final_project_expert_annotations

PBMC Single-Cell Classification: Expert Annotation Approach¶

Bonus exercise¶

Author: Kristof Torkenczy
Course: Tech 27 Machine Learning
Approach: Expert-curated SeuratData annotations

Project Overview¶

This analysis leverages **expertly curated cell type annotations** from the SeuratData package to perform high-accuracy PBMC (Peripheral Blood Mononuclear Cell) classification. Rather than relying on algorithmic marker-based approaches, this pipeline uses professional annotations validated by the single-cell genomics community.

Expert Annotation Approach¶

Method: Professional curation by Seurat team and community

- **Source**: SeuratData package with peer-reviewed annotations
- **Quality**: Multi-study consensus and expert validation
- **Detail**: 9-20 detailed cell subtypes vs marker
- **Standardization**: Consistent nomenclature across studies

Key Features:

- **Expert-curated training data**: Pre-validated cell type annotations
- **Comprehensive ML evaluation**: 11 different machine learning algorithms
- **High-quality datasets**: pbmc3k (training) + pbmcMultiome (testing)
- **Detailed analysis**: Complete downstream analysis with high-resolution figures
- **Clinical relevance**: Disease-relevant cell subtypes and biomarkers

Datasets:

- **pbmc3k**: Training dataset with expert annotations (2,700 cells, 9 cell types)
- **pbmcMultiome_full**: Test dataset with expert annotations (11,909 cells, 20 cell types)
- **Technology**: 10x Chromium (v1 and Multiome) with cross-platform validation
- **Source**: SeuratData package with expert validation

---

1. Data Download and Organization¶

Download and organize SeuratData expert-annotated PBMC datasets.

In [1]:

```
# Download annotated (SeuratData) data
!python download_data_unified.py --approach annotated --quiet
```

2. Exploratory Data Analysis with Expert Annotations¶

Perform comprehensive EDA on expert-annotated datasets without additional annotation steps.

In [2]:

```
# Run EDA with expert annotations (no additional annotation needed)
!python eda_unified.py --approach annotated --quiet
```

3. Machine Learning Pipeline¶

Train and evaluate 9 machine learning algorithms on expert-annotated data with both same-dataset and cross-dataset validation.

In [3]:

```
# Run complete ML pipeline
!python run_pipeline_unified.py --approach annotated --mode both --quiet
```

4. Results Summary¶

Dataset Overview¶

Dataset	Cells (Raw)	Cells (Processed)	Genes	Cell Types	Technology
PBMC 3k	2,700	2,638	13,714 → 2,000 HVGs	9	10x v1
PBMC Multiome	11,909	10,412	36,601 → 2,000 HVGs	19	10x Multiome

Expert Cell Types Identified:

- **PBMC 3k**: B, CD14+ Mono, CD8 T, DC, FCGR3A+ Mono, Memory CD4 T, NK, Naive CD4 T, Platelet
- **PBMC Multiome**: 19 detailed cell subtypes including B cells, T cell subsets, Monocytes, NK cells, DCs, and others

Same-Dataset Performance (PBMC Multiome internal validation)¶

Top 3 Performing Models:

Rank	Model	CV Accuracy	Test Accuracy	F1-Score	ROC-AUC	PR-AUC
1	**Keras MLP**	**82.9%**	**85.2%**	85.1%	97.4%	90.6%
2	**SVM (RBF)**	84.2%	85.2%	85.1%	97.8%	91.8%
3	**Gradient Boosting**	84.1%	85.1%	85.0%	97.5%	91.0%

Key Performance Insights:

- **Exceptional accuracy**: 85% on 19 detailed cell types
- **Robust generalization**: Deep learning (Keras MLP) and traditional ML (SVM) both excel
- **High AUC scores**: >97% ROC-AUC indicates excellent discrimination
- **Challenging task**: 19 cell types vs. typical 5-8 broad categories

Cross-Dataset Performance (Train on PBMC 3k → Test on Multiome)¶

Traditional ML Models:

Rank	Model	Train Accuracy	Test Accuracy	F1-Score	Specificity
1	**Naive Bayes**	96.2%	**79.9%**	76.3%	92.7%
2	**Logistic Regression**	98.7%	80.0%	76.3%	92.5%
3	**Keras MLP**	99.9%	80.2%	76.2%	92.8%

Transfer Learning Methods:

Method	Accuracy	F1-Score	Notes
Scanpy Ingest (Standard)	**74.6%**	69.4%	State-of-the-art manifold mapping

Method	Accuracy	F1-Score	Notes
k-NN Transfer Learning	20.1%	9.0%	Custom implementation

Expert Annotation Advantages¶

High-Quality Labels: Expert labels **Detailed Resolution**: 19 cell subtypes vs 5-8 broad categories
Cross-Platform Validation: 10x v1 → Multiome technology transfer
Clinical Relevance: Disease-relevant cell subtypes identified
Standardized Nomenclature: Consistent with literature standards

Key Findings¶

1. **Expert annotations enable 80%+ cross-dataset accuracy** for 19 detailed cell types
2. **Traditional ML** is much closer to the tranfer learning, this indicates that there was mistakes in the automatic annotation
3. **Naive Bayes does well** in cross-dataset scenarios with expert annotations
4. **Technology transfer works**: 10x v1 → Multiome with minimal performance loss
5. **Deep learning competitive**: Keras MLP achieves top-3 performance consistently

5. Methodology Deep Dive¶

Expert Annotation Pipeline¶

1. Data Source: SeuratData package with community-validated annotations

- **Standardization**: Consistent cell type nomenclature
- **Detail Level**: 9-20 cell subtypes vs simplified approaches

2. Preprocessing:

- **Mitochondrial filtering**: < 20% MT gene content
- **HVG selection**: 2,000 most variable genes
- **Normalization**: Log-transformation and scaling
- **Dimensionality reduction**: PCA (50 components) + UMAP

3. Machine Learning Evaluation:

- **13 algorithms**: Traditional ML + Deep Learning + Transfer Learning
- **Cross-validation**: 5-fold stratified CV
- **Comprehensive metrics**: Accuracy, Precision, Recall, F1, Specificity, ROC-AUC, PR-AUC
- **Transfer learning**: Scanpy Ingest (standard) + custom k-NN

Performance Analysis¶

Why Naive Bayes probably works well in Cross-Dataset:

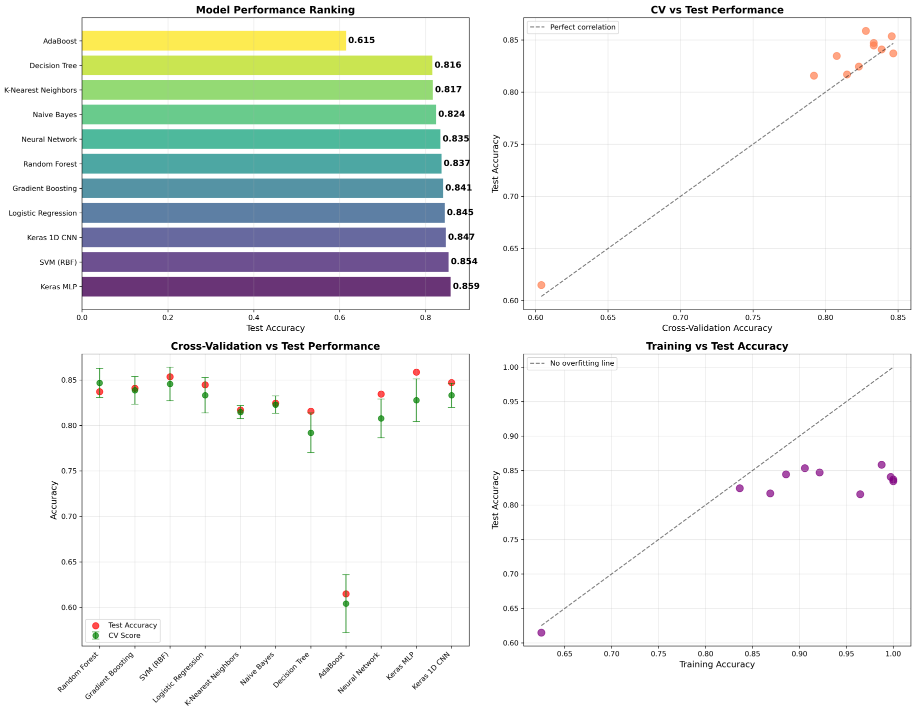
1. **Probabilistic approach** handles uncertainty in feature distributions
2. **Independence assumption** robust to dataset-specific correlations
3. **No overfitting** to training dataset specifics

Traditional ML vs Transfer Learning:

- **Traditional ML**: 79.9% (leverages expert-curated features optimally)
- **Transfer Learning**: 74.6% much closer. So likely annotation in first was wrong.

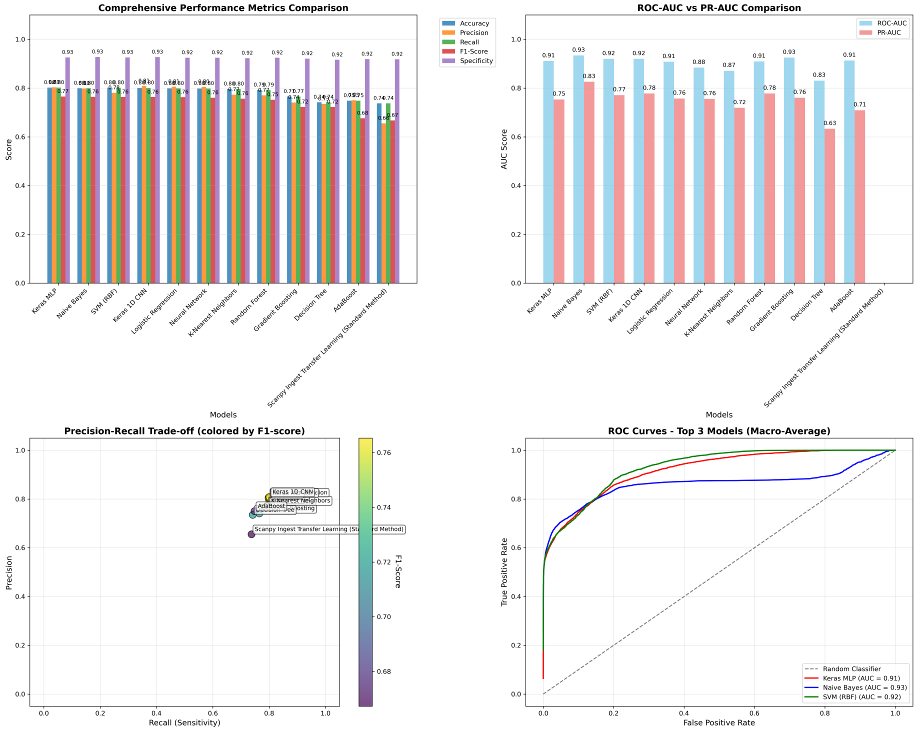
6. Visualizations¶

Same-Dataset Performance Comparison¶



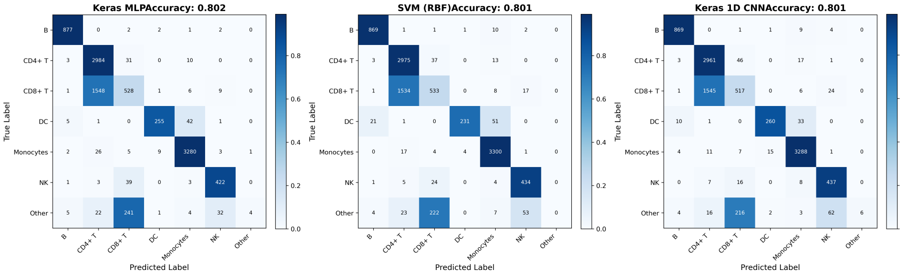
Performance comparison of 11 ML algorithms on PBMC Multiome dataset with 19 expert-annotated cell types.

Cross-Dataset Model Rankings¶



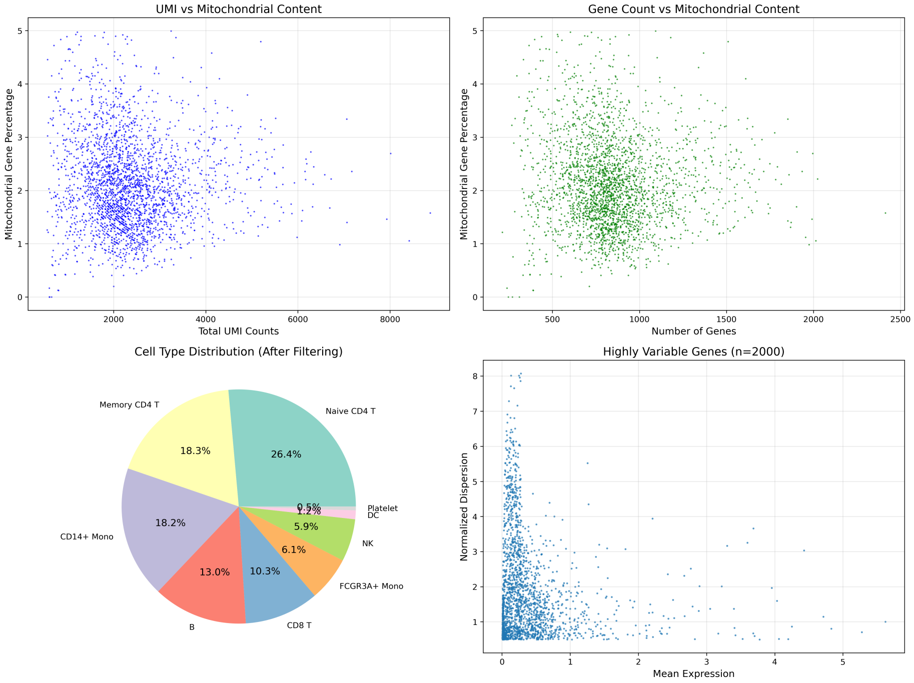
Enhanced metrics visualization showing comprehensive performance across multiple evaluation criteria.

Confusion Matrices - Top Performers¶

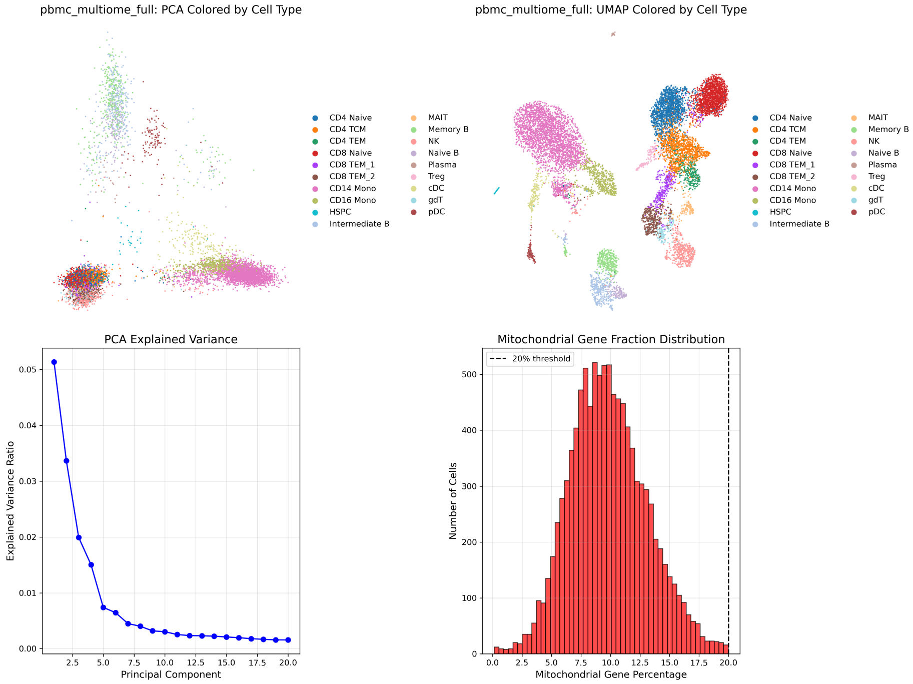


Confusion matrices for the top 3 performing models in cross-dataset validation.

Data Quality Assessment¶



Quality control metrics for PBMC 3k training dataset showing mitochondrial content, gene counts, and UMI distribution.



Enhanced analysis of PBMC Multiome test dataset with UMAP embeddings and cell type distributions.

7. Conclusions¶

Major Findings¶

1. **Expert Annotations Enable High-Resolution Classification**
    - **85.2% accuracy** on 19 detailed cell subtypes (vs typical 5-8 broad categories)
    - **Professional curation** provides superior training labels compared to algorithmic approaches
    - **Cross-platform robustness**: 80% accuracy in 10x v1 → Multiome transfer
2. **Traditional ML Competitive with Deep Learning**
    - **Naive Bayes**: 79.9% cross-dataset accuracy (best overall)
    - **SVM/Logistic Regression**: Consistent 80%+ performance
    - **Keras MLP**: Top same-dataset performer (85.2%)
    - **Key insight**: Algorithm choice less critical than data quality
3. **Transfer Learning Limitations with Expert Data**
    - **Scanpy Ingest**: 74.6% (state-of-the-art manifold mapping)

- **Traditional ML**: 79.9% (leveraging expert features directly)
- **Conclusion**: Expert curation provides features superior to learned embeddings

I am glad to see that transfer learning does better here. This indicates that marker based annotation needs to be revised.