Evaluation Report

1. Overview

This report compares the performance of twelve machine learning models trained to classify four football skill actions: **Heading**, **Juggling**, **Pass**, and **Shots**. The dataset contains 200 balanced windows (50 per class) with 72 engineered features per window. A 5-fold cross-validation was used on the training set, and a held-out test set of 40 windows (10 per class) was used for final evaluation.

For each model, we report test accuracy, cross-validation accuracy (as a proxy for stability), confusion matrix, false positives/false negatives, main confusion patterns, inference time per sample, ability to support real-time deployment, batch size assumptions, and learning rate settings (where applicable). All inference times were measured using a batch size of 64 samples; most algorithms train in a full-batch or closed-form manner.

2. Global Comparison Summary

The table below summarizes the key quantitative metrics across all models.

Model	Test Accuracy	CV Accuracy (Stability)	Average Inference Time per Sample (ms)	Real-Time Comment	CSV Verdict
AdaBoost	97.50%	88.50% (Unstable)	0.222	Real-time capable but noticeably slower than the fastest models.	Most Accurate (Fixes Pass/Shot)
CatBoost	95.00%	96.00%	0.032	Excellent real-time performance (	Most Stable & Robust

XGBoost	95.00%	95.50%	0.029	Excellent real-time performance (	Best All-Rounder
SVM (RBF)	95.00%	93.50%	0.021	Excellent real-time performance (	Speed Demon (High Accuracy)
Random Forest	95.00%	96.50%	0.650	Still real-time for this dataset, but significantly heavier; may limit deployment on low-power devices or at higher sampling rates.	Too Slow
KNN	95.00%	92.50%	0.264	Real-time capable but noticeably slower than the fastest models.	Memory Heavy
LightGBM	92.50%	94.50%	0.058	Real-time capable but noticeably slower than the fastest models.	Good, but beaten by XGBoost
Logistic Regression	92.50%	95.50%	0.013	Excellent real-time performance (	Excellent Backup
Decision Tree	92.50%	90.50%	0.019	Excellent real-time performance (	Unstable
LDA	90.00%	92.50%	0.009	Excellent real-time performance (	Fastest (Low Accuracy)
Naive Bayes	90.00%	91.00%	0.014	Excellent real-time performance (	Low Accuracy
QDA	42.50%	60.00%	0.012	Excellent real-time performance (	Failed (Do not use)

Overall, **AdaBoost** achieved the highest test accuracy (97.50%), while **CatBoost**, **XGBoost**, **Random Forest**, **KNN**, and **SVM**

(RBF) all reached 95.00% test accuracy with varying trade-offs in stability and speed. **LDA** was the fastest model, but at the cost of lower accuracy, and **QDA** clearly failed on this dataset. All models meet basic real-time constraints for this problem size, although Random Forest and KNN are markedly heavier at inference.

3. Model-by-Model Analysis

3.1 AdaBoost

Key metrics

- Test accuracy: **97.50%** (39 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **88.50% (Unstable)**.
- Average inference time per sample: **0.222 ms** (batch size 64). Real-time capable but noticeably slower than the fastest models.
- Learning rate: **0.1**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Most Accurate (Fixes Pass/Shot)**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	1	9	0
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 1 sample(s) of true 'Pass' predicted as 'Juggling'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Juggling**.

3.2 CatBoost

Key metrics

- Test accuracy: **95.00%** (38 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **96.00%**.
- Average inference time per sample: **0.032 ms** (batch size 64). Excellent real-time performance (
- Learning rate: **0.05**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Most Stable & Robust**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	0	8	2
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 2 sample(s) of true 'Pass' predicted as 'Shots'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.3 XGBoost

Key metrics

- Test accuracy: **95.00%** (38 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **95.50%**.
- Average inference time per sample: **0.029 ms** (batch size 64). Excellent real-time performance (
- Learning rate: **0.1**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Best All-Rounder**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	0	8	2
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 2 sample(s) of true 'Pass' predicted as 'Shots'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.4 SVM (RBF)

Key metrics

- Test accuracy: **95.00%** (38 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **93.50%**.
- Average inference time per sample: **0.021 ms** (batch size 64). Excellent real-time performance (

- Learning rate: **N/A (quadratic programming; no explicit learning rate)**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Speed Demon (High Accuracy)**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	0	8	2
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 2 sample(s) of true 'Pass' predicted as 'Shots'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.5 Random Forest

Key metrics

- Test accuracy: **95.00%** (38 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **96.50%**.
- Average inference time per sample: **0.650 ms** (batch size 64). Still real-time for this dataset, but significantly heavier; may limit deployment on low-power devices or at higher sampling rates.
- Learning rate: **N/A (ensemble of trees, no learning rate)**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Too Slow**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	0	8	2
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 2 sample(s) of true 'Pass' predicted as 'Shots'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.6 KNN

Key metrics

- Test accuracy: **95.00%** (38 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **92.50%**.
- Average inference time per sample: **0.264 ms** (batch size 64). Real-time capable but noticeably slower than the fastest models.
- Learning rate: **N/A (lazy learner; no gradient descent)**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Memory Heavy**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	0	8	2
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 2 sample(s) of true 'Pass' predicted as 'Shots'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.7 LightGBM

Key metrics

- Test accuracy: **92.50%** (37 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **94.50% ± 5.34%** (fold scores: 85.00%, 97.50%, 97.50%, 100.00%, 92.50%).
- Average inference time per sample: **0.058 ms** (batch size 64). Real-time capable but noticeably slower than the fastest models.
- Learning rate: **0.05**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Good, but beaten by XGBoost**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	1	7	2
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 1 sample(s) of true 'Pass' predicted as 'Juggling'.
- 2 sample(s) of true 'Pass' predicted as 'Shots'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.8 Logistic Regression

Key metrics

- Test accuracy: **92.50%** (37 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **95.50%**.
- Average inference time per sample: **0.013 ms** (batch size 64). Excellent real-time performance (
- Learning rate: **N/A (LBFGS line search; no fixed learning rate)**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Excellent Backup**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	0	7	3
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 3 sample(s) of true 'Pass' predicted as 'Shots'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.9 Decision Tree

Key metrics

- Test accuracy: **92.50%** (37 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **90.50%**.
- Average inference time per sample: **0.019 ms** (batch size 64). Excellent real-time performance (
- Learning rate: **N/A (tree grown in one pass; no learning rate)**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Unstable**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	8	2	0	0
Juggling	0	10	0	0
Pass	0	1	9	0
Shots	0	0	0	10

Error analysis (false positives / false negatives)

- 2 sample(s) of true 'Heading' predicted as 'Juggling'.
- 1 sample(s) of true 'Pass' predicted as 'Juggling'.

What the model gets confused about: The dominant confusion pattern is **Heading** being predicted as **Juggling**.

3.10 LDA

Key metrics

- Test accuracy: **90.00%** (36 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **92.50%**.
- Average inference time per sample: **0.009 ms** (batch size 64). Excellent real-time performance (
- Learning rate: **N/A (closed-form solution; no learning rate)**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Fastest (Low Accuracy)**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	0	7	3
Shots	0	0	1	9

Error analysis (false positives / false negatives)

- 3 sample(s) of true 'Pass' predicted as 'Shots'.
- 1 sample(s) of true 'Shots' predicted as 'Pass'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.11 Naive Bayes

Key metrics

- Test accuracy: **90.00%** (36 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **91.00%**.
- Average inference time per sample: **0.014 ms** (batch size 64). Excellent real-time performance (
- Learning rate: **N/A (analytic parameter estimation; no learning rate)**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Low Accuracy**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	10	0	0	0
Juggling	0	10	0	0
Pass	0	1	7	2
Shots	1	0	0	9

Error analysis (false positives / false negatives)

- 1 sample(s) of true 'Pass' predicted as 'Juggling'.
- 2 sample(s) of true 'Pass' predicted as 'Shots'.
- 1 sample(s) of true 'Shots' predicted as 'Heading'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Shots**.

3.12 QDA

Key metrics

- Test accuracy: **42.50%** (17 / 40 test windows classified correctly).
- Cross-validation accuracy (5-fold): **60.00%**.
- Average inference time per sample: **0.012 ms** (batch size 64). Excellent real-time performance (
- Learning rate: **N/A (closed-form Gaussian parameters; no learning rate)**.
- Training / batch size: trained on the full available training set (160 windows); inference timing used batches of 64 samples.
- Overall verdict (from CSV summary): **Failed (Do not use)**.

Confusion matrix (rows = true class, columns = predicted class):

True \ Pred	Heading	Juggling	Pass	Shots
Heading	4	6	0	0
Juggling	0	10	0	0
Pass	1	7	0	2
Shots	1	6	0	3

Error analysis (false positives / false negatives)

- 6 sample(s) of true 'Heading' predicted as 'Juggling'.
- 1 sample(s) of true 'Pass' predicted as 'Heading'.
- 7 sample(s) of true 'Pass' predicted as 'Juggling'.
- 2 sample(s) of true 'Pass' predicted as 'Shots'.
- 1 sample(s) of true 'Shots' predicted as 'Heading'.
- 6 sample(s) of true 'Shots' predicted as 'Juggling'.

What the model gets confused about: The dominant confusion pattern is **Pass** being predicted as **Juggling**.

3. Conclusion & Final Model

Across the twelve models, most achieved very strong performance on the four skills, with Heading and Juggling almost always predicted correctly. The main difficulty for all models was telling **Pass** and **Shots** apart, which is expected because their motion patterns are very similar. All models were fast enough for real-time use, but some (like Random Forest and KNN) were noticeably heavier at inference time. LightGBM, CatBoost and XGBoost all showed strong

cross-validation results, while QDA clearly failed and will not be used.

After weighing accuracy, stability, speed and practicality, **we choose XGBoost as the final model**. It offers a **high test accuracy (95%)**, **strong and stable cross-validation performance (≈95.5%)**, and a **very low inference time (~0.029 ms per window)**. Its confusion matrix is clean: it handles Heading and Juggling perfectly and only makes a few remaining mistakes between Pass and Shots.

We deliberately did not select AdaBoost, even though it has the highest single test accuracy, because its lower cross-validation score suggests it is more sensitive to the exact data split (higher risk of overfitting). CatBoost performs similarly to XGBoost but is heavier to train and integrate, and Random Forest, while accurate, is slower at inference.

In short, **XGBoost gives the best overall balance** between accuracy, robustness and real-time performance, making it the most suitable choice for deployment in the Footskillz skill classification system.