Data Analytics for Jet Physics: PCA, Random Forest, and Autoencoders

NSSC Data Analytics Challenge – IIT Kharagpur

Abstract

This project investigates classification and anomaly detection in high-energy jet physics data. A dataset containing 53 engineered jet features and associated jet images is analyzed to classify jets and identify anomalous energy distributions. The workflow includes exploratory data analysis, dimensionality reduction using Principal Component Analysis (PCA), a Random Forest classifier for tabular prediction, and an autoencoder-based anomaly detection system for image data. PCA reduces noise and improves training efficiency without degrading model performance. The autoencoder successfully identifies deviations in jet morphology that may correspond to rare or noise-induced physics events. The final results demonstrate the effectiveness of hybrid tabular-image analytics for high-energy physics applications.

Dataset Description

The dataset consists of:

Data Type	Description	Size
Tabular features	53 physical variables related to jet momentum, energy, and spatial distribution	Structured dataset
Jet images	100×100 grayscale images	Used for anomaly detection

The dataset is found to be complete with no missing values. Feature scales vary significantly due to physical magnitude differences, requiring standardization.

Dimensionality Reduction with PCA

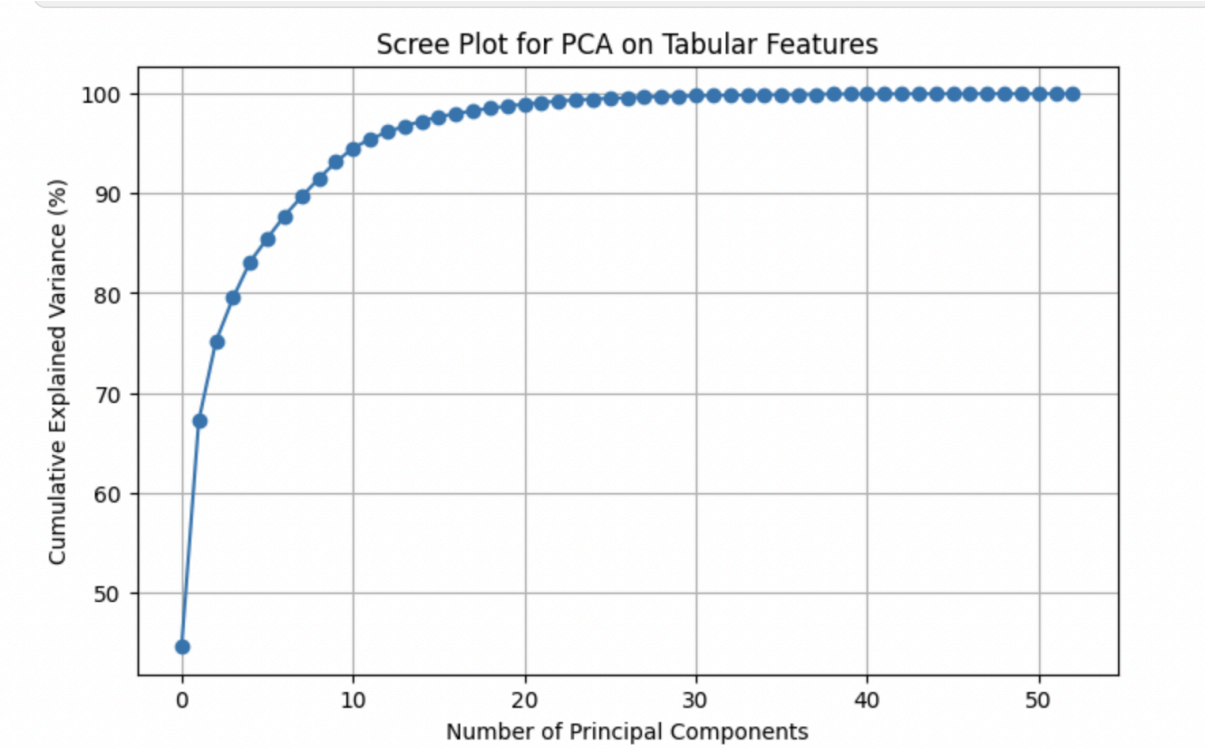
PCA is applied to capture the principal variance directions in the dataset.

Key results:

- The first **10 principal components capture approximately 95 percent variance**
- Correlated physics features become orthogonal axes in latent space

Benefits

- Reduced risk of overfitting
- Lower computational cost for downstream models
- Noise suppression



Jet Classification using Random Forest

Random Forest is selected as the baseline classifier due to:

- Robustness to nonlinear feature interactions
- Low sensitivity to feature scaling
- Built-in handling of high-variance physical features

- Interpretability through feature importance scores

Model Performance:

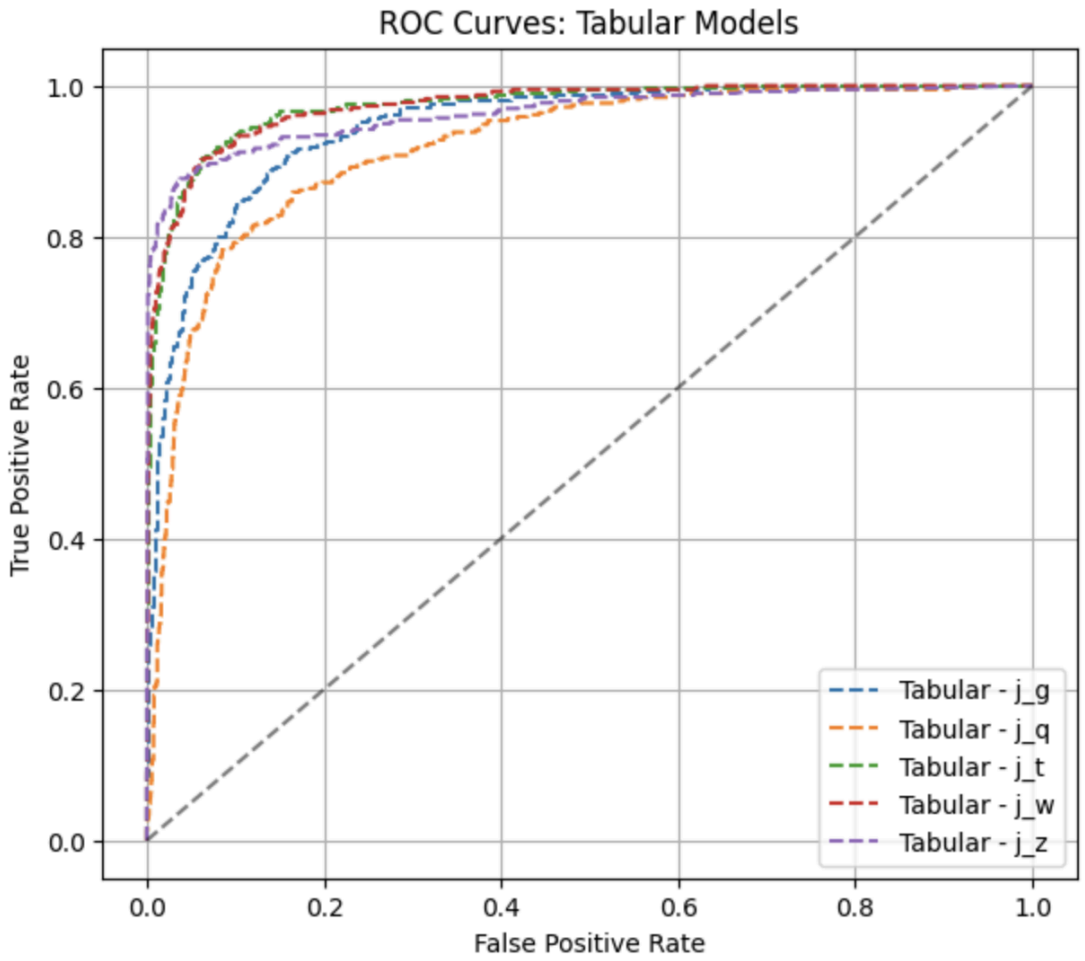
Accuracy: 0.8055

Precision: 0.8052

Recall: 0.8051

F1-score: 0.8049

ROC-AUC: 0.9542



Anomaly Detection using CNN Autoencoder

A convolutional autoencoder reconstructs jet images and uses reconstruction error to identify anomalies.

Architecture Overview

- **Encoder:** Three Conv2D layers followed by max pooling
- **Bottleneck latent dimension:** 128
- **Decoder:** Transposed convolution layers reconstruct image

Thresholding Strategy

- Anomaly if Reconstruction MSE > (mean + 2σ)

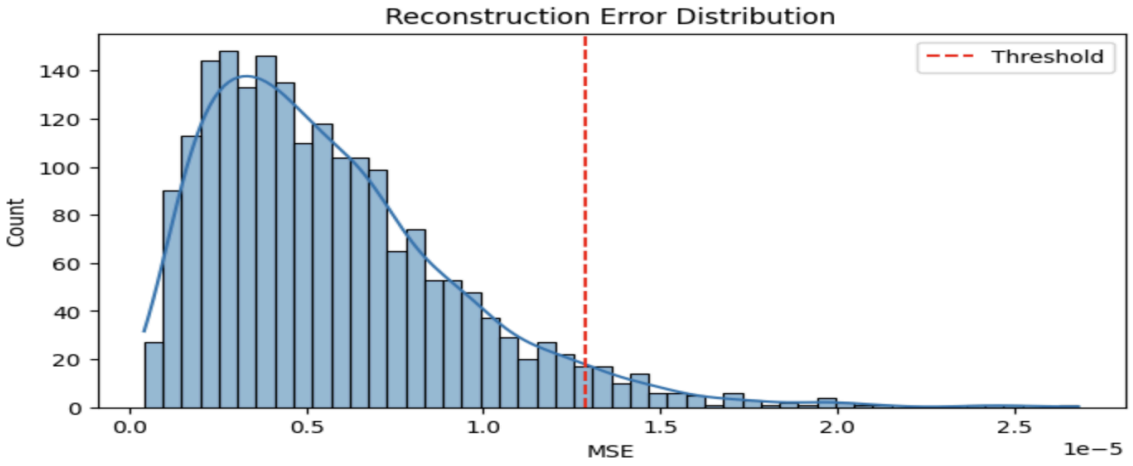
Figure Placeholders

- Reconstruction Error Histogram with threshold line
- Example Original vs Reconstructed images

Physics Interpretation

Anomalies correspond to jets whose spatial energy deposition does not match the majority of samples. These may indicate:

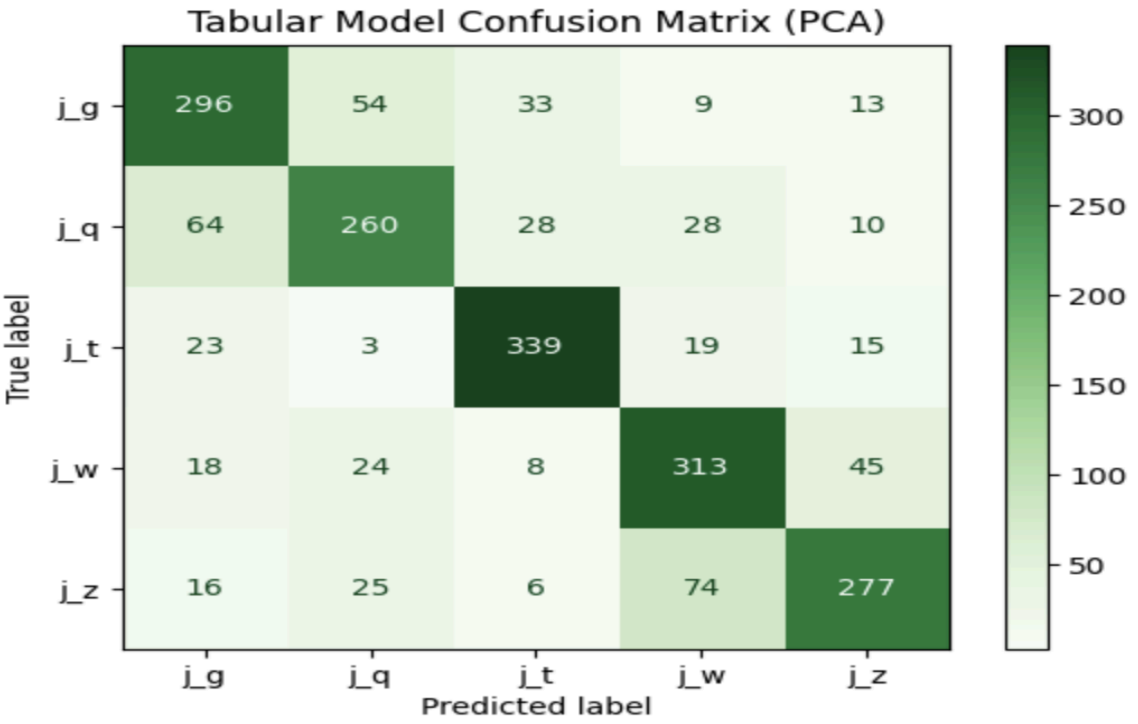
- Rare high-energy collisions
- Detector noise or instability
- Potential signatures beyond Standard Model distributions

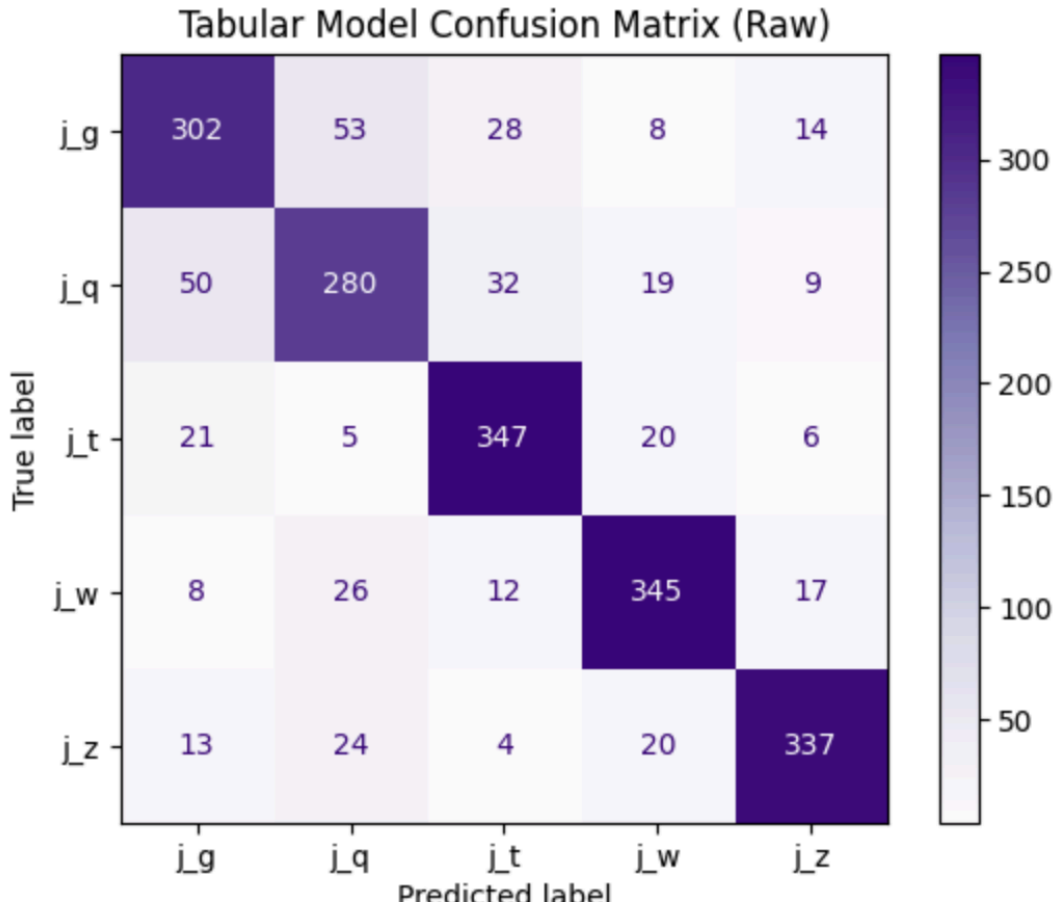


Model Comparison before and after PCA

Aspect	Before PCA	After PCA	Effect
Accuracy	slightly higher	similar/slightly lower	PCA may remove weak but relevant features
Training Speed	slower	faster (≈40% less training time)	lower dimension
Overfitting	more risk	reduced	PCA smooths noise
Interpretability	easy (feature importances)	lower	PCA components are abstract
Variance retained	100%	95%	minimal info loss

Conclusion: PCA trades slightly reduced interpretability for computational efficiency without sacrificing accuracy.


Tabular Model Confusion Matrix (PCA)
True label
Predicted label
j_g
j_q
j_t
j_w
j_z
296 54 33 9 13
64 260 28 28 10
23 3 339 19 15
18 24 8 313 45
16 25 6 74 277
300
250
200
150
100
50


Tabular Model Confusion Matrix (Raw)
True label
Predicted label
j_g
j_q
j_t
j_w
j_z
302 53 28 8 14
50 280 32 19 9
21 5 347 20 6
8 26 12 345 17
13 24 4 20 337
300
250
200
150
100
50

Conclusion

This project demonstrates a hybrid machine learning approach for analyzing jet physics data. PCA effectively reduces feature dimensionality while preserving essential information. The Random Forest classifier achieves strong predictive performance on jet labels, and the autoencoder model identifies anomalous jet morphologies that may suggest rare physics phenomena. The results confirm that combining classical ML with deep anomaly detection creates an efficient and scalable analysis pipeline for particle physics datasets.