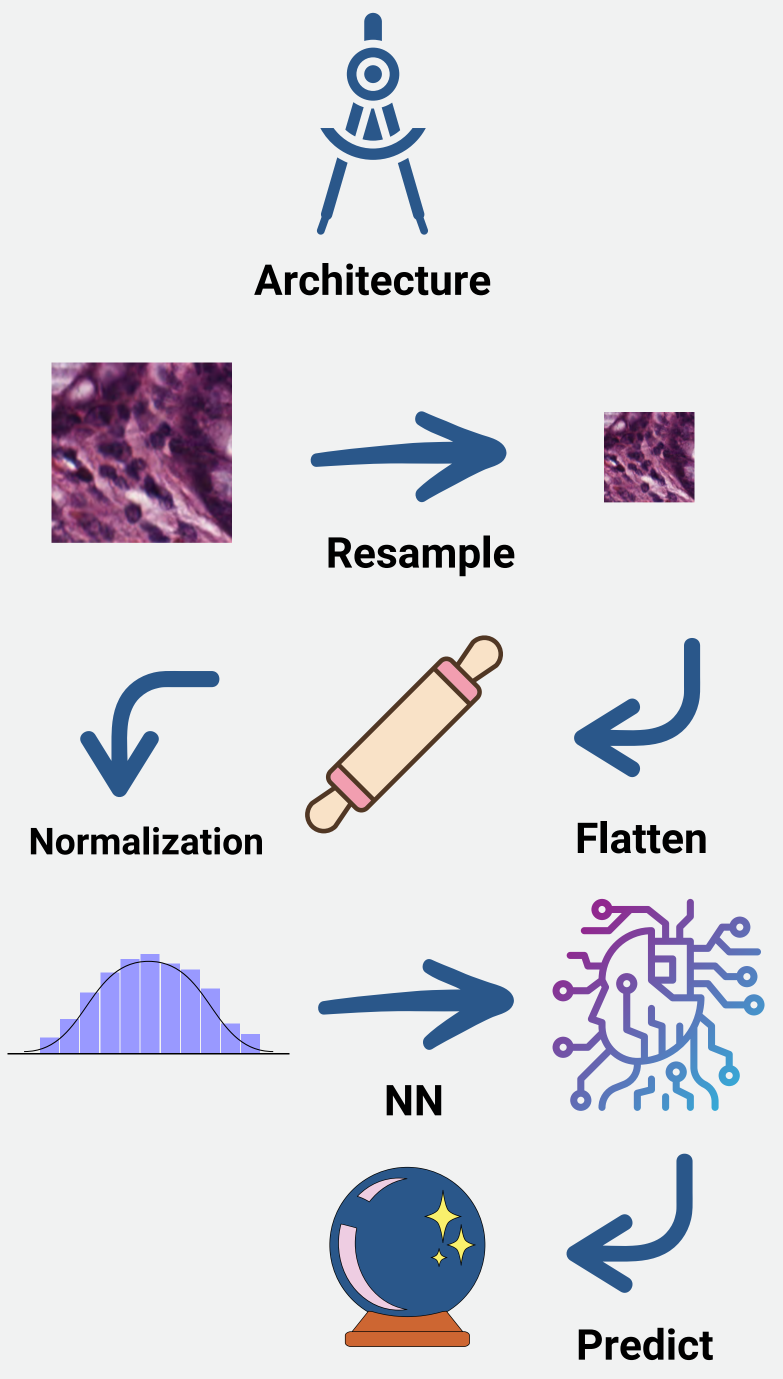
SECOND EXPERIMENT

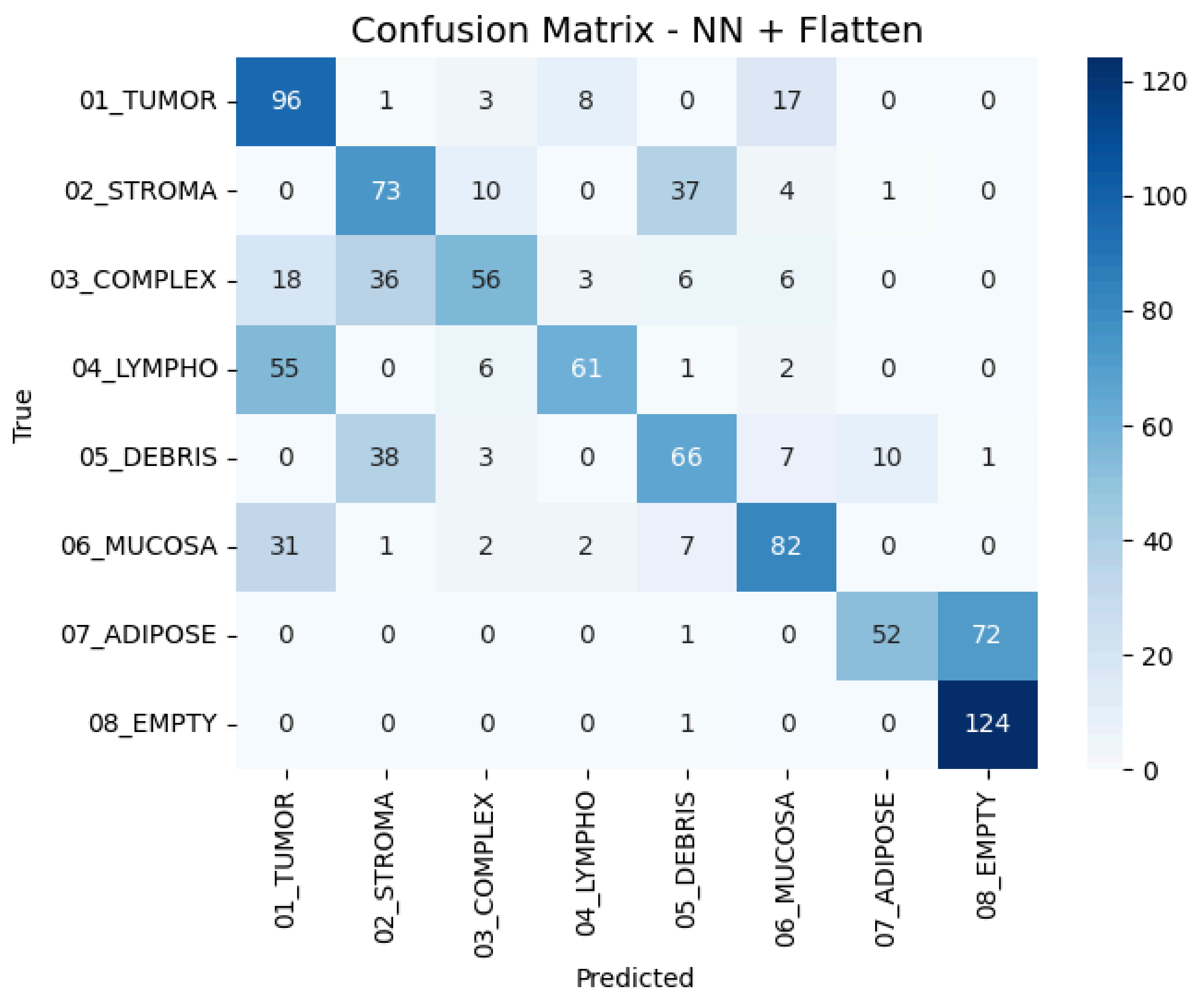
SOFTMAX + FLATTEN

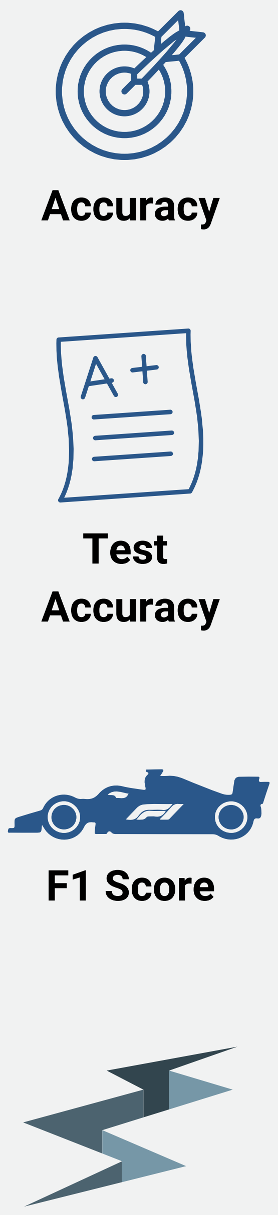


THIRD EXPERIMENT

NN + FLATTEN







Accuracy → 0.94350

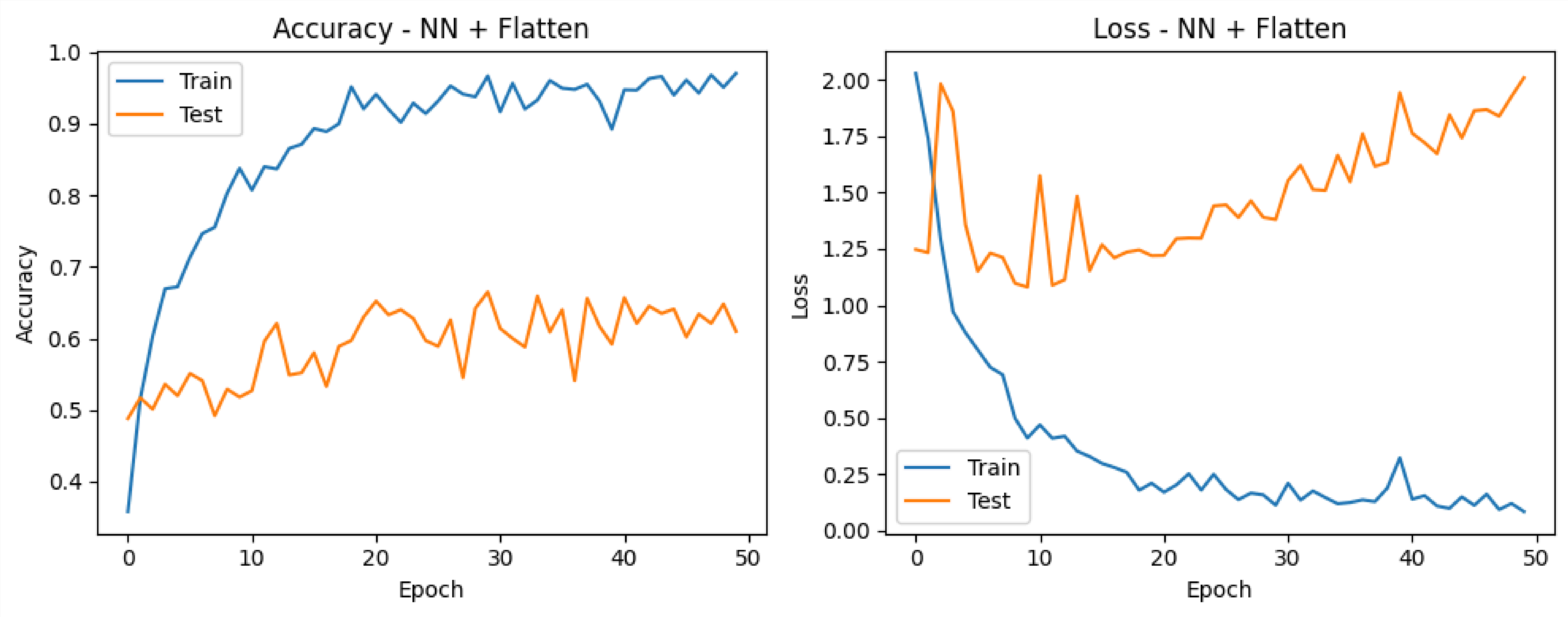
Test Accuracy → 0.610

F1 Score → 0.602785

Overfitting Gap → 0.33350

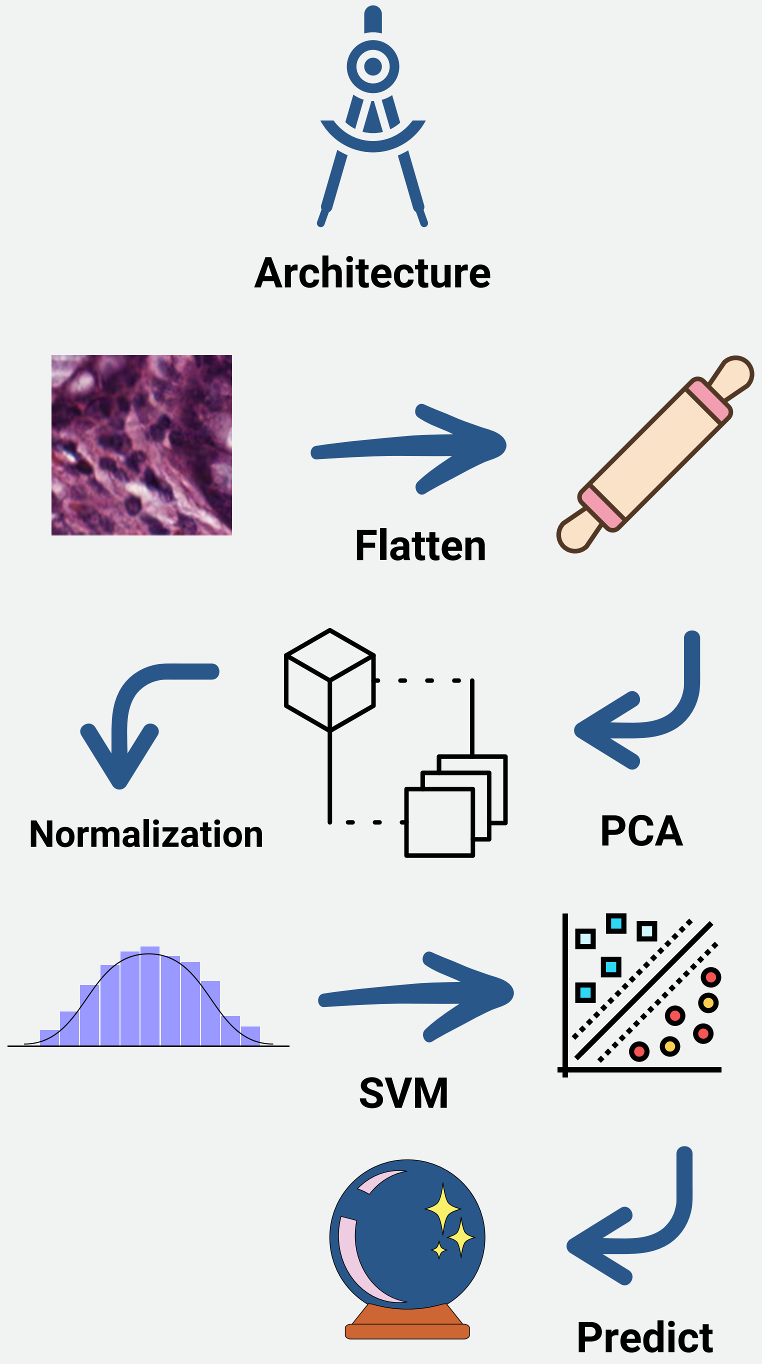
THIRD EXPERIMENT

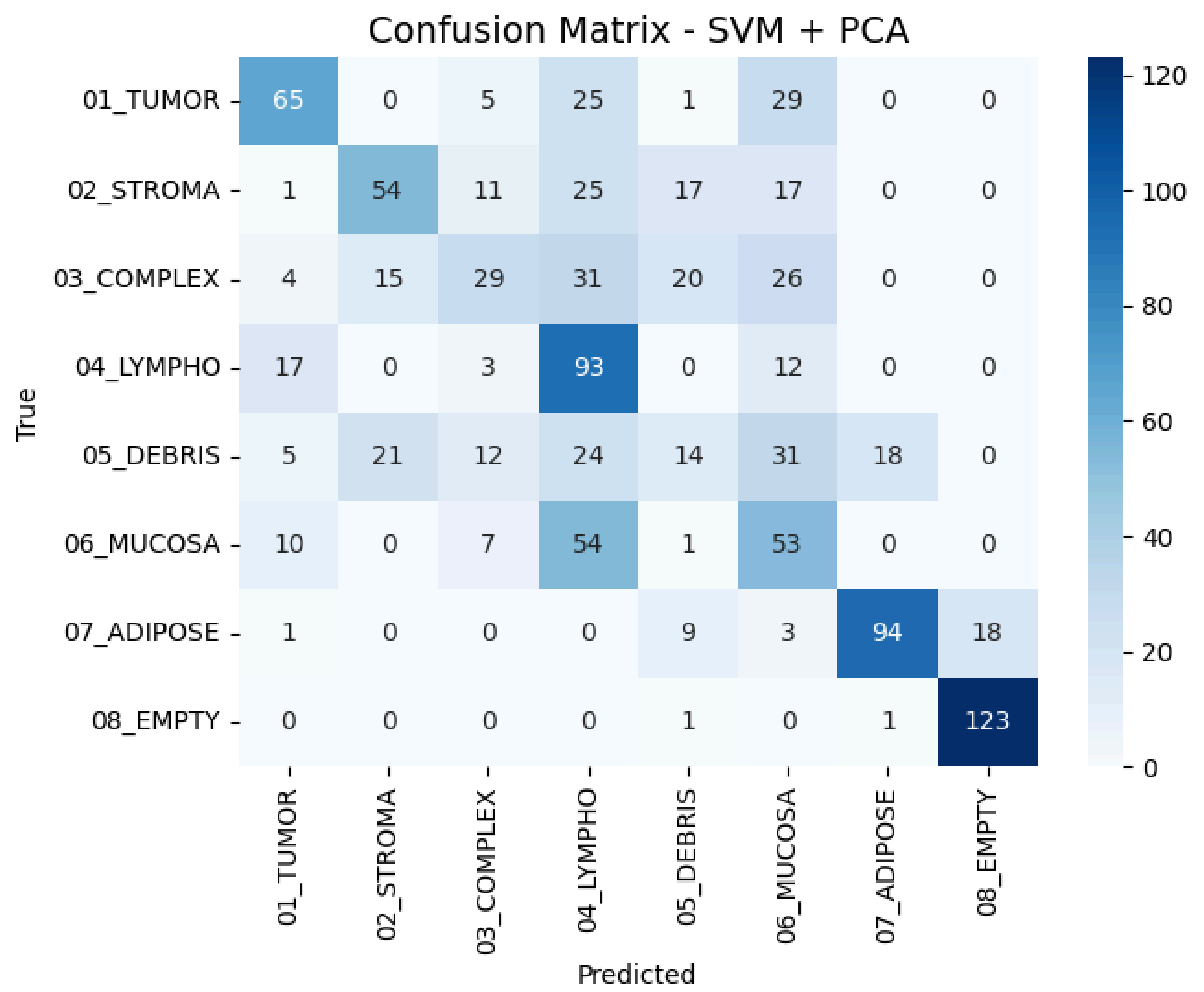
NN + FLATTEN



FOURTH EXPERIMENT

SVM + PCA





Accuracy → 0.95050

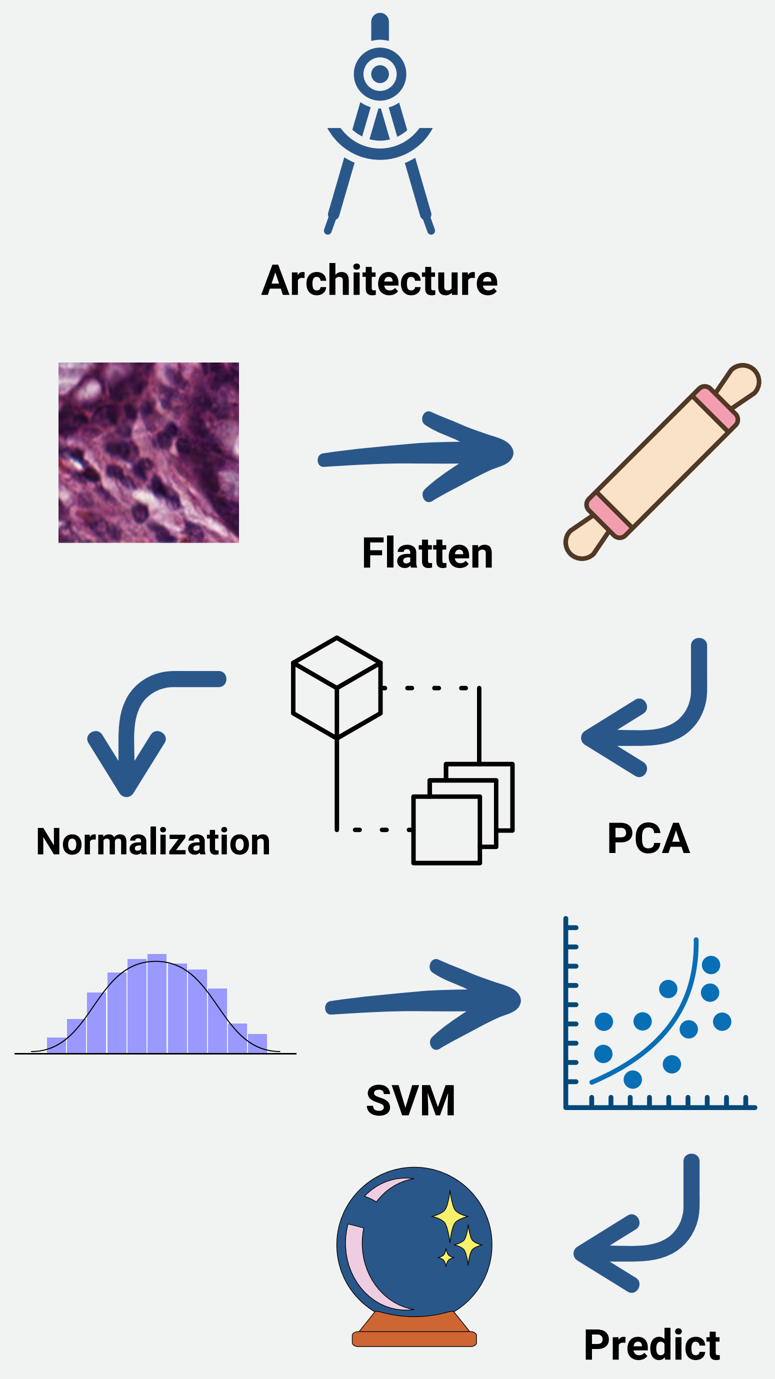
Test Accuracy → 0.525

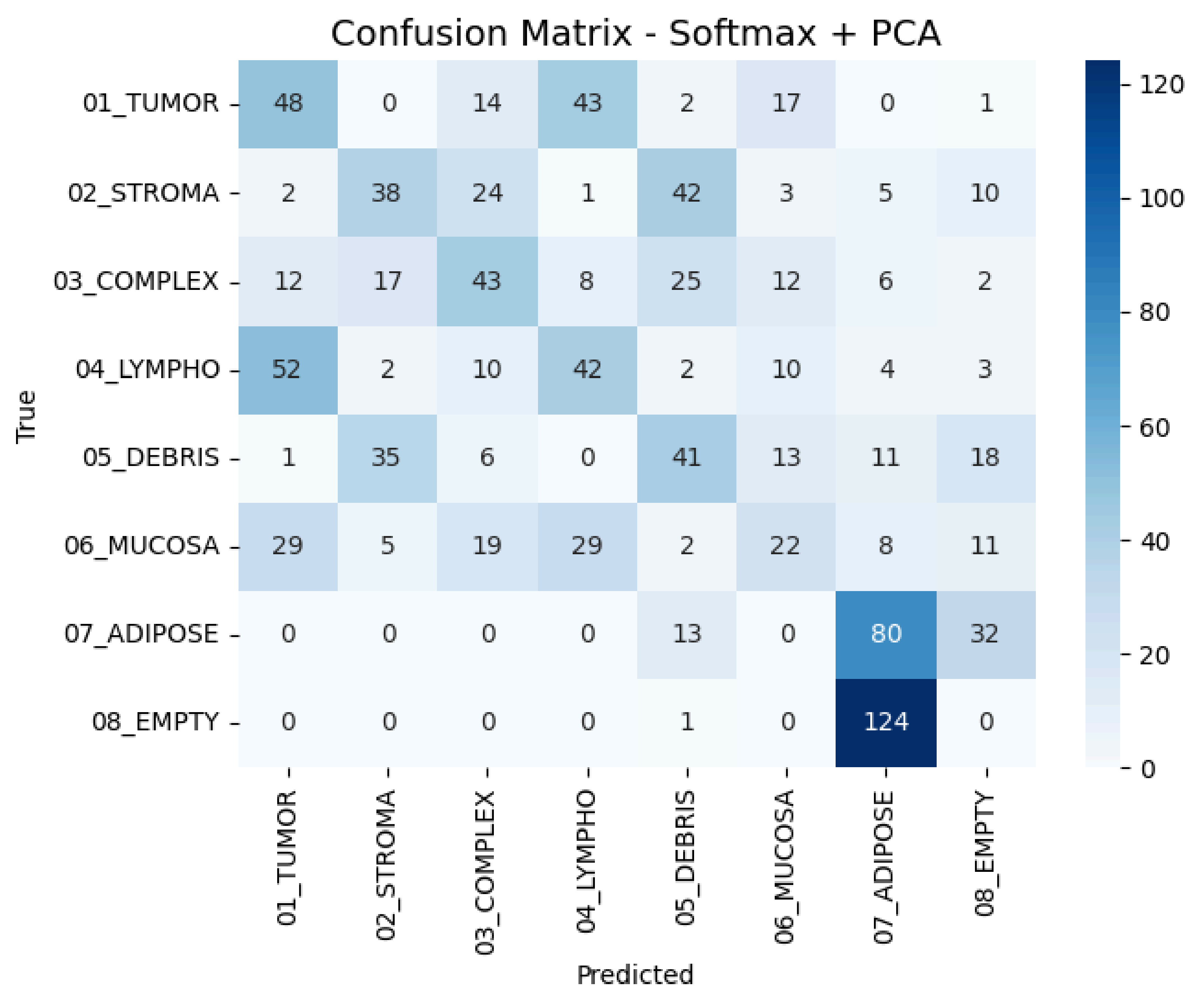
F1 Score → 0.511216

Overfitting Gap → 0.42550

FIFTH EXPERIMENT

SOFTMAX + PCA

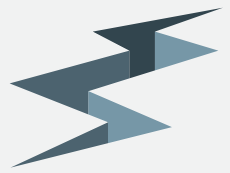




Accuracy → 0.70900

Test Accuracy → 0.314

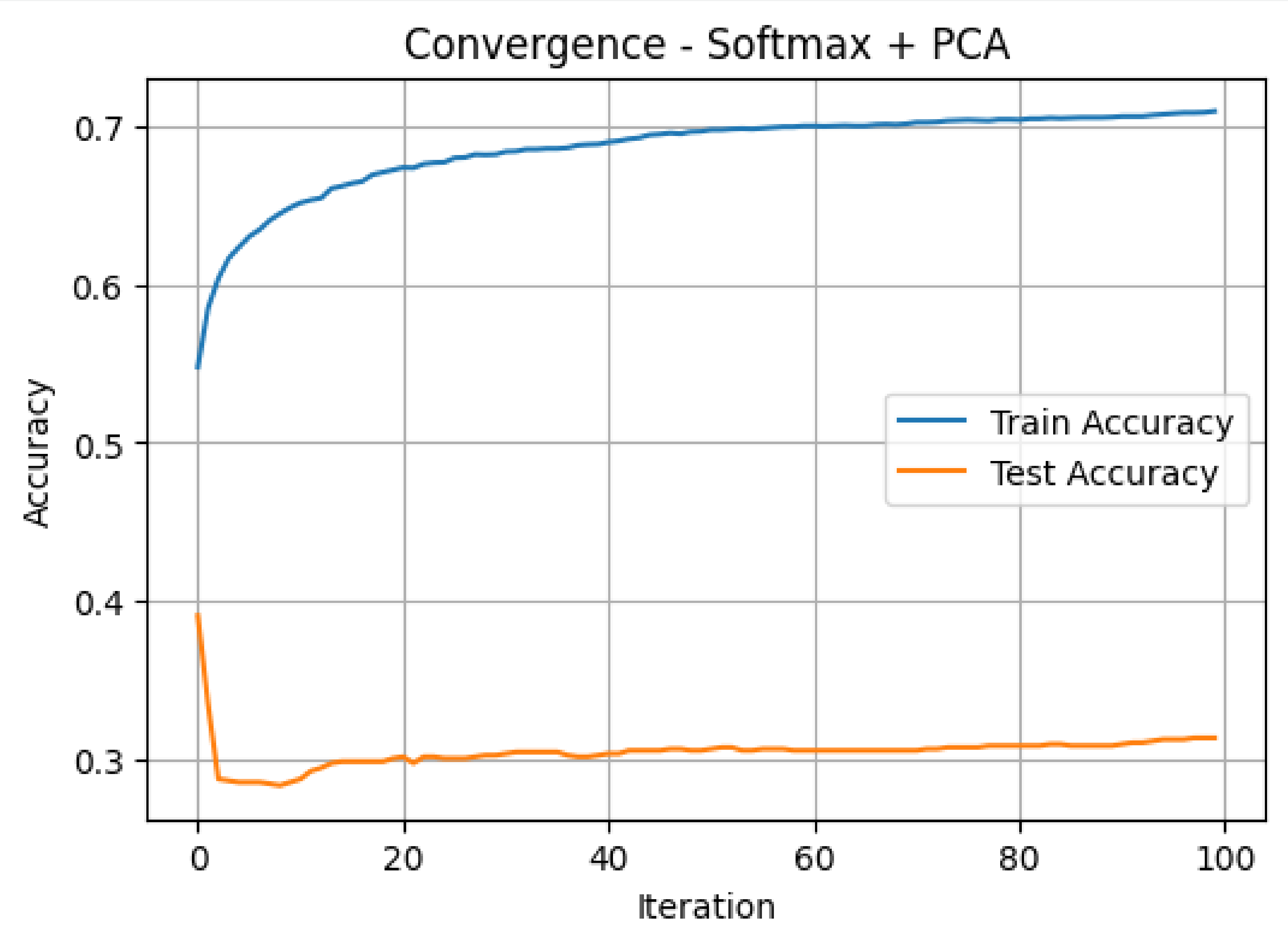
F1 Score → 0.297185



Overfitting Gap → 0.39500

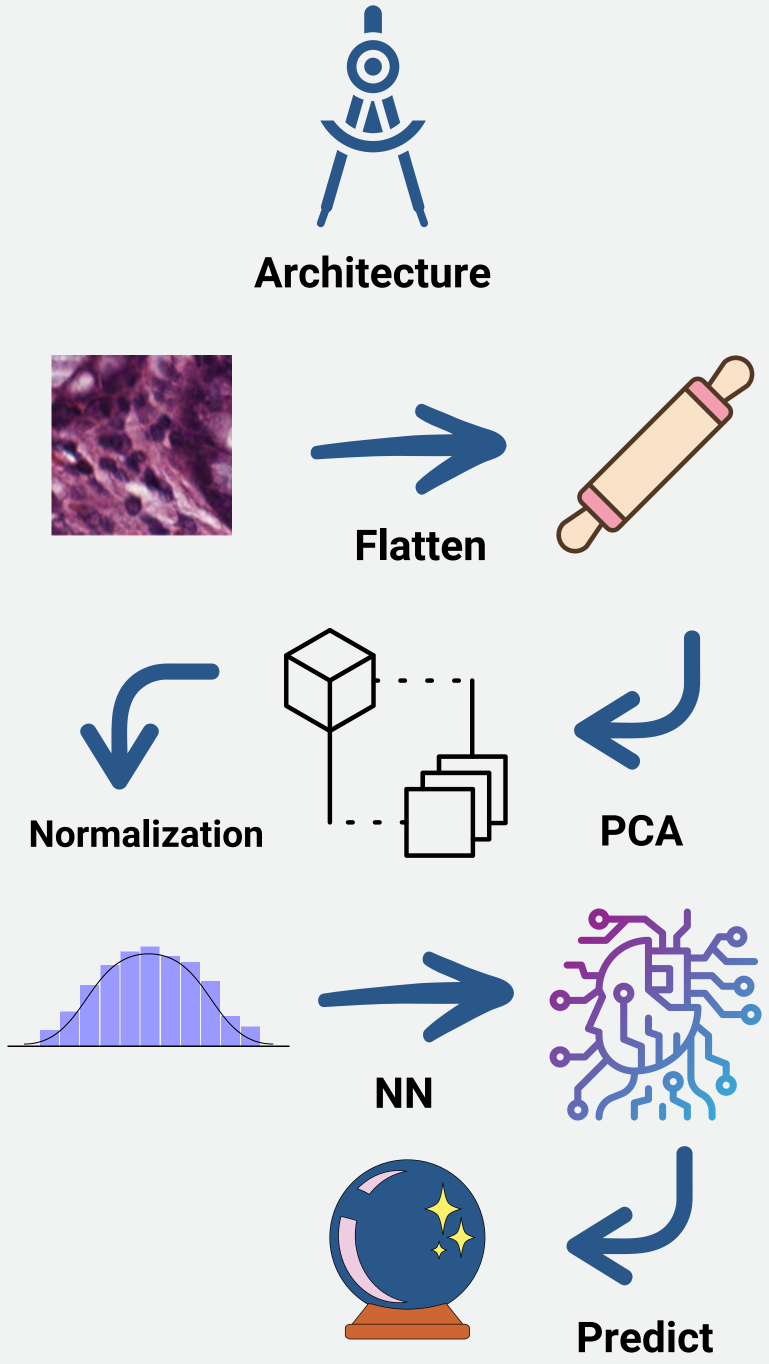
FIFTH EXPERIMENT

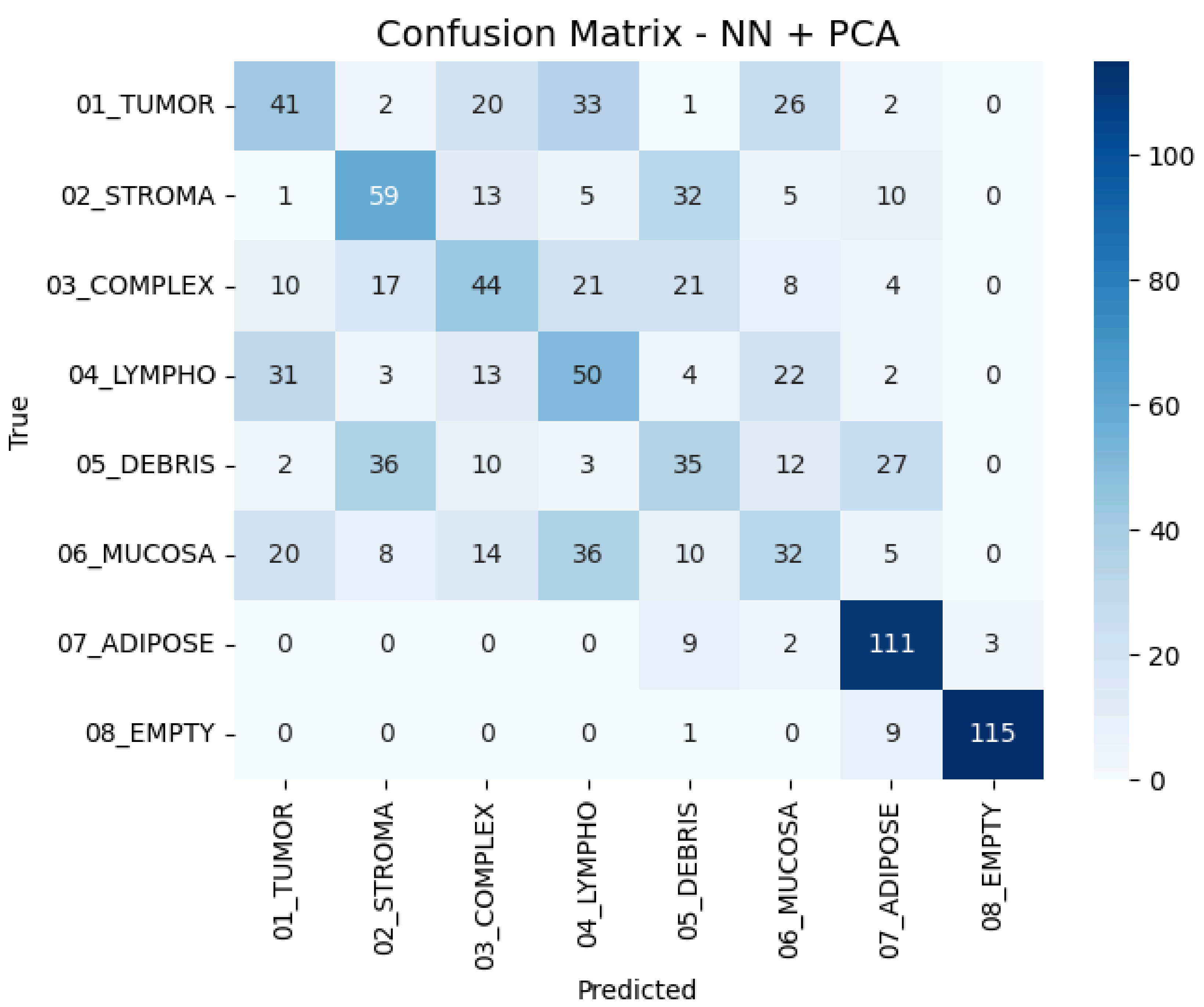
SOFTMAX + PCA



SIXTH EXPERIMENT

NN + PCA





Accuracy → 0.99850

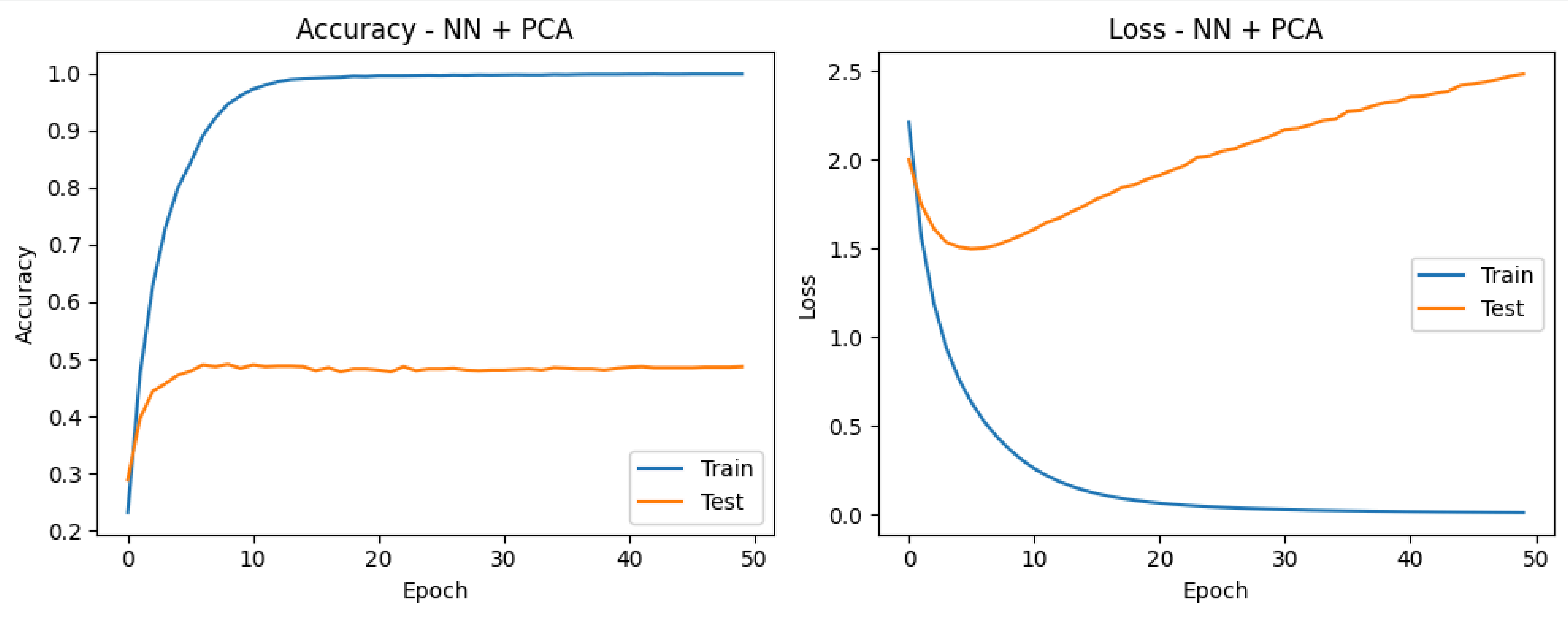
Test Accuracy → 0.487

F1 Score → 0.479006

Overfitting Gap → 0.51150

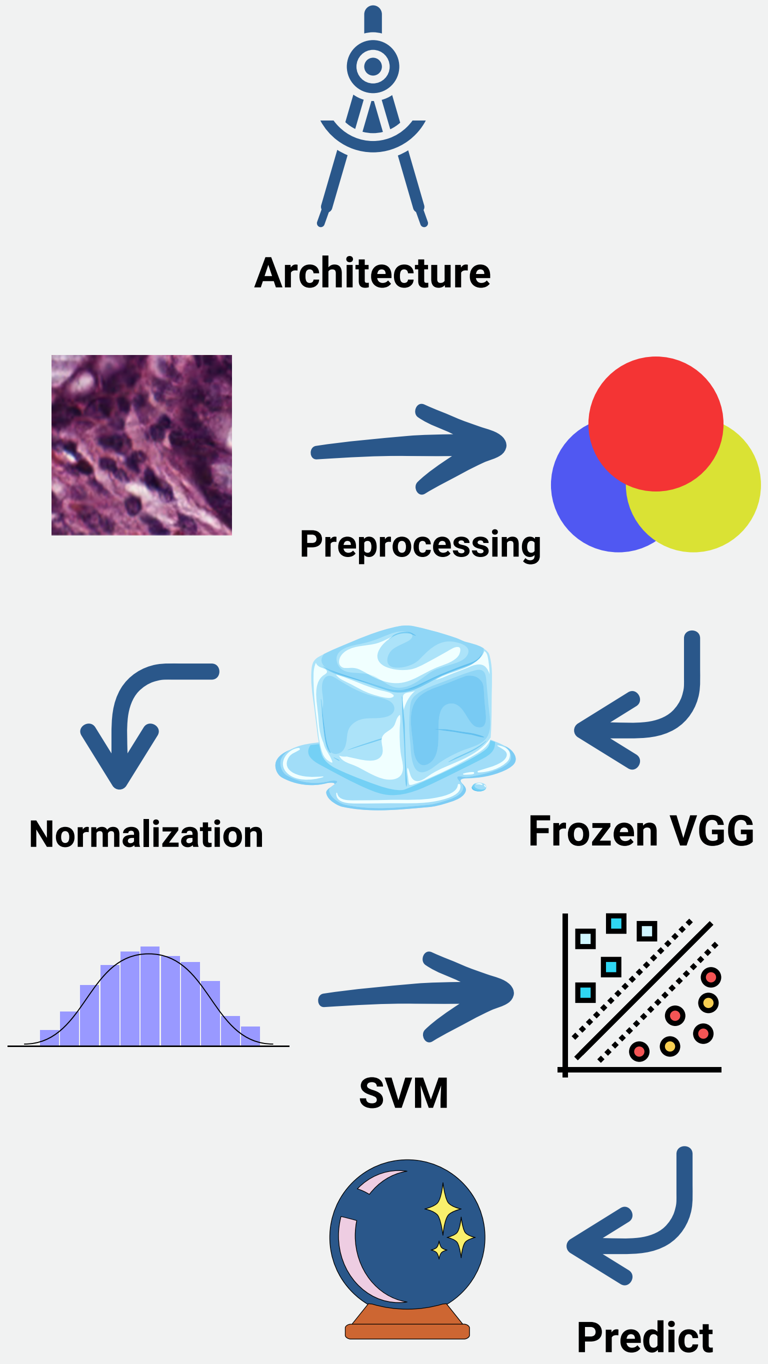
SIXTH EXPERIMENT

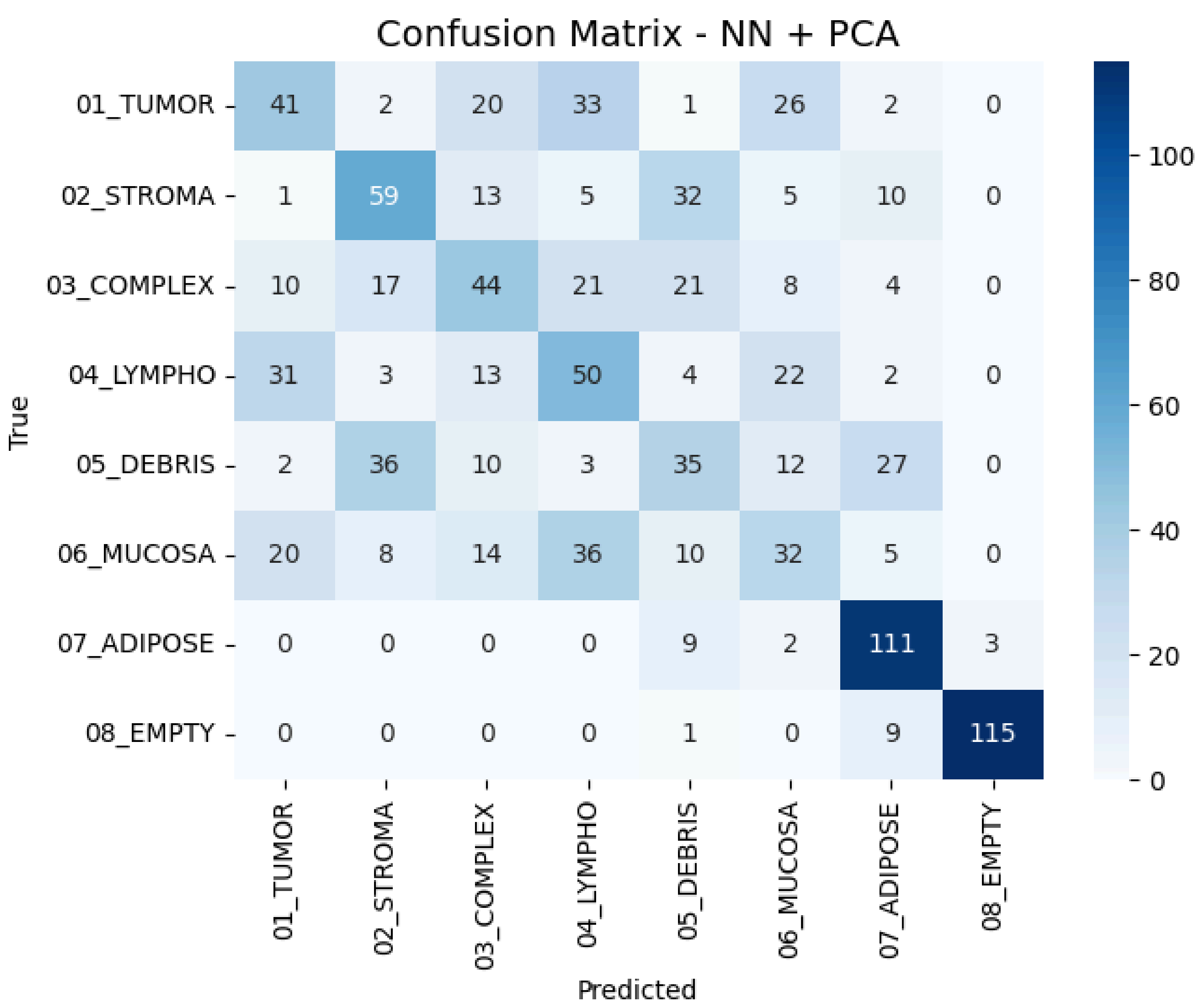
NN + PCA



SEVENTH EXPERIMENT

SVM + VGG16





True \ Predicted	01_TUMOR	02_STROMA	03_COMPLEX	04_LYMPHO	05_DEBRIS	06_MUCOSA	07_ADIPOSE	08_EMPTY
01_TUMOR	41	2	20	33	1	26	2	0
02_STROMA	1	59	13	5	32	5	10	0
03_COMPLEX	10	17	44	21	21	8	4	0
04_LYMPHO	31	3	13	50	4	22	2	0
05_DEBRIS	2	36	10	3	35	12	27	0
06_MUCOSA	20	8	14	36	10	32	5	0
07_ADIPOSE	0	0	0	0	9	2	111	3
08_EMPTY	0	0	0	0	1	0	9	115

Accuracy → 0.99850

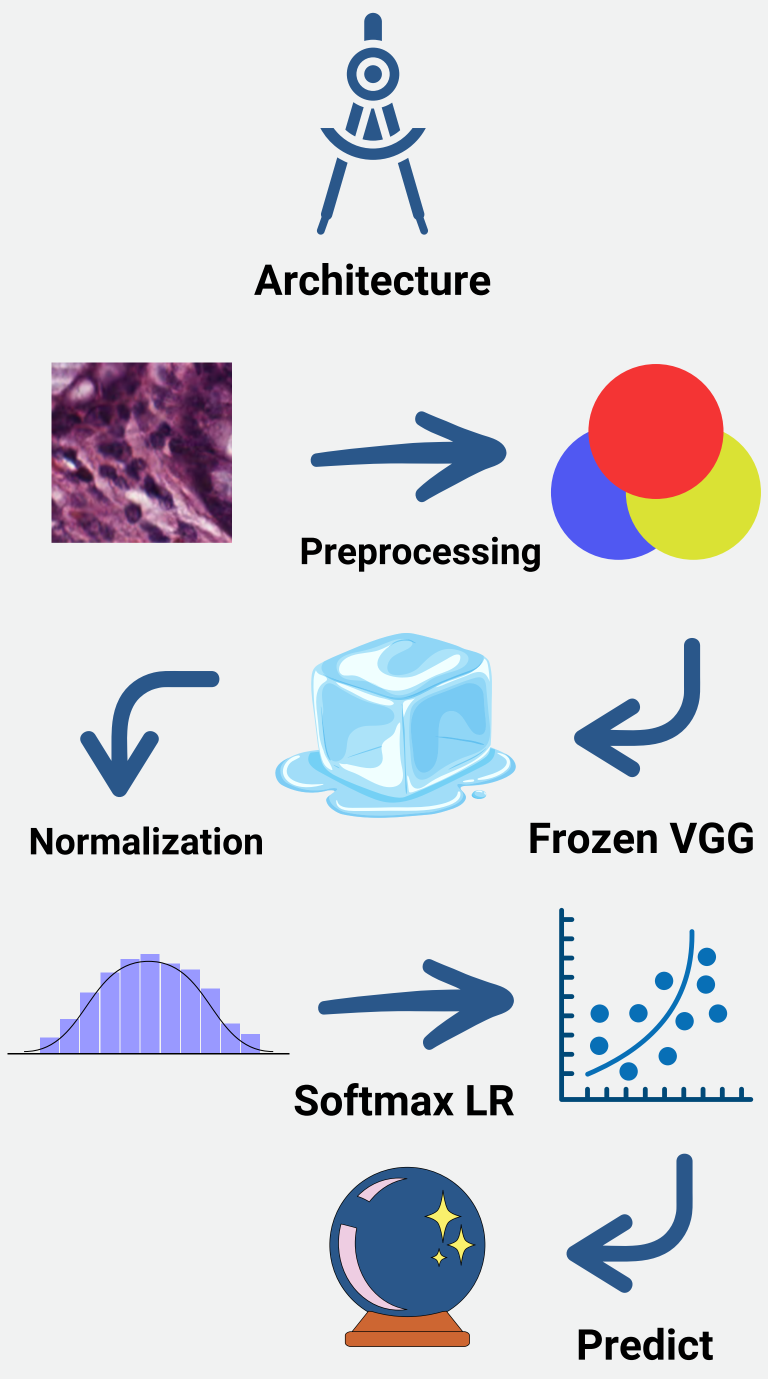
Test Accuracy → 0.487

F1 Score → 0.479006

Overfitting Gap → 0.51150

EIGHTH EXPERIMENT

SOFTMAX + VGG16



Confusion Matrix - Softmax + VGG16

True \ Predicted	01_TUMOR	02_STROMA	03_COMPLEX	04_LYMPHO	05_DEBRIS	06_MUCOSA	07_ADIPOSE	08_EMPTY
01_TUMOR	115	0	3	2	2	3	0	0
02_STROMA	2	91	16	2	14	0	0	0
03_COMPLEX	3	12	98	8	3	1	0	0
04_LYMPHO	0	3	16	103	3	0	0	0
05_DEBRIS	4	12	5	1	100	1	2	0
06_MUCOSA	4	0	1	4	3	113	0	0
07_ADIPOSE	0	0	0	0	0	0	121	4
08_EMPTY	0	0	0	0	0	0	1	124

Accuracy → 0.95150

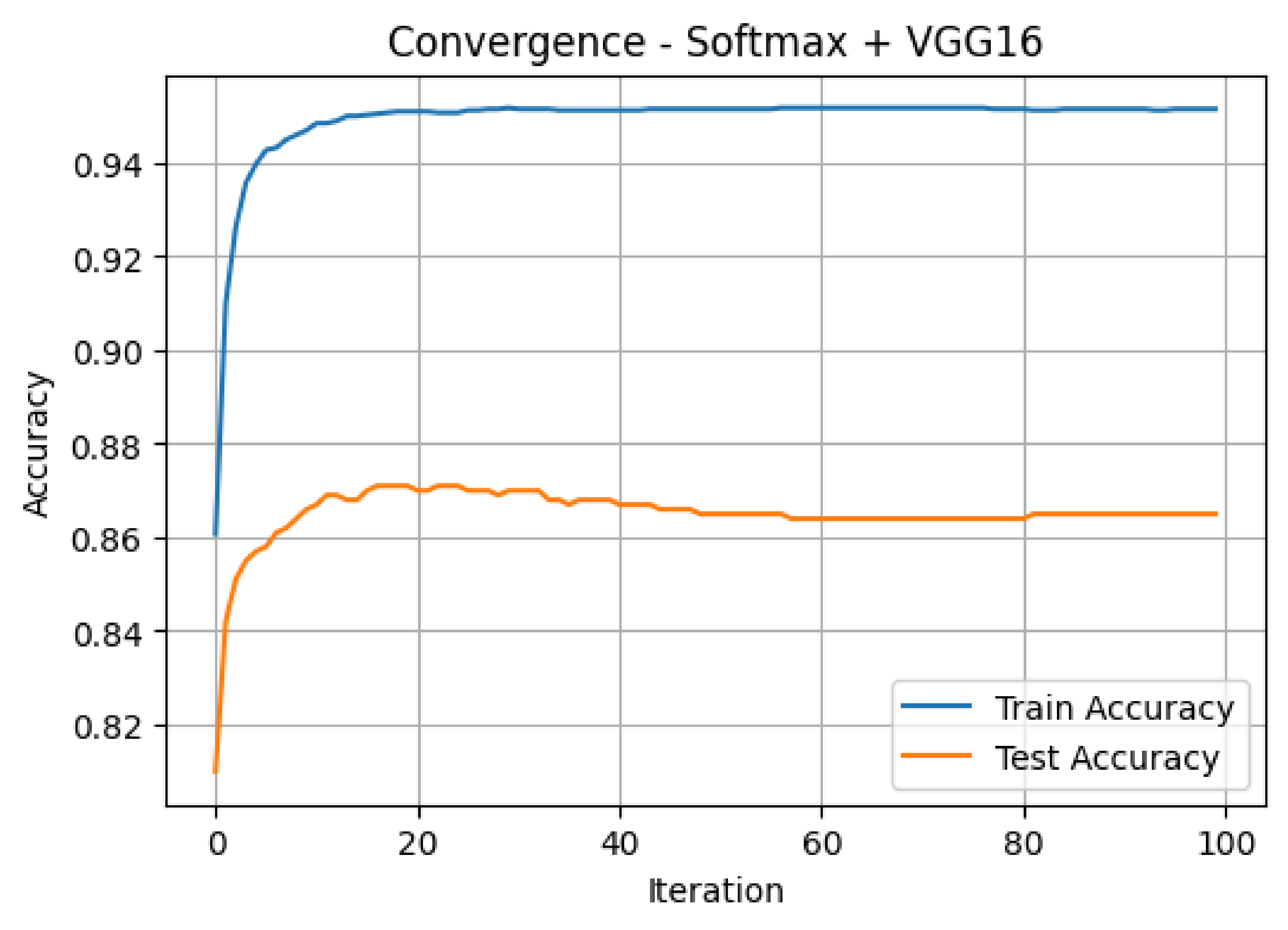
Test Accuracy → 0.865

F1 Score → 0.865434

Overfitting Gap → 0.08650

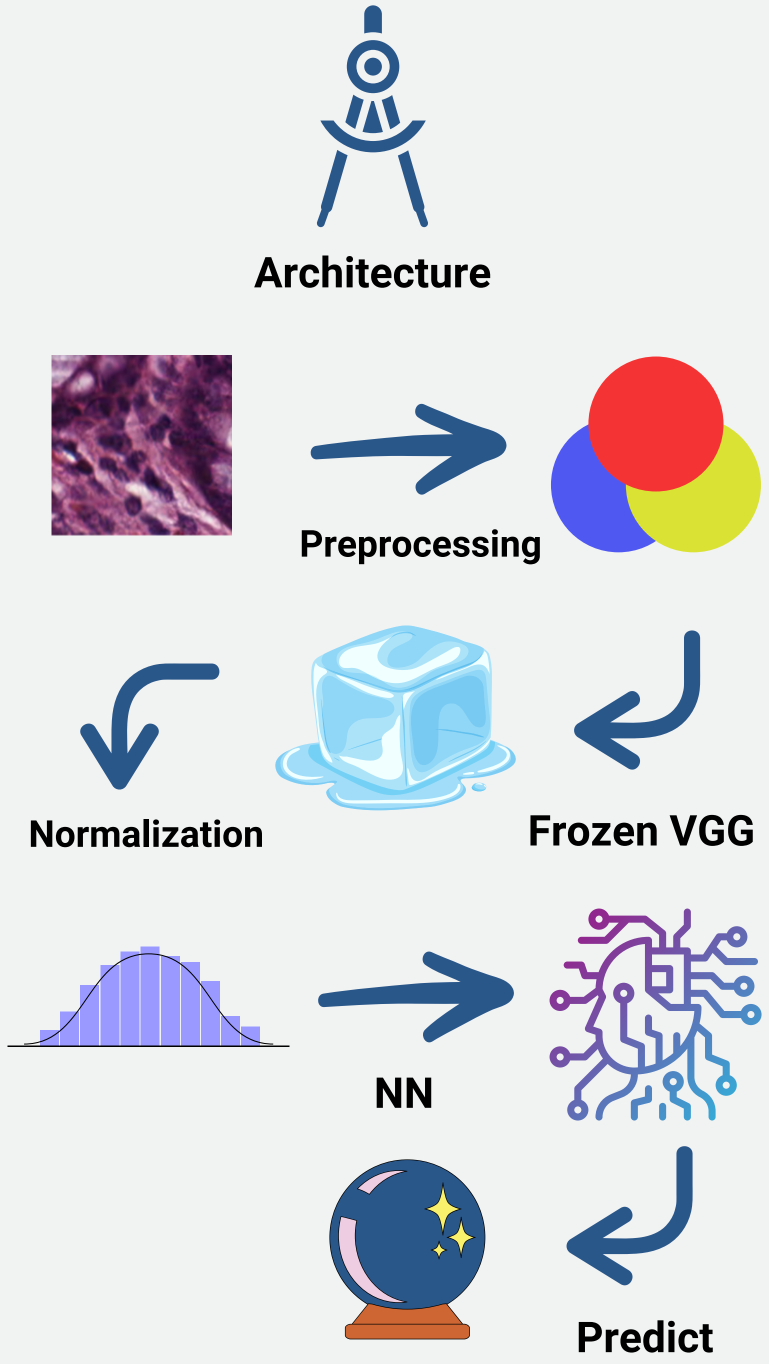
EIGHTH EXPERIMENT

SOFTMAX + VGG16



NINTH EXPERIMENT

NN + VGG16



Confusion Matrix - NN + VGG16

True \ Predicted	01_TUMOR	02_STROMA	03_COMPLEX	04_LYMPHO	05_DEBRIS	06_MUCOSA	07_ADIPOSE	08_EMPTY
01_TUMOR	114	0	4	2	2	3	0	0
02_STROMA	1	105	6	3	10	0	0	0
03_COMPLEX	2	18	92	9	2	2	0	0
04_LYMPHO	0	3	7	113	2	0	0	0
05_DEBRIS	3	13	5	0	101	1	0	2
06_MUCOSA	5	0	1	5	3	111	0	0
07_ADIPOSE	0	0	0	0	1	0	121	3
08_EMPTY	0	1	0	0	0	0	1	123

Accuracy → 0.99950

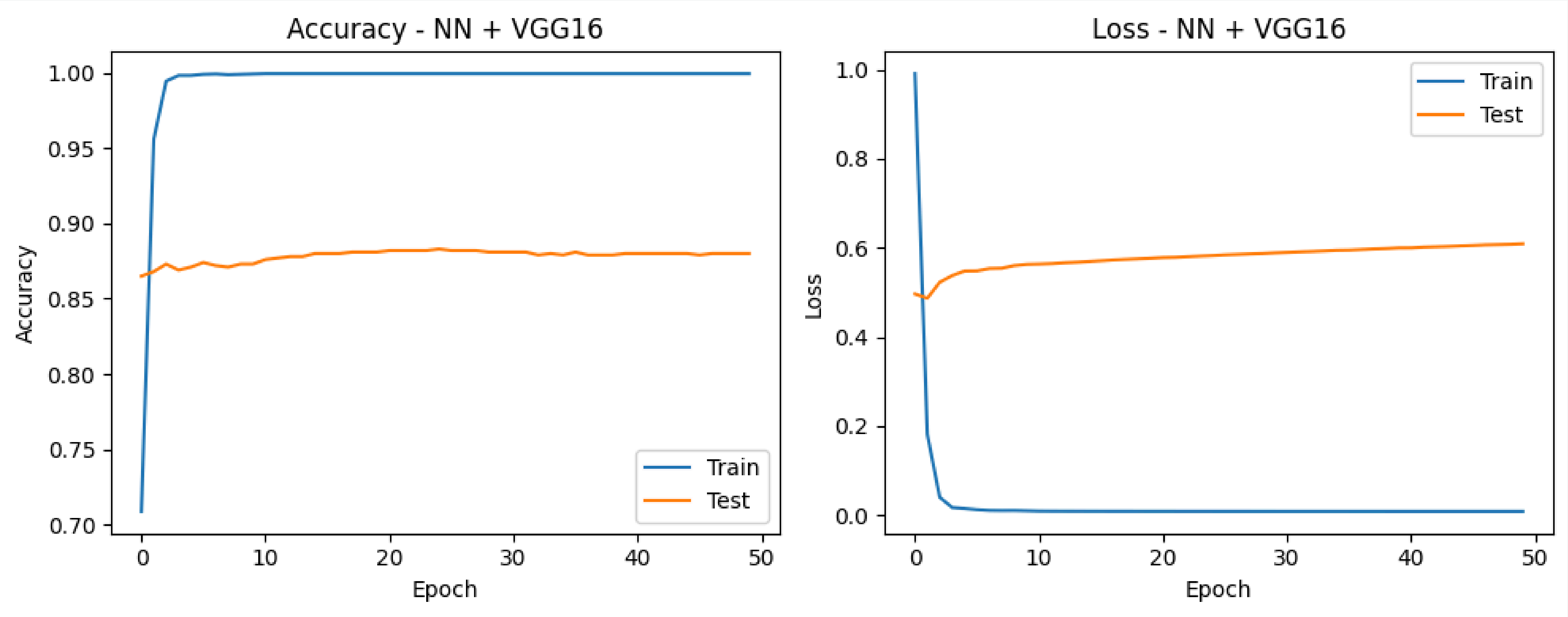
Test Accuracy → 0.880

F1 Score → 0.880135

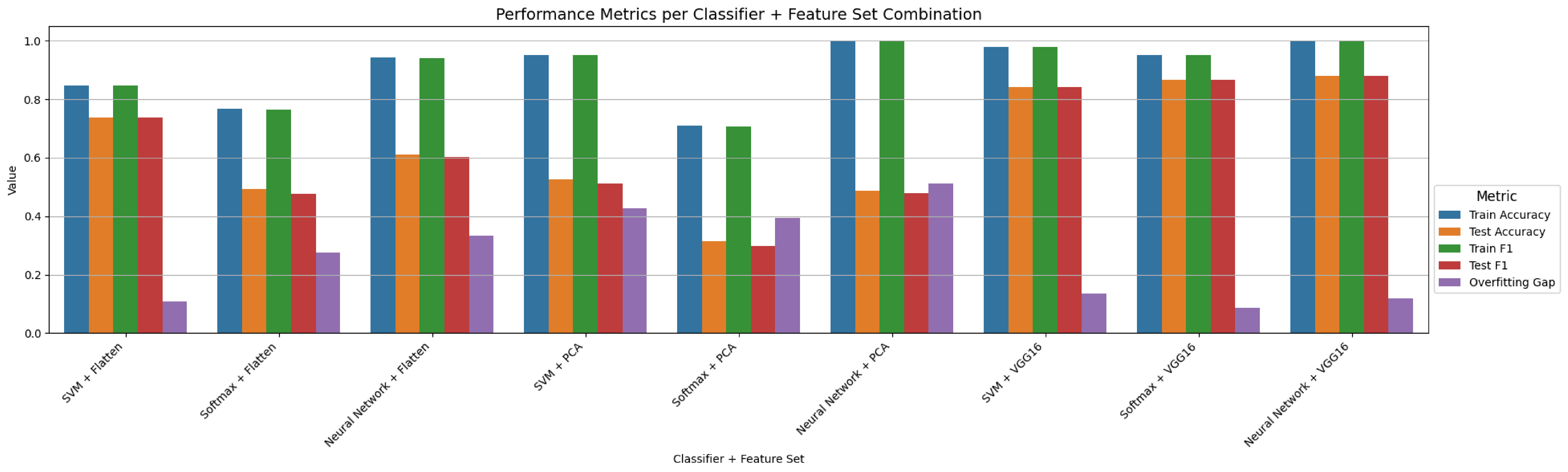
Overfitting Gap → 0.11950

NINTH EXPERIMENT

NN + VGG16



EXPERIMENTS 1-9 SUMMARY



Obstacles Along The Road

01

Classes Stroma & Complex

Were the most difficult to classify with the highest confusion rate between them.

02

Pin Point Fine-Tuning

Finding the “sweet spot” of each hyperparameter to maximize model performence.

03

Overfitting and Unstable Training

With more advances techniques - the more the model could memorize the training.

04

Balancing Complexity

Finding the middle ground between too complex and too simplified.

05

Inconsistent Data Splitting

Data leakage danger since the files upload in a random order. Leading to different train-test split each session

06

The “Other”s

Even though the main problem was classes 2&3 - other classes added to the already difficult task

The Solutions

Hierarchy Modelling

Separating the task to two models - a general classification model and an expert model.

Optuna

Using the optuna library to help us effectively identify the best methods and hyperparameters

Regularization & Augmentation

Using advanced augmentation with albumentations, while using early stopping and ReduceLROnPlateau

Balancing Complexity

The general model was trained on a simple augmentation while the expert learned through difficult patterns.

Permanent Loading

Sorting the dataframe by path and random_state = 42.

Confusion Class Grouping

Training the expert model to also classify a class of “Other” containing the most confusing classes.

ON THE WAY TO THE BEST F1

All the trials and mostly failures

STEP 01

Multiple Optunas

As much as 5 optuna trials to try and find the best strategy.

STEP 02

Augmentations

Different, advanced augmentations to each “difficult” class.

STEP 03

Ensemble Models

Gathering an ensemble of models to try and tackle the task together.

STEP 04

Strategy Pivot

A more focused approach trying to improve as much as possible the Stroma and Complex predictions.

STEP 05

Attention Mechanism

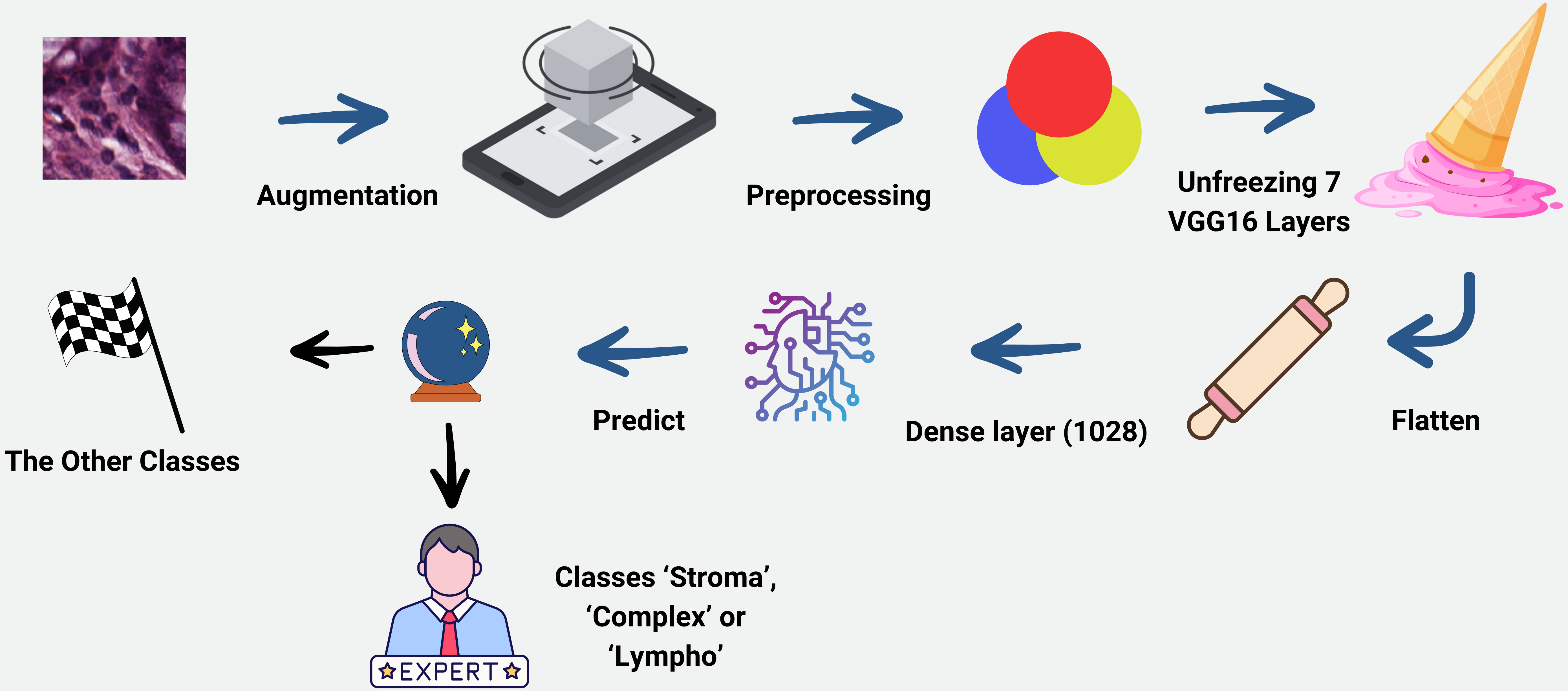
Integrating an attention mechanism to force the model to focus on the most subtle features.

STEP 06

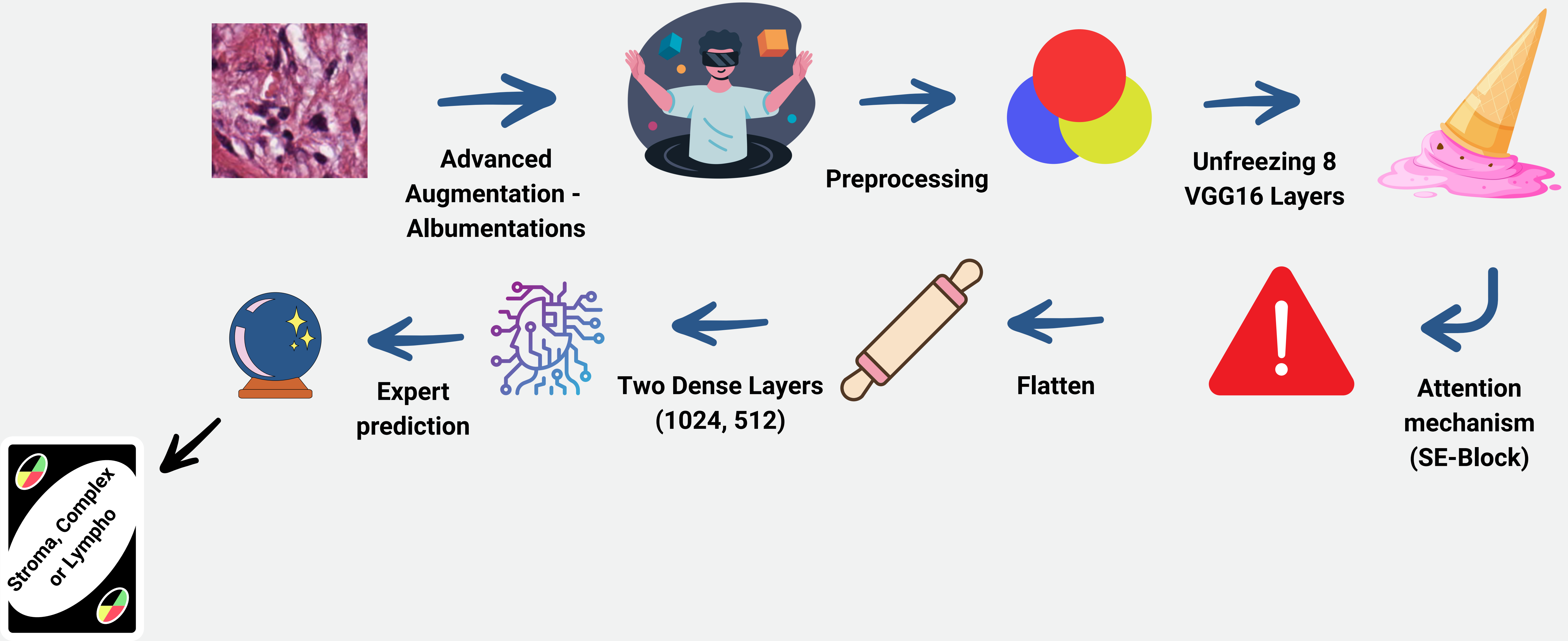
Hierarchy Modelling

Training a Stroma-Complex-Lympho expert and applying it to boost the best regular model found in the optunas.

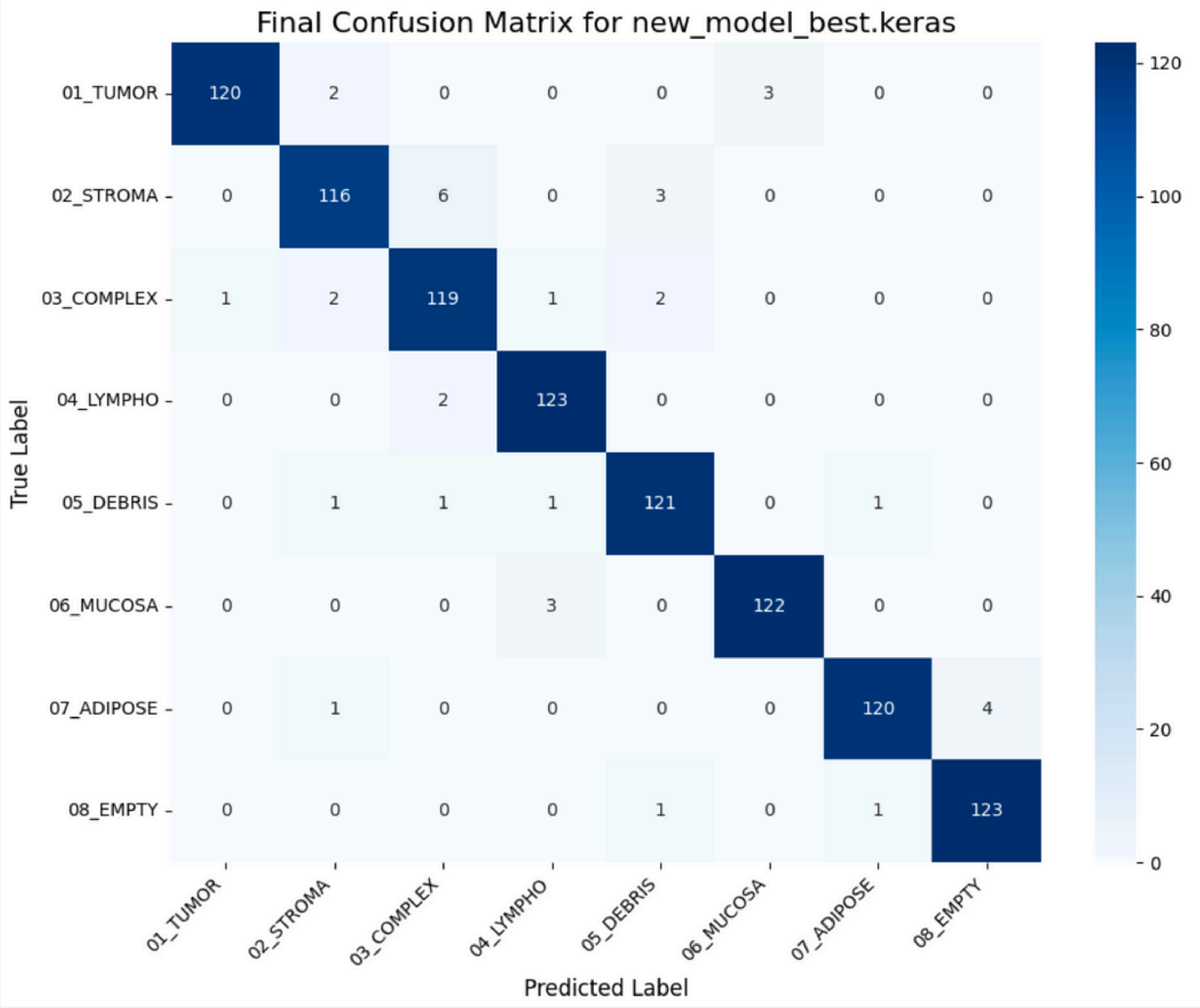
BEST MODEL - ARCHITECTURE



EXPERT MODEL - ARCHITECTURE



RESULTS

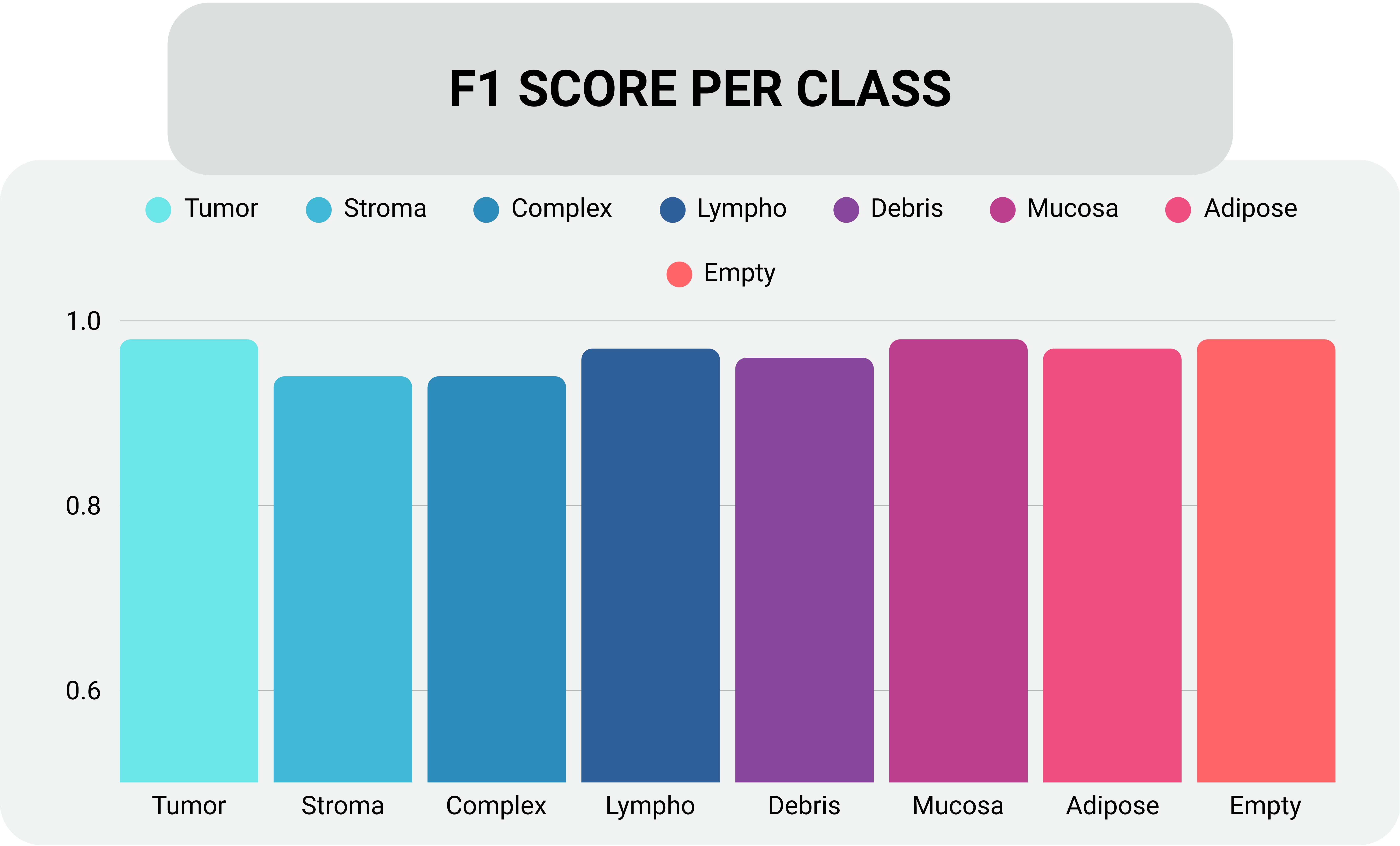


Original F1 (no specialist) → **0.9522**

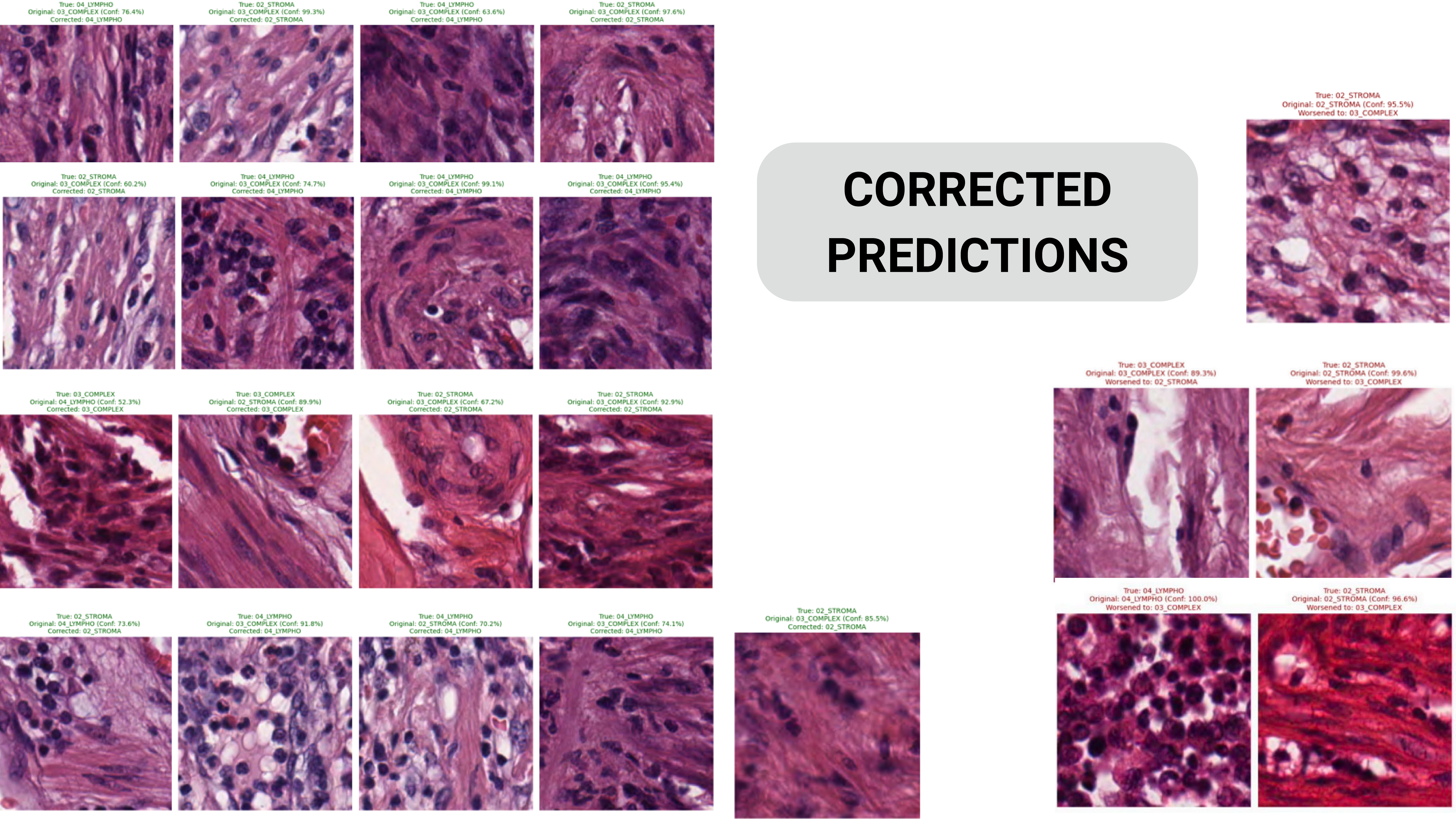
Final F1 (with specialist) → **0.9640**

Avg. confidence of main model in the corrected predictions → **0.8023**

Changes made by Specialist → **23**



F1 SCORE PER CLASS
Tumor
Stroma
Complex
Lympho
Debris
Mucosa
Adipose
Empty
1.0
0.8
0.6
Tumor
Stroma
Complex
Lympho
Debris
Mucosa
Adipose
Empty


True: 04_LYMPHO
Original: 03_COMPLEX (Conf: 76.4%)
Corrected: 04_LYMPHO
True: 02_STROMA
Original: 03_COMPLEX (Conf: 99.3%)
Corrected: 02_STROMA
True: 04_LYMPHO
Original: 03_COMPLEX (Conf: 63.6%)
Corrected: 04_LYMPHO
True: 02_STROMA
Original: 03_COMPLEX (Conf: 97.6%)
Corrected: 02_STROMA
True: 02_STROMA
Original: 03_COMPLEX (Conf: 60.2%)
Corrected: 02_STROMA
True: 04_LYMPHO
Original: 03_COMPLEX (Conf: 74.7%)
Corrected: 04_LYMPHO
True: 04_LYMPHO
Original: 03_COMPLEX (Conf: 99.1%)
Corrected: 04_LYMPHO
True: 04_LYMPHO
Original: 03_COMPLEX (Conf: 95.4%)
Corrected: 04_LYMPHO
True: 03_COMPLEX
Original: 04_LYMPHO (Conf: 52.3%)
Corrected: 03_COMPLEX
True: 03_COMPLEX
Original: 02_STROMA (Conf: 89.9%)
Corrected: 03_COMPLEX
True: 02_STROMA
Original: 03_COMPLEX (Conf: 67.2%)
Corrected: 02_STROMA
True: 02_STROMA
Original: 03_COMPLEX (Conf: 92.9%)
Corrected: 02_STROMA
True: 02_STROMA
Original: 04_LYMPHO (Conf: 73.6%)
Corrected: 02_STROMA
True: 04_LYMPHO
Original: 03_COMPLEX (Conf: 91.8%)
Corrected: 04_LYMPHO
True: 04_LYMPHO
Original: 02_STROMA (Conf: 70.2%)
Corrected: 04_LYMPHO
True: 04_LYMPHO
Original: 03_COMPLEX (Conf: 74.1%)
Corrected: 04_LYMPHO
CORRECTED PREDICTIONS
True: 02_STROMA
Original: 03_COMPLEX (Conf: 85.5%)
Corrected: 02_STROMA
True: 02_STROMA
Original: 02_STROMA (Conf: 95.5%)
Worsened to: 03_COMPLEX
True: 03_COMPLEX
Original: 03_COMPLEX (Conf: 89.3%)
Worsened to: 02_STROMA
True: 02_STROMA
Original: 02_STROMA (Conf: 99.6%)
Worsened to: 03_COMPLEX
True: 04_LYMPHO
Original: 04_LYMPHO (Conf: 100.0%)
Worsened to: 03_COMPLEX
True: 02_STROMA
Original: 02_STROMA (Conf: 96.6%)
Worsened to: 03_COMPLEX

CONCLUSION

Among all the advanced techniques tested, the combination of advanced augmentations, attention mechanisms, and training a specialized model provided a major performance boost compared to the initial results.

Hyperparameter tuning with Optuna helped us focus on the most effective configurations for the more advanced experiments.

A hierarchical combination of models (main + specialist) improved classification in ambiguous cases and reduced confusion between the most similar classes.

IDEAS FOR THE FUTURE

Implementing a more advanced uncertainty estimation method for the main model. In our experiments, we mainly used a confidence threshold based on softmax probability, which doesn't always reflect true model certainty. Techniques like Monte Carlo Dropout could offer a deeper understanding of model confidence.

Further improving the model and evaluating its potential for real-world deployment, particularly in clinical decision-making processes.

Exploring alternative architectures beyond VGG and simple neural networks to push the model's accuracy even further.

THANK YOU