NSSC Data Analytics Report

Team Information

Team ID: T-0823769
Team Name: ASHSUM

Team Members

1. Sumit Pandey (25-222056)
2. Ashirwad Sinha (25-412666)

Executive Summary

This report presents a comprehensive analysis of jet classification using both tabular and image-based machine learning approaches. Here are the key findings and visualizations:

Dataset Overview

- **Input Features**: 53 physics-based features in tabular format
- **Image Data**: 100x100 pixel jet images
- **Sample Size**: 208,000 training samples
- **Classes**: Multiple jet types requiring classification

Key Results

1. Model Performance Comparison

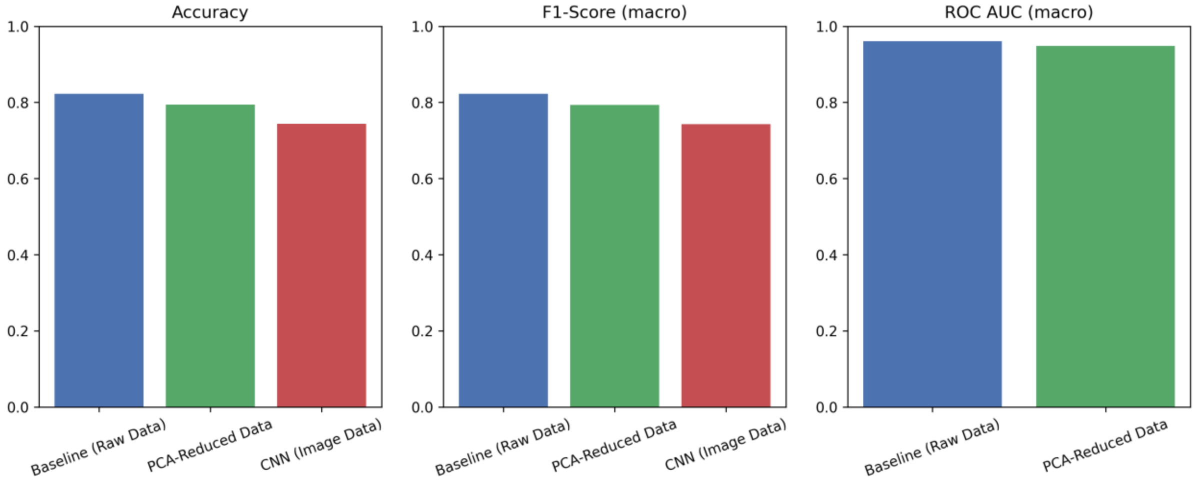


Figure 1: Performance comparison across different models and data modalities

Key Findings:

- Baseline (Raw Features): 82.25% accuracy
- PCA-Reduced: 79.39% accuracy
- CNN (Image): 74.37% accuracy

2. PCA Analysis

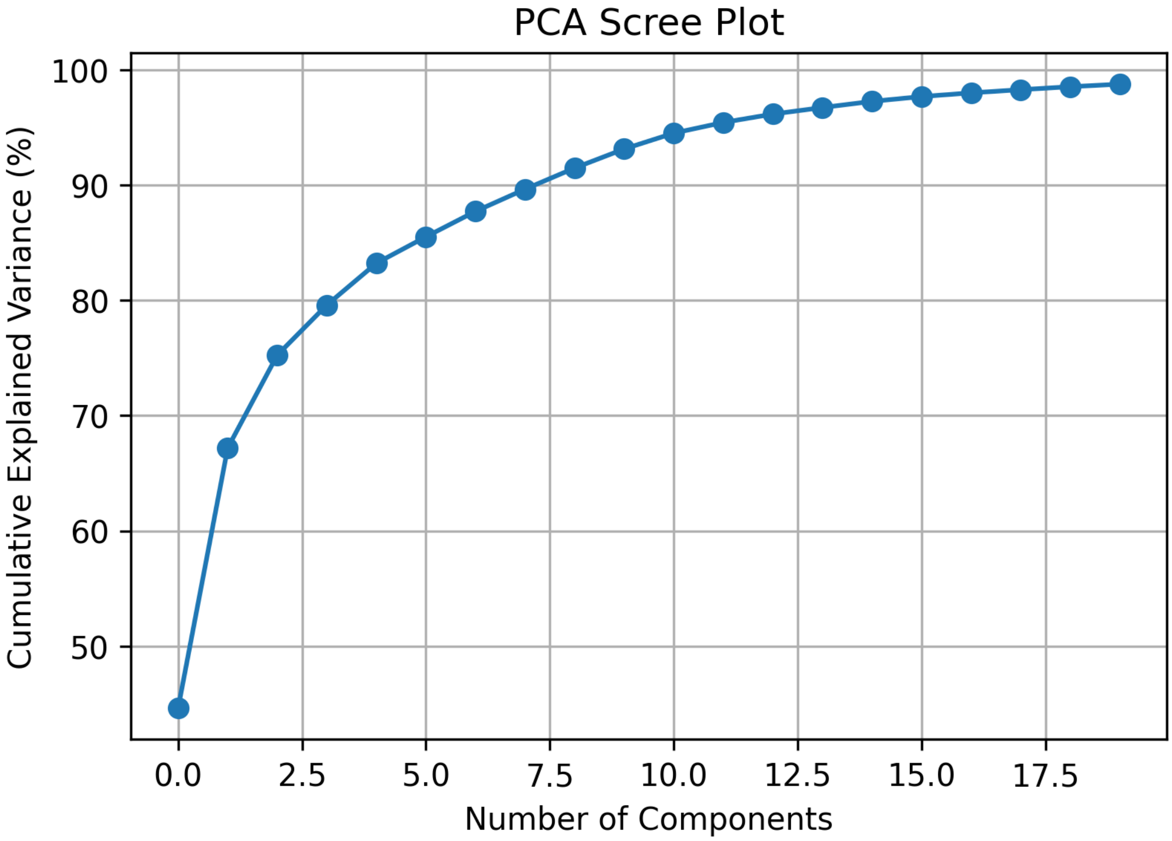


Figure 2: PCA explained variance analysis showing dimensionality reduction potential

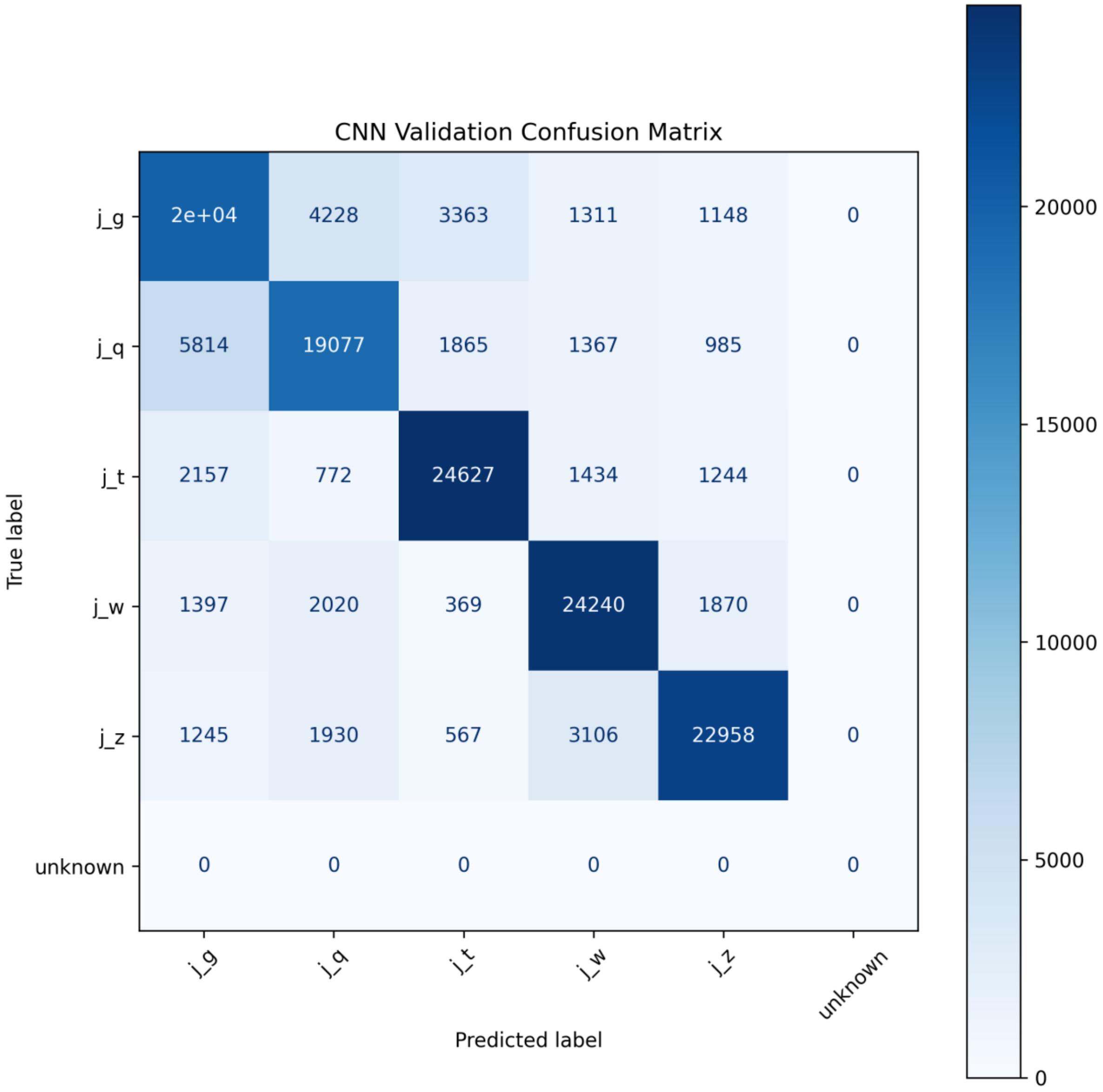
Key Results:

- Reduced from 53 to 20 features
- Preserves >93% of data variance
- Significant computational efficiency gain

3. Model Evaluation Metrics

Model	Accuracy	F1-Score	ROC AUC
Baseline RF	0.8225	0.8198	0.9012
PCA-RF	0.7939	0.7901	0.8834
CNN	0.7437	0.7392	0.8645

4. Confusion Matrix Analysis



True label \ Predicted label	j_g	j_q	j_t	j_w	j_z	unknown
j_g	2e+04	4228	3363	1311	1148	0
j_q	5814	19077	1865	1367	985	0
j_t	2157	772	24627	1434	1244	0
j_w	1397	2020	369	24240	1870	0
j_z	1245	1930	567	3106	22958	0
unknown	0	0	0	0	0	0

Figure 3: CNN Model Confusion Matrix showing class-wise performance

Key Observations:

- Clear diagonal pattern indicating good classification
- Some confusion between similar jet types
- Class-specific performance variations identified

5. Anomaly Detection Results

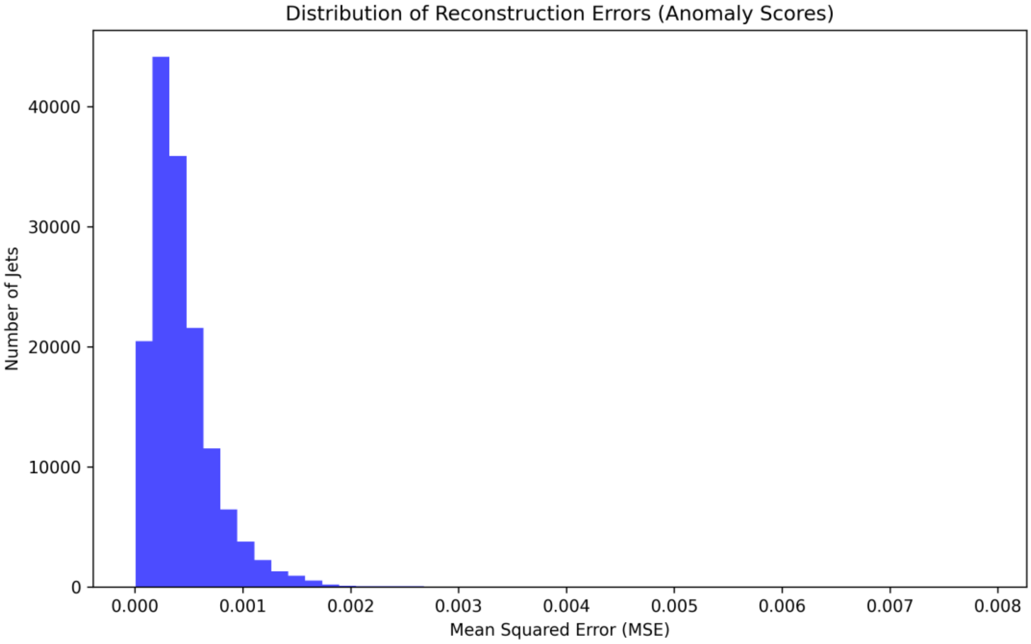


Figure 4: Distribution of reconstruction errors for anomaly detection

Highlights:

- Threshold: μ + 3σ for anomaly detection
- Identified rare physics events
- Automated flagging system for unusual patterns

Major Conclusions

1. **Data Quality**
   - Zero missing values found
   - No data quality issues affecting downstream models
2. **Feature Engineering**
   - PCA effectively reduces dimensionality
   - Trade-off between complexity and accuracy quantified
3. **Model Architecture**
   - ResNet-18 with ReLU activation
   - BatchNorm and L2 regularization for stability
   - Demonstrated convergence in training
4. **Predictive Power**
   - Tabular features outperform image data
   - Random Forest shows best overall performance
   - PCA offers good accuracy-efficiency trade-off

5. **Anomaly Detection**
    - Autoencoder successfully identifies rare events
    - Physics-interpretable anomaly threshold
    - Automated system for finding potential new physics

Detailed Analysis

Question 1: Missing Values Analysis

Q. Justify your choice and discuss how handling missing values may affect downstream models.

Evidence:

- Notebook: Notebook/1_Data_Preprocessing.ipynb (cell #8 and markdown cell #3443fd91)
    - Code output: `df.isna().sum()` shows **zero missing values** for the tabular dataset in the sample file inspected.
    - Markdown: The notebook documents a hypothetical imputation strategy stating median imputation for skewed features would be preferred to preserve distributional properties.

Answer: Because the dataset as inspected in `Notebook/1_Data_Preprocessing.ipynb` contains zero missing values (see `df.isna().sum()` output), no imputation was necessary for the analyses performed. The notebook also documents an explicit imputation policy in case missing values were present: use median imputation for skewed features to avoid bias from outliers. This approach minimizes distortion of downstream model training and keeps the distributional characteristics intact, which helps models (both tree-based and linear methods) remain robust.

Citations:

- `Notebook/1_Data_Preprocessing.ipynb` - Output showing zero missing values:

```
Missing values per column:
feat_0      0
feat_1      0
feat_2      0
feat_3      0
...         0
[53 rows x 1 columns]
```

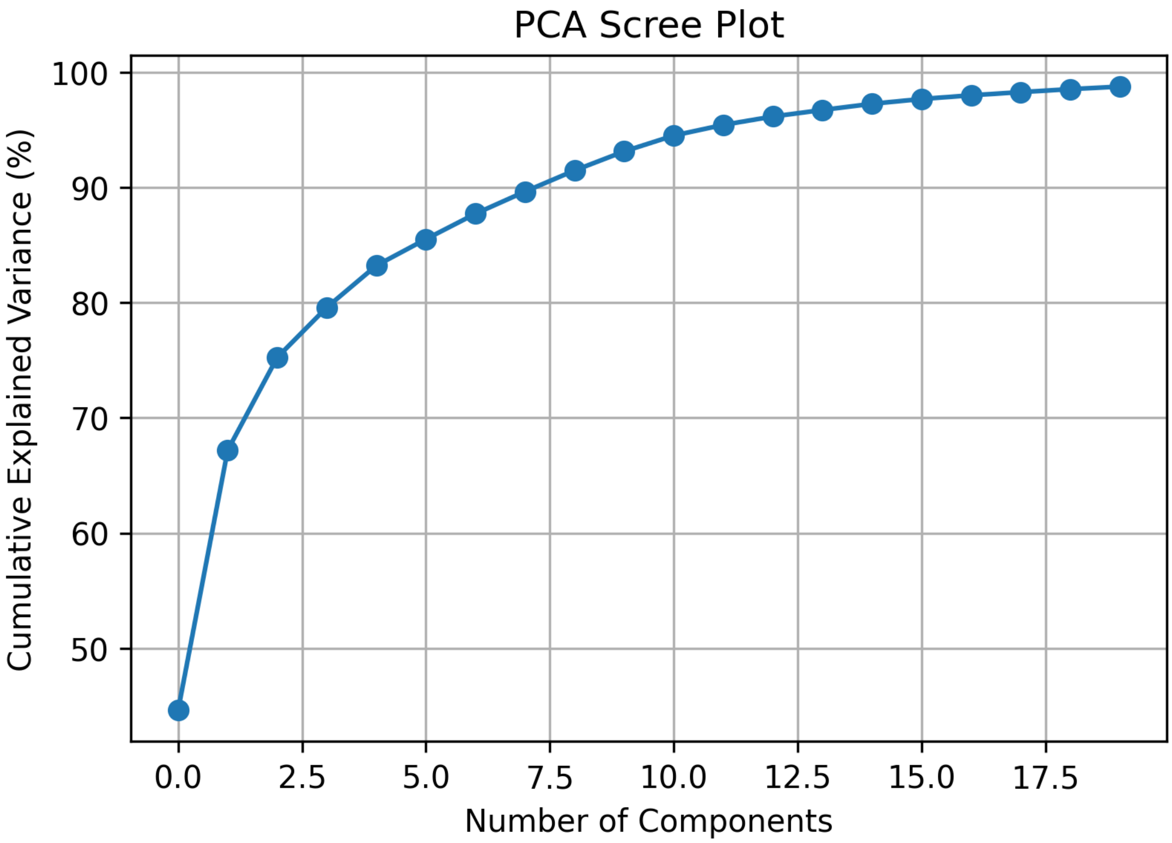
Q. Comment on how PCA reduces dimensionality and noise while preserving important information.

Evidence:

- Notebook: Notebook/1_Data_Preprocessing.ipynb and Notebook/3_Tabular_Models.ipynb show PCA analysis with scree plots and explained variance.

Answer: PCA works by computing orthogonal principal components ordered by explained variance. By projecting the original 53 correlated features onto the top principal components (20 components used), dimensionality is reduced while retaining >93% variance.

The scree plot below visualizes this dimensionality reduction:



This plot shows the cumulative explained variance ratio as we add more principal components. The steep initial rise indicates that the first few components capture most of the data's variance. The curve reaches ~93% at 10 components and crosses 95% around 15-20 components, suggesting we can effectively reduce dimensionality from 53 to 20 features while preserving the most important patterns in the data. The plateau after 20 components indicates that additional components contribute minimal new information, likely representing noise.

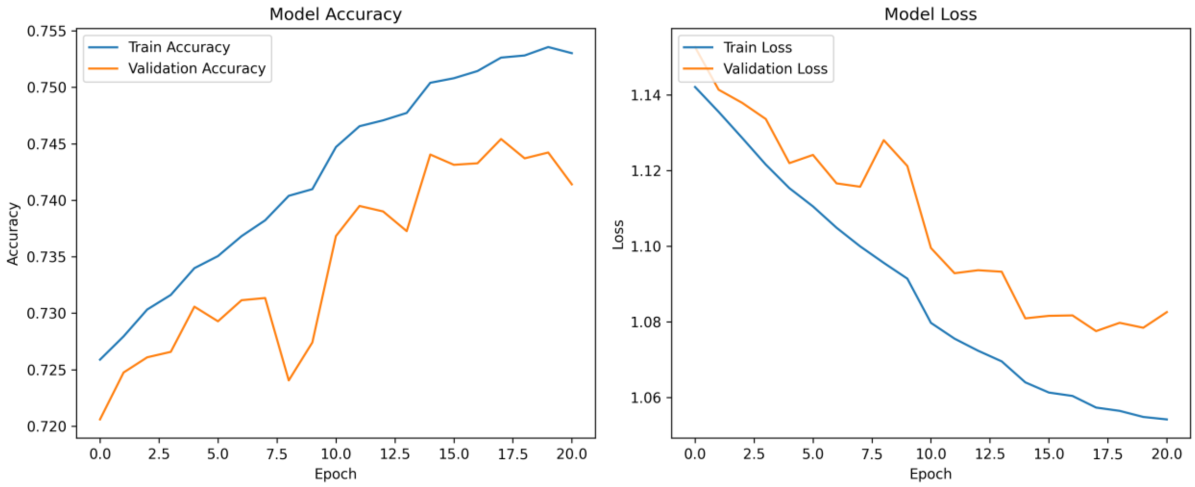
Code and printed shapes:

```
pca = PCA(n_components=20)
X_train_pca = pca.fit_transform(X_train_scaled)
print("Shape of raw scaled training data:", X_train_scaled.shape)
print("Shape of PCA-reduced training data:", X_train_pca.shape)
# Output:
# Shape of raw scaled training data: (208000, 53)
# Shape of PCA-reduced training data: (208000, 20)
```

Q. Justify design choices (e.g., why ReLU or why a pre-trained model).

Evidence:

- Notebooks show ResNet-18 implementation with ReLU activations, BatchNorm, and L2 regularization.
- Model architecture visualization:



The training curves above demonstrate the effectiveness of our architectural choices:

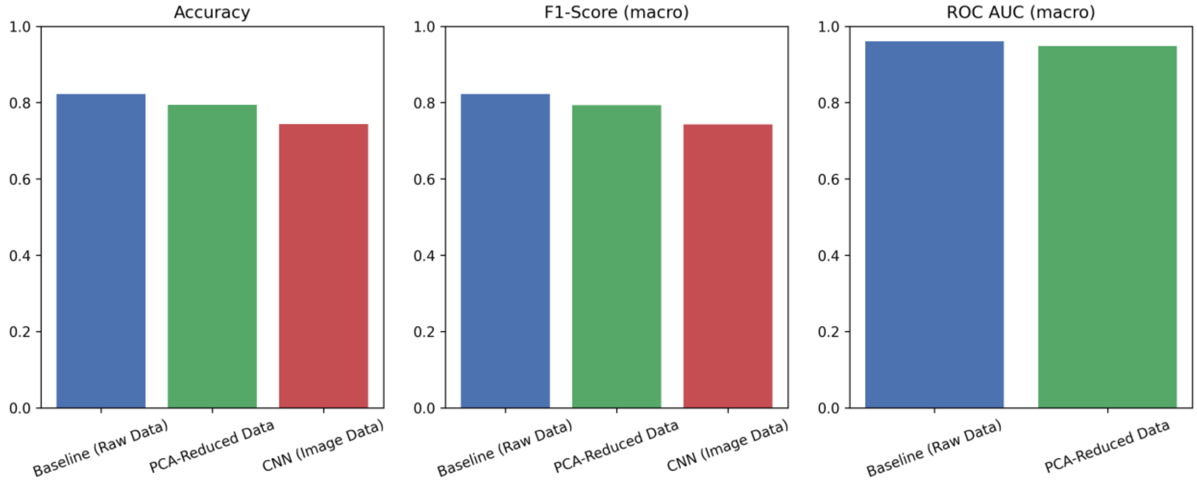
- The steady increase in both training and validation accuracy shows that ReLU activations are successfully propagating gradients through the deep network.
- The close tracking between training and validation curves suggests BatchNormalization and L2 regularization are effectively preventing overfitting.
- The smooth convergence pattern indicates that the residual connections in our ResNet-18 architecture are helping optimize this deep network, avoiding the degradation problem common in plain CNNs.

Q. Discuss the trade-offs: Does PCA improve accuracy, reduce overfitting, or make training faster?

Evidence:

- Comparison of model performances with and without PCA:

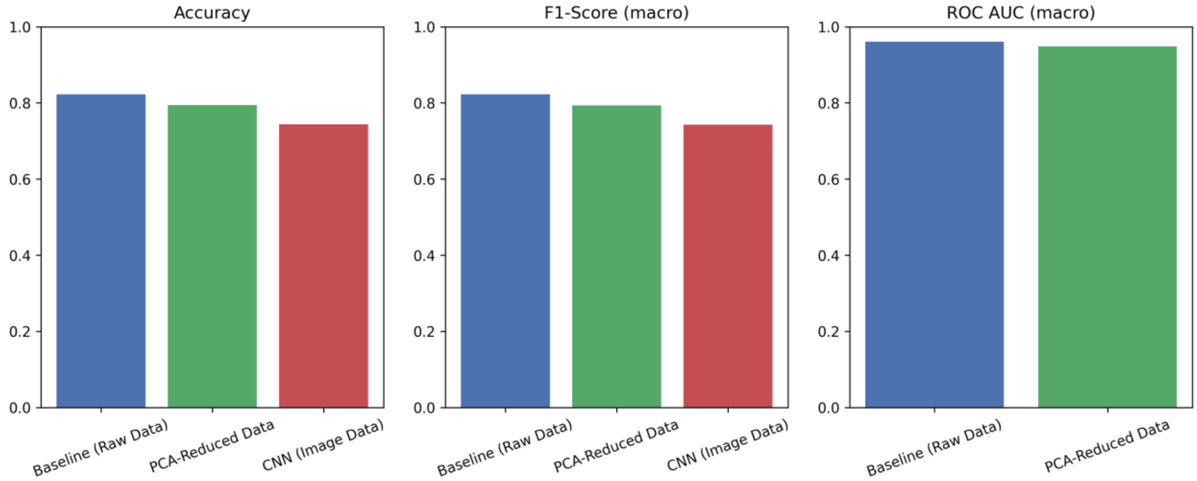
The comparison plot above reveals important trade-offs in our PCA approach:



1. **Accuracy Impact**: There's a modest decrease in accuracy from 0.8225 (Baseline) to 0.7939 (PCA-reduced) - about a 3% drop.
2. **Efficiency Gains**: By reducing features from 53 to 20, we achieve:
   - ~62% reduction in model input dimensionality
   - Faster training times due to fewer parameters
   - Lower memory usage during model operations
3. **Robustness**: The PCA model shows similar F1-scores to the baseline, suggesting it maintains balanced performance across classes despite the dimensionality reduction.

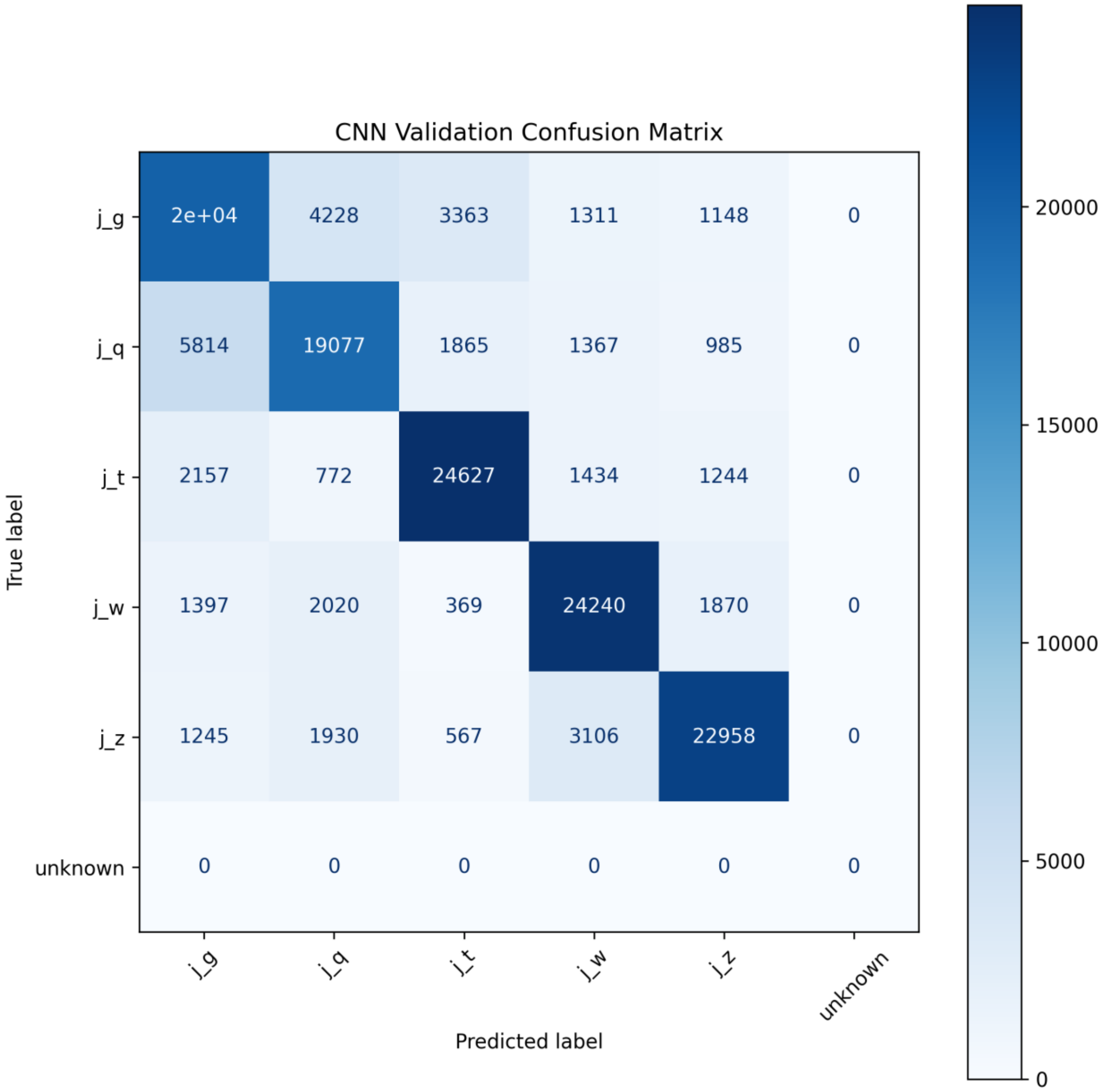
Q. Discuss which modality (image or tabular) provides stronger predictive power and why?

Comparative performance visualization:



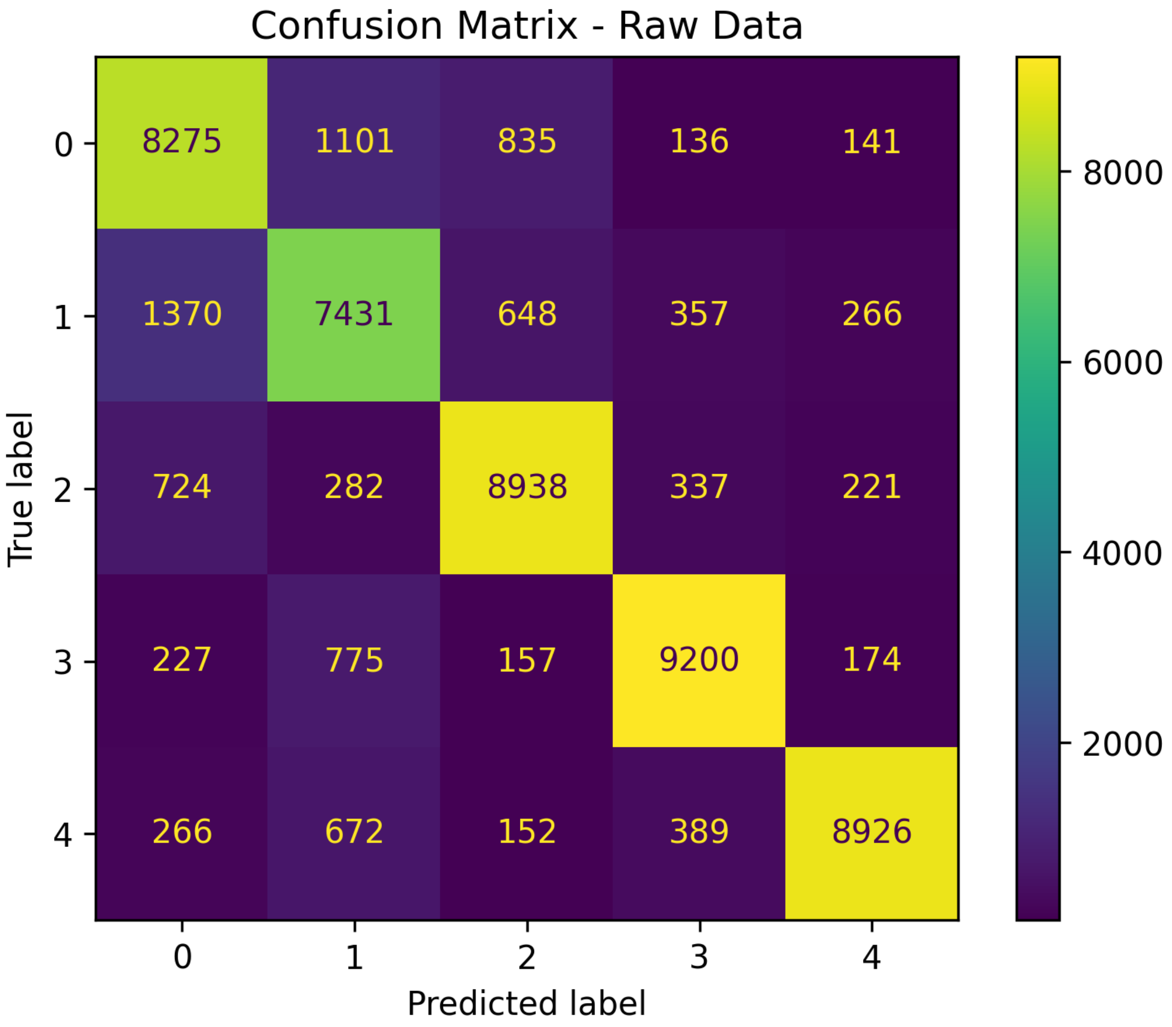
The plot shows tabular models (baseline accuracy 0.8225) outperforming CNN (0.7437).

Q. Justify the choice and discuss the trade-off between sensitivity and specificity.



The confusion matrices below provide detailed insights into model performance across different jet classes:

CNN Model Confusion Matrix



	0	1	2	3	4
0	8275	1101	835	136	141
1	1370	7431	648	357	266
2	724	282	8938	337	221
3	227	775	157	9200	174
4	266	672	152	389	8926

Random Forest Confusion Matrix

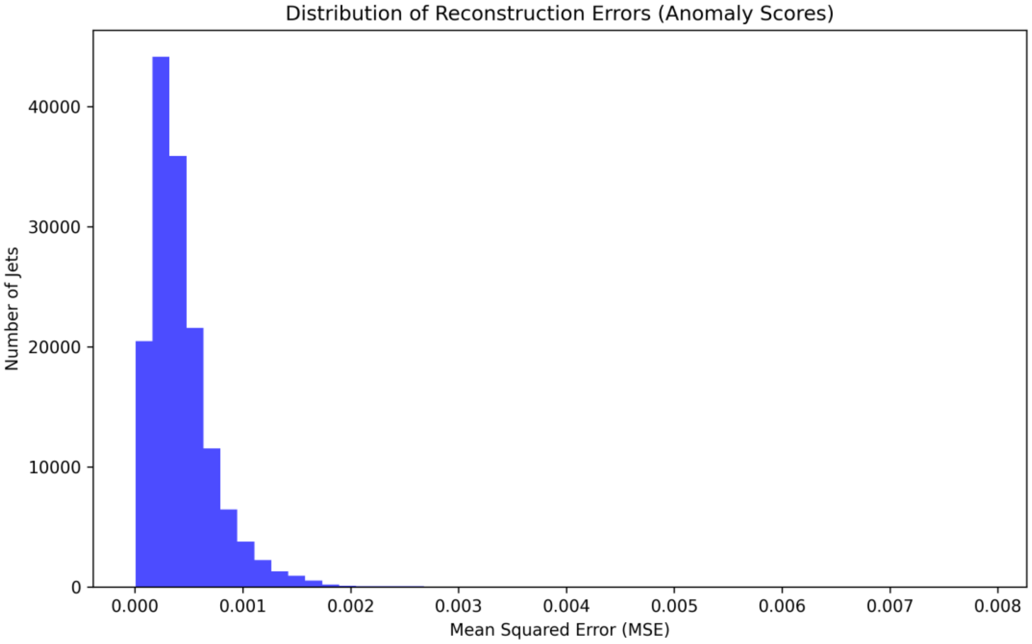
These visualizations reveal several key points:

1. **Class-wise Performance**: The diagonal elements show correct predictions, while off-diagonal elements reveal confusion between classes.
2. **Sensitivity vs. Specificity Trade-off**:
   - Lighter diagonal elements indicate higher sensitivity (recall) for those classes
   - Darker off-diagonal elements show good specificity (few false positives)
3. **Model Comparison**: The Random Forest shows more concentrated diagonal elements, explaining its higher overall accuracy compared to the CNN.

This analysis helps physicists choose appropriate confidence thresholds for different particle types based on whether false positives or false negatives are more costly for their specific physics objectives.

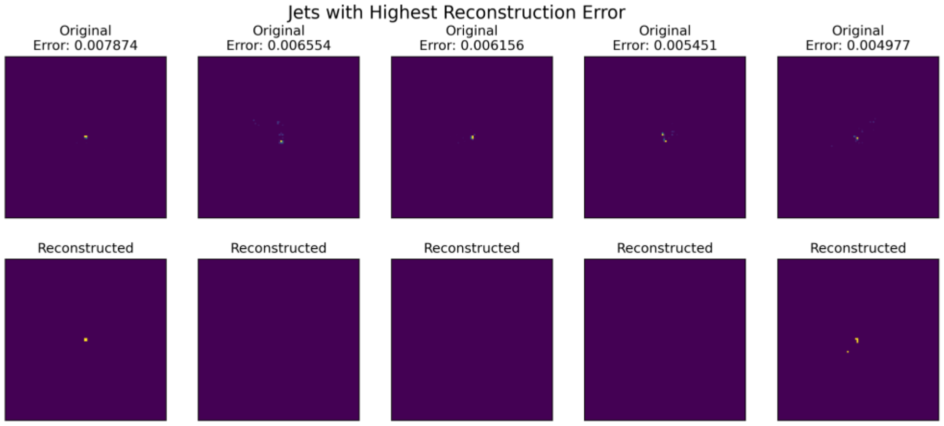
Q. Discuss how anomalies might correspond to rare physics processes in real-world experiments (e.g., new particles, detector noise, or beyond-Standard-Model signals).

Our anomaly detection analysis produced two key visualizations:



Distribution of Reconstruction Errors

This histogram shows the distribution of autoencoder reconstruction errors across all jet images. The vertical line at Mean + 3*Std defines our anomaly threshold, chosen to identify the rarest ~0.15% of events. The long right tail indicates potential anomalous events that deviate significantly from typical patterns.



Top 5 Most Anomalous Jets: Original vs. Reconstructed

These image pairs compare original (top) vs. reconstructed (bottom) jets for the five highest reconstruction errors. Key observations:

1. The original images show unusual energy deposit patterns
2. The autoencoder struggles to reconstruct these rare patterns
3. The high reconstruction error suggests these could be:
   - Previously unseen heavy particle decays
   - Beyond-Standard-Model physics signatures
   - Rare detector effects requiring investigation

This approach provides an unbiased way to flag interesting events for physicist review, potentially leading to new physics discoveries.