
Coffee
bean quality prediction system and
leaf disease classification.
Dubai UG 2

R2-Data Analysis and Exploration

Preprocessing

Feature Engineering and Selection:

·One-Hot Encoding:
Categorical features such as Country of Origin, Processing Method, and Color were transformed into binary vectors

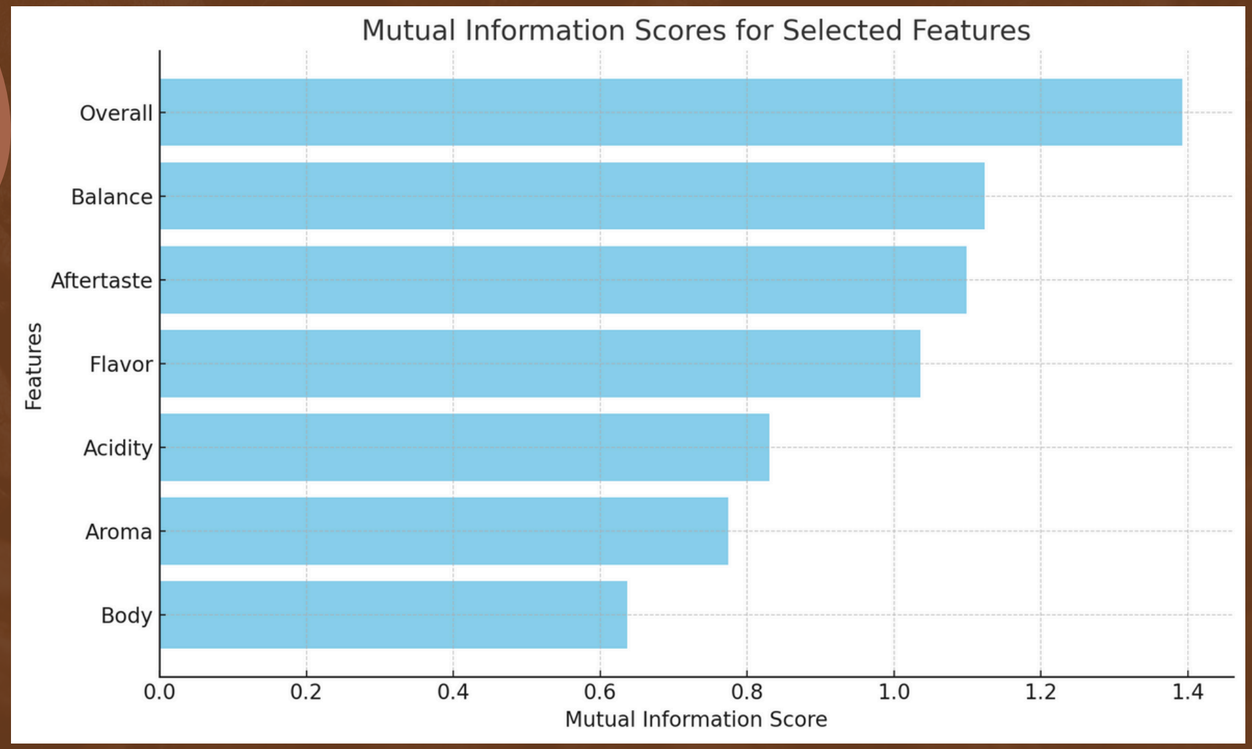
·Standard Scaling:
Numeric features, including Aroma, Flavor, and Total Cup Points, were scaled using StandardScaler to normalize values and prevent dominance by larger-scale features.

·Mutual Information Scores:
Calculated to evaluate feature relevance.

·Categorical Changes:
Mapped Country of origin to Continents, Categorize processing methods & Colors

Image Dataset:

- Normalizing Pixel Values
- Resizing to 128*128



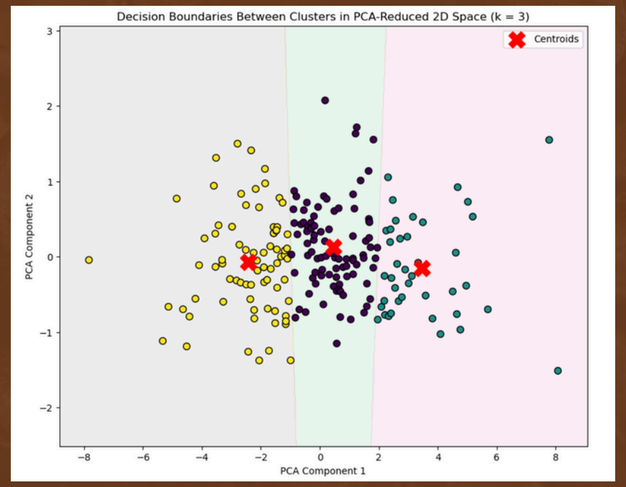
R3- Clustering

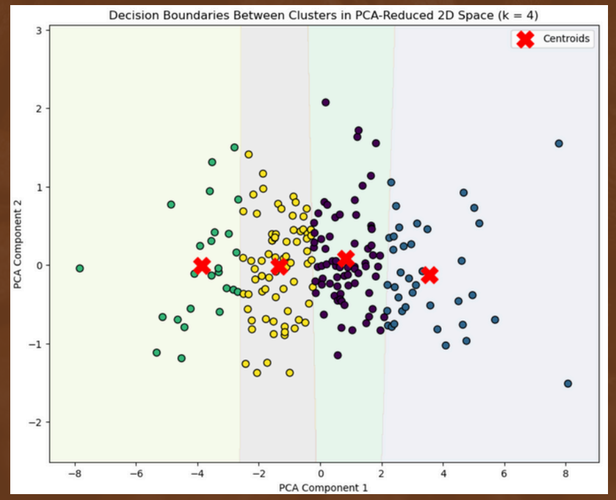
Preprocessing

- Selected features: Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Overall.
- StandardScaler applied for normalization.

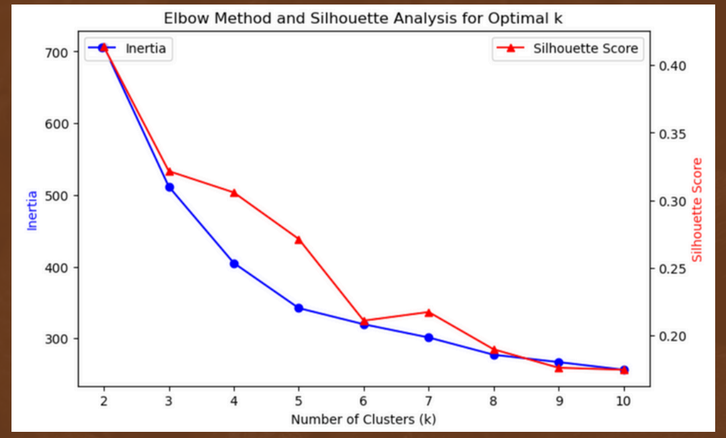
Methods Used

- Silhouette & Elbow Method: Identified optimal clusters (k=3).
- KMeans Clustering: Visualized clusters using PCA.
- Gaussian Mixture Model (GMM): Soft clustering comparison.
- Visualization techniques t-SNE, and UMAP

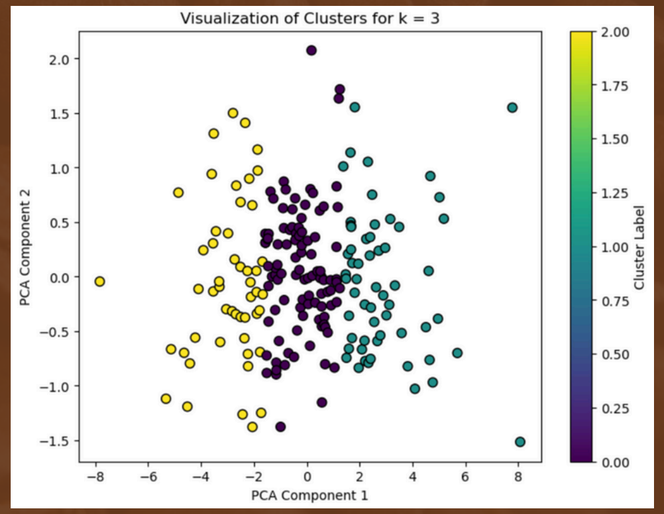




Decision boundary visualization (k=3 and k=4).

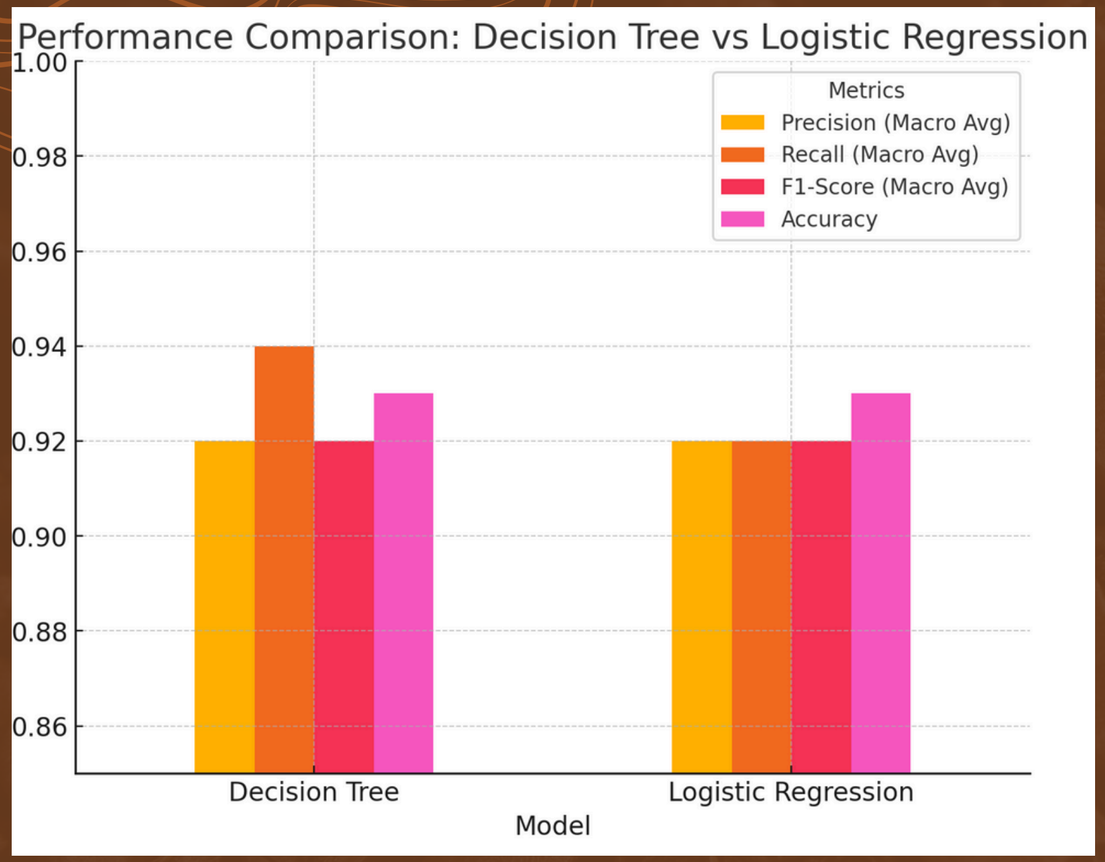


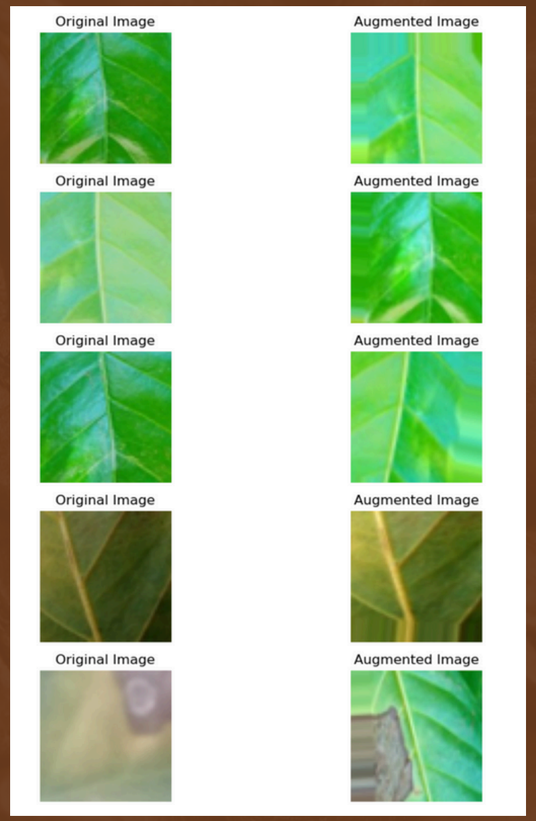
Silhouette scores for tested cluster sizes.

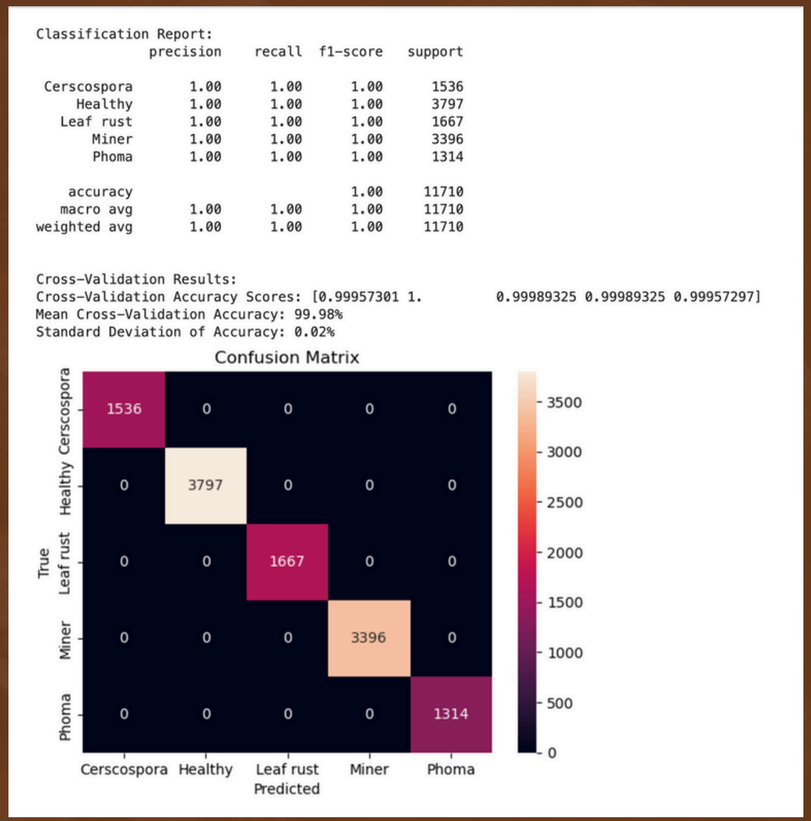


PCA plot for k=3 clusters.

R4-Baseline Training and Evaluation Experiments







Decision Trees- Used to classify coffee quality into Low, Moderate, and High categories

Max Depth, Min Samples Split, Min Samples Leaf, Max Leaf Nodes, Test Size,-were the Hyperparameters we experimented with.

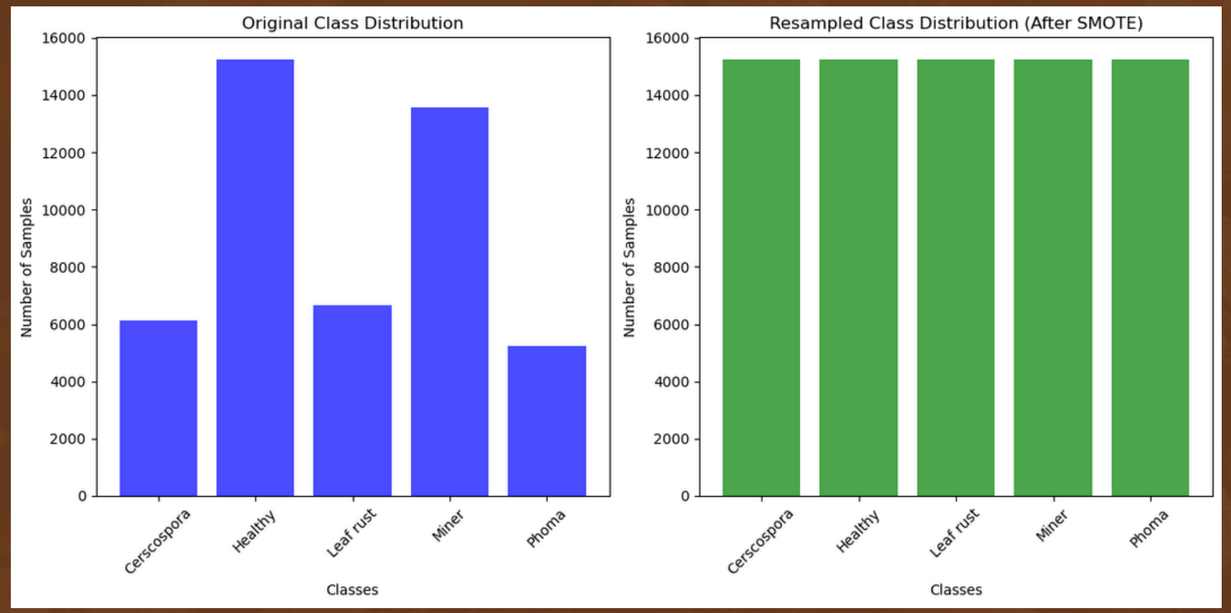
Multinomial logistic regression was implemented by splitting total cup points into 3 categories based on percentiles. The categories were encoded to integers and the model was trained on six features.

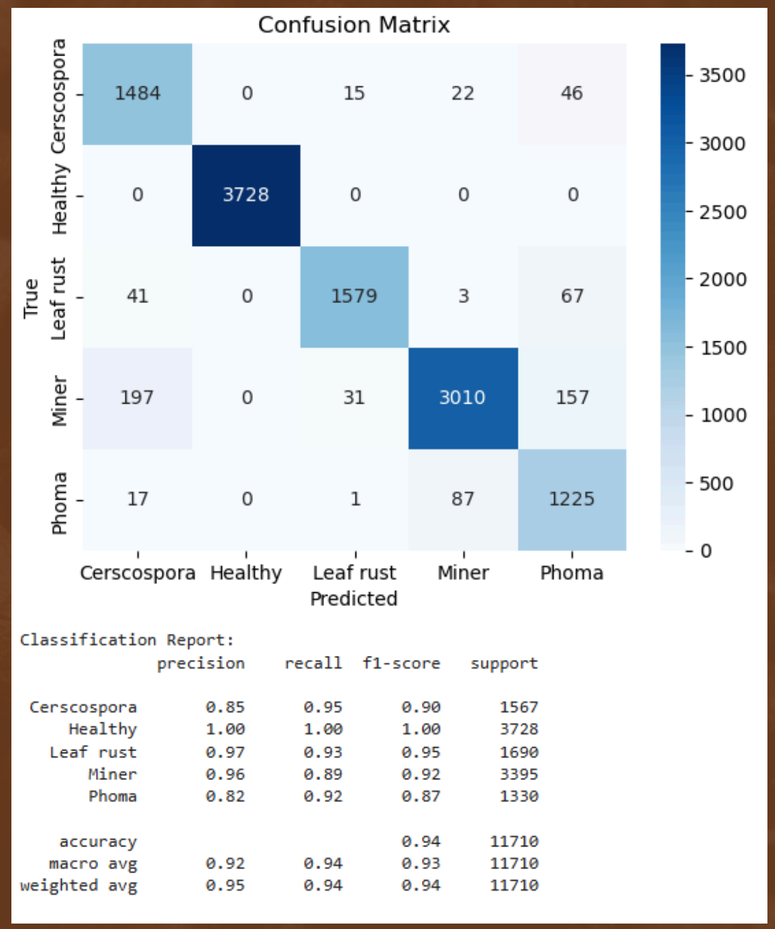
KNN was applied to the image dataset image classifcation and also for comparison of a simpler model with more complex models in R5,

HyperParameter = K Tuned using gridsearch to find best K=1

Data augmentation and data imbalance checks were also carried out to check the model against overfitting

R5-Neural Networks
Multi-Layer Perceptron (MLP)





Classification Report:

	precision	recall	f1-score	support
Cerscospora	0.85	0.95	0.90	1567
Healthy	1.00	1.00	1.00	3728
Leaf rust	0.97	0.93	0.95	1690
Miner	0.96	0.89	0.92	3395
Phoma	0.82	0.92	0.87	1330
accuracy			0.94	11710
macro avg	0.92	0.94	0.93	11710
weighted avg	0.95	0.94	0.94	11710

To implement an MLP on the image dataset, the following steps were taken:

1. Images were resized to 64x64 for simplicity.
2. SMOTE was applied to address the imbalance in class sizes.
3. Data augmentation was carried out on the train and validation set.

The model flattens input images into vector. It is then passed to two separate layers with ReLU activation. The output layer uses the Softmax activation.

The model uses Stochastic Gradient Descent Optimizer and after 5 epochs, the learning rate is reduced by 10% each epoch. This is done so that the model doesnt overestimate.

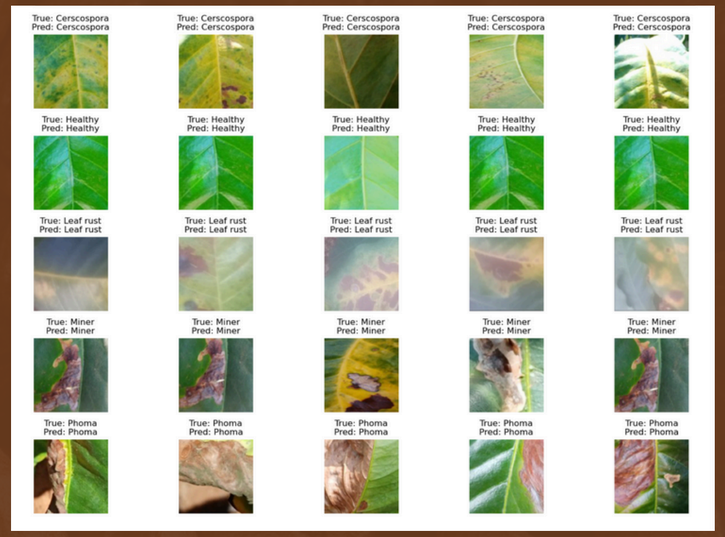
R5-Neural Networks
Convolutional Neural Network (CNN) Implementation

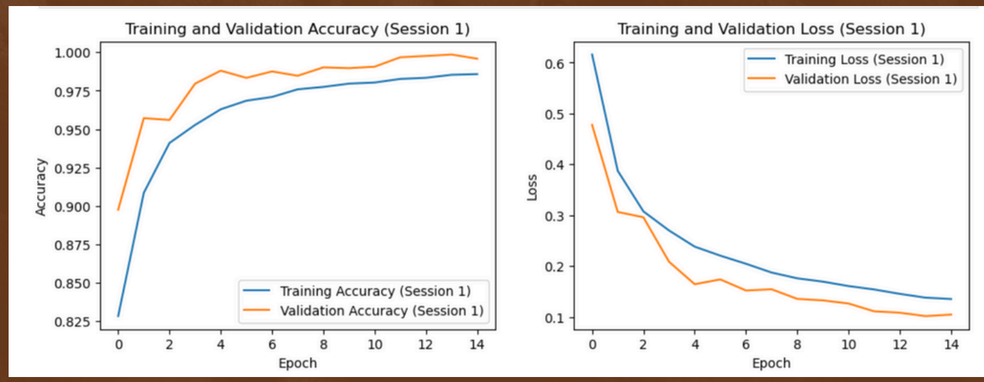
Key Steps:
- Data Augmentation:
  - Used ImageDataGenerator for data augmentation (e.g., rotation, flipping) to enhance generalization.
- Label Encoding and Dataset Splitting:
  - Labels encoded numerically; dataset split into 80% training and 20% testing with stratified sampling.

Model Architecture:
- Convolutional Layers:
  - Two layers (32 and 64 filters, 3x3 kernel, ReLU activation) for feature extraction.
- MaxPooling Layers:
  - 2x2 kernel to reduce spatial dimensions.
- Fully Connected Layers:
  - Dense layers with dropout (rate = 0.5) for classification.
- Softmax Activation:
  - For multi-class classification.

Training and Optimization:
- Adam optimizer with categorical cross-entropy loss.
- ReduceLROnPlateau callback for dynamic learning rate adjustments.

Performance Evaluation:
- Metrics: Accuracy, Precision, Recall, and F1-scores.
- Tools: Confusion Matrix and Classification Reports.





Aspect	KNN	CNN	MLP
Learning	Lazy, memorizes.	Spatial features.	General patterns.
Best For	Small datasets.	Image tasks.	Tabular data.
Strength	Simple, fast setup.	High image accuracy.	Versatile.
Weakness	Slow inference.	High computation.	Lacks spatial focus.