**BIBE INTERNATIONAL CONFERENCE**

# FROM PEN TO PREDICTION: HANDWRITING-BASED ALZHEIMER'S DETECTION

Maria Boumpi [1], Kalliopi V. Dalakleidi[1 2], John Pavlopoulos[1 2]
[1]Department of Informatics, Athens University of Economics and Business, Greece
[2]Archimedes/Athena Research Center, Greece

# Overview

| | | | |
|---|---|---|---|
| Problem | Research Question | Related Work | Datasets |
| Tabular Classification | Image Classification | Fusion Model | Results |
| Erros Analysis | Discussion | Conclusion | Thank You |

# Problem

**Alzheimer's disease (AD)** is a progressive neurodegenerative disorder that affects memory, cognition, and daily functioning.

- Early detection is vital to slow progression and improve treatment outcomes.
- Current diagnostic tools (neuroimaging) are costly, invasive, and not scalable.
- Subtle handwriting patterns can reveal neuromotor and cognitive decline.
- Handwriting analysis offers a simple, non-invasive, and low-cost alternative for early screening.

# Research Question

Can visual handwriting signals and motion-derived features result in more effective early Alzheimer's diagnosis compared to traditional medical history information?

**OBJECTIVE 1**

Comparative analysis of clinical, tabular, and image-based handwriting data

**OBJECTIVE 2**

Development of a multimodal fusion model

**OBJECTIVE 3**

Evaluation of performance and robustness

# Related Work

### Handwriting Features

Classical ML models (SVM, RF, KNN) used kinematic handwriting features to detect cognitive decline.

### CNN-Hybrid Approaches

CNN-based methods combined spatial handwriting images with temporal motion data, improving sensitivity to cognitive decline.

### Explainable

Ensemble and SHAP-based studies identified key handwriting features, providing interpretable insights into cognitive decline.

### Digital Drawing Tests

Systems like DCTclock showed higher sensitivity to mild cognitive impairment compared to standard cognitive tests (e.g., MMSE).

# DARWIN Dataset

Contains handwriting data collected from 174 participants, including 89 patients diagnosed with Alzheimer's disease (AD) and 85 healthy controls (HC).

## Tabular Data

- 25 handwriting tasks (graphic, copying, memory, and dictation activities)
- Each of them exported 18 features (450 features)

| | |
|---|---|
| Pen-up time (ms) | 6,085 |
| Mean pen pressure | 1,851.08 |
| Task duration (ms) | 24,870 |

Table: Sample features for task 2

## Images

- 6 tasks ( 2, 3, 4, 5, 21, and 24)
- Join points (vertical, horizontal), retrace circles (6 cm and 3 cm),  a complex form, and draw a clock (11:05)
- 88 AD / 78 HC  likely due to consent limitations

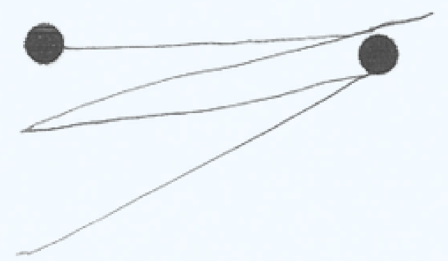


Figure: Join two points

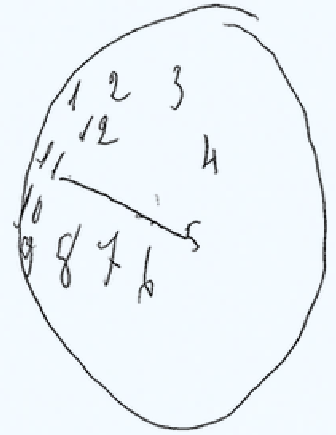


Figure: Draw clock (11:05)

Aligned handwriting tasks and participant data (88 AD / 78 HC) enabled consistent multimodal comparison and fusion.

# ADD Dataset

Contains clinical and demographic data for 2,149 individuals (760 with AD, 1,389 HC). Each record contains 34 attributes, including demographic, lifestyle, and clinical data.

- Key features: MMSE, ADL, memory complaints, family history

| Feature | Value |
|---|---|
| Family history of AD | **0** |
| Functional Assessment | **6.52** |
| MemoryComplaints | **0** |

Table: Sample features

Subsampled to 166 participants (88 AD, 78 healthy) to match the size and class balance of the DARWIN dataset. Not participant alignment.

# Tabular classification

## Step 1 – Model Selection

Six standard binary classifiers were used:

- SVM,
- Logistic Regression (LR),
- Random Forest (RF)
- Gaussian Naive Bayes (GNB),
- K-Nearest Neighbors (KNN),
- XGBoost (XGB)

## Step 2 –Training Setup

- Models were trained with five Monte Carlo cross-validation (80% train / 20% test).
- The mean accuracy and SEM were reported to indicate variability.

## Step 3 – Hyperparameter Tuning

Hyperparameters were optimized using:

- Grid Search,
- Optuna,
- Default configurations

## Step 4 – Class Balance

Stratified sampling was applied throughout to maintain balanced classes between AD and healthy participants.

# Results

**Clinical (ADD) data achieved slightly higher accuracy, likely due to its greater feature diversity.**

Best performance:

- ADD → XGB: 83.53 % ± 3.44
- DARWIN → RF: 83.03 % ± 1.18

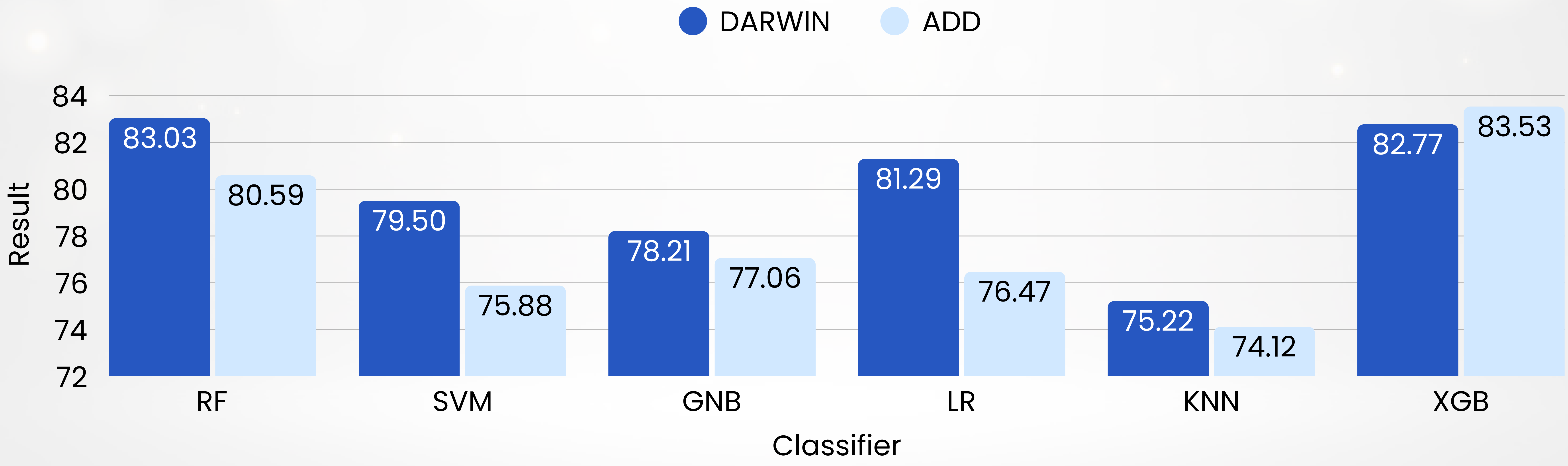



Table: Mean accuracy of classifiers on DARWIN and ADD

# Image classification

### Step 1 – Model

- A fine-tuned Swin Transformer was used.
- Images were resized to 224×224.
- Normalized using ImageNet statistics.

### Step 2 – Data Split

Data was divided into 80% training/validation (72% training, 8% validation) and 20% testing, matching the tabular setup.

### Step 3 – Training Setup

- Training used AdamW (lr = 5e-5), cosine annealing, and cross-entropy loss with label smoothing ($\varepsilon = 0.1$).
- Images were shuffled each epoch; validation/test sets stayed fixed.
- Early stopping (patience = 10) prevented overfitting.

### Step 4 – Evaluation

- The best model per seed (selected by lowest validation loss) was evaluated on the test set.
- Performance was averaged over Monte Carlo runs (seeds 42–46), reported as mean accuracy ± SEM.

# Results

It achieved an average accuracy of **80.02% ± 0.87**, ranging from 77.27% to 82.29%, slightly below the top tabular results (RF: 83.03%, XGB: 83.53%)

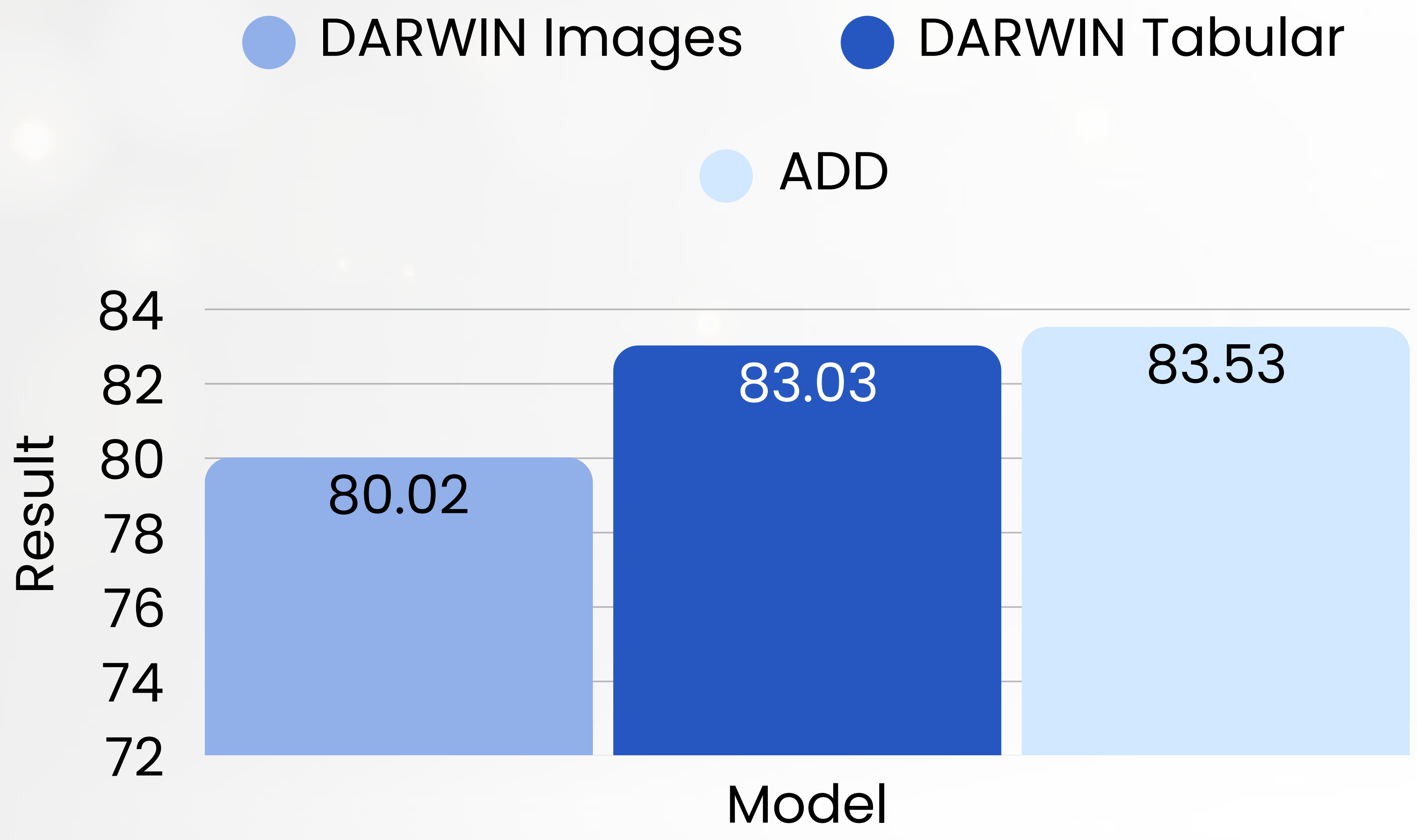



Table: Compare the mean accuracy of DARWIN handwriting images, DARWIN tabular, and ADD

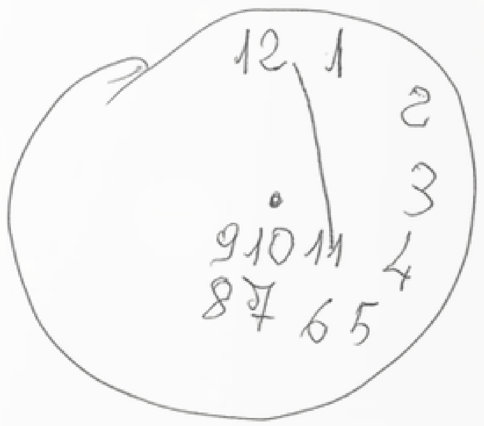


Figure: Clock Drawing AD

Figure: Clock Drawing Healthy

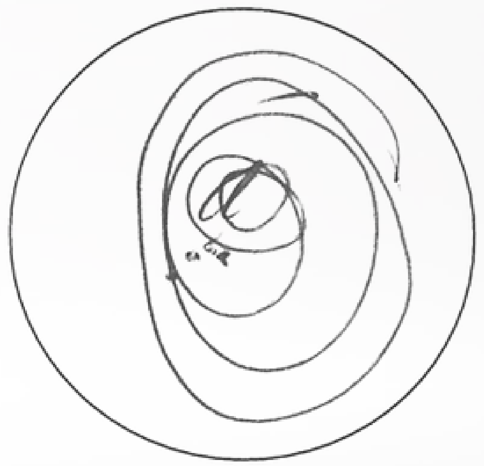


Figure: Circle Drawing AD

Figure: Circle Drawing Healthy

# Multimodal Fusion Model

### Step 1 – Model Design

- A fine-tuned Swin Transformer analyzed handwriting images.
- Random Forest processed tabular handwriting features (the best tabular performer).

### Step 2 – Fusion Strategy

Both models produced softmax-normalized probabilities, which were averaged to form the final prediction (late mean fusion), giving equal weight to each modality.

### Step 3 – Data Alignment

Participant IDs were matched across image and tabular data to ensure identical test splits and seed assignments.

### Step 4 – Evaluation

Performance was assessed through Monte Carlo cross-validation (seeds 42–46), reporting mean accuracy ± SEM to ensure stability and generalization.

# Results

The fusion model outperformed all single-modality models, achieving a mean accuracy of **89.15% ± 1.73.**
This highlights the strength of combining visual and tabular handwriting data, surpassing the performance of clinical models.

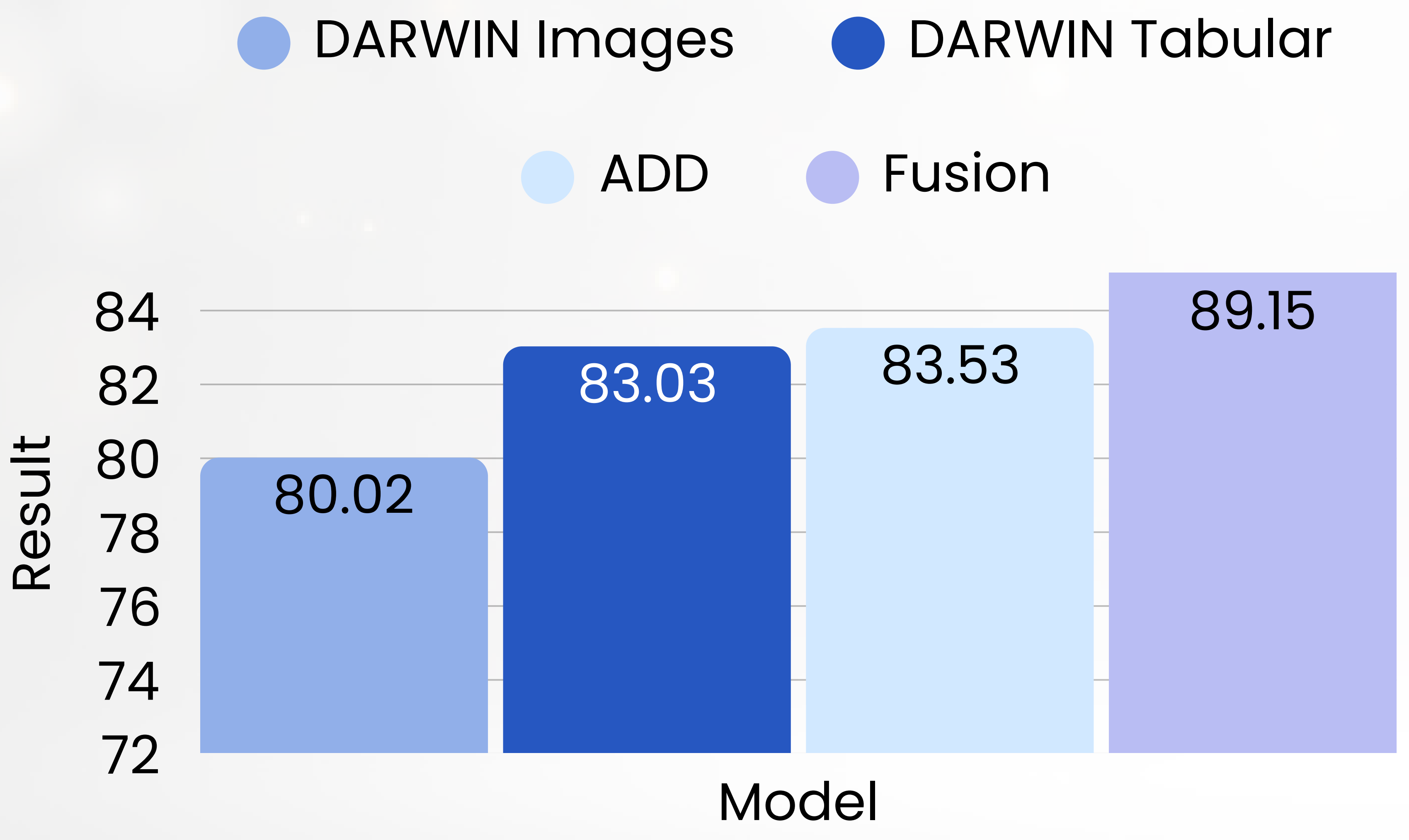



Table: Compare the mean accuracy of DARWIN handwriting images, DARWIN tabular, ADD, and Fusion model

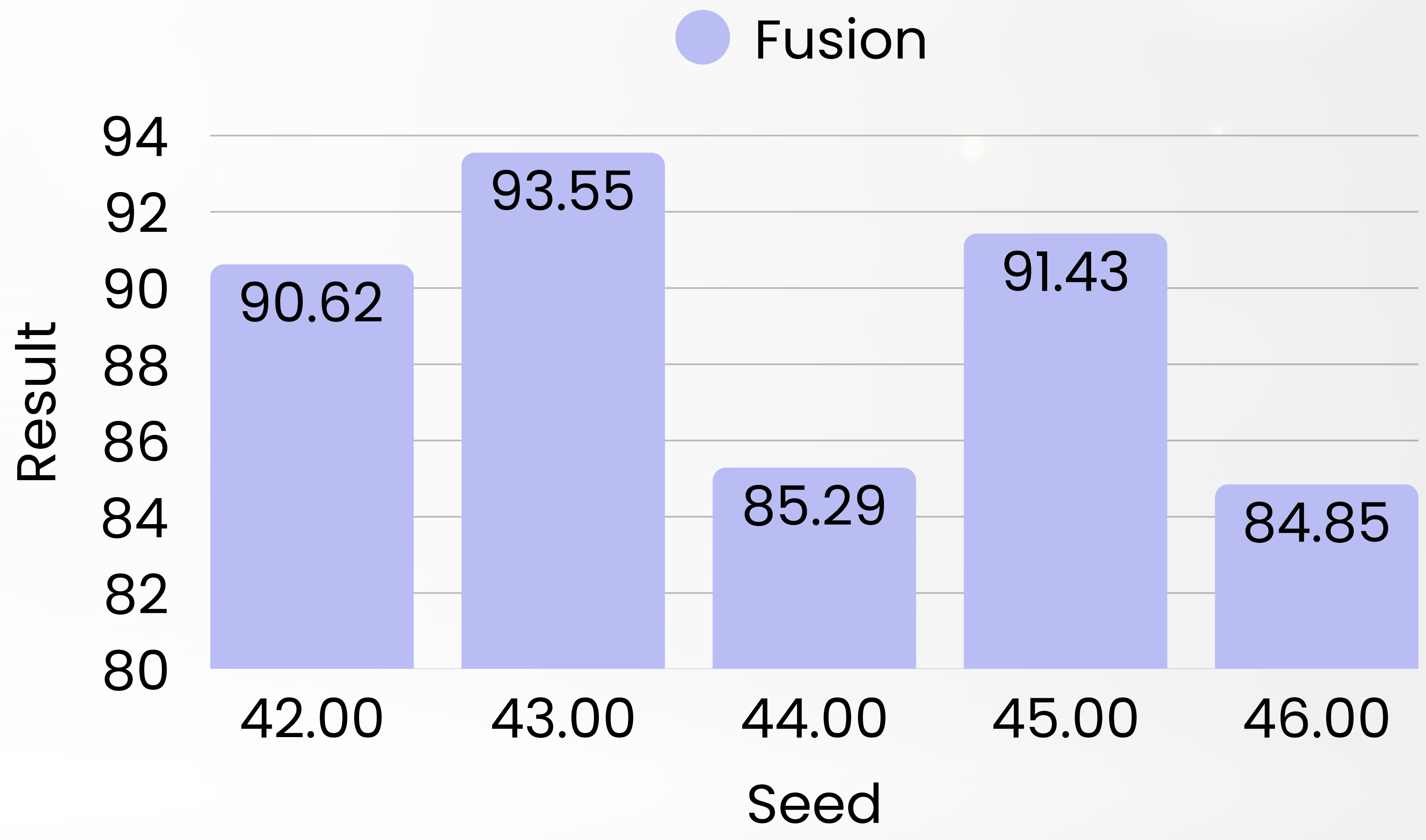
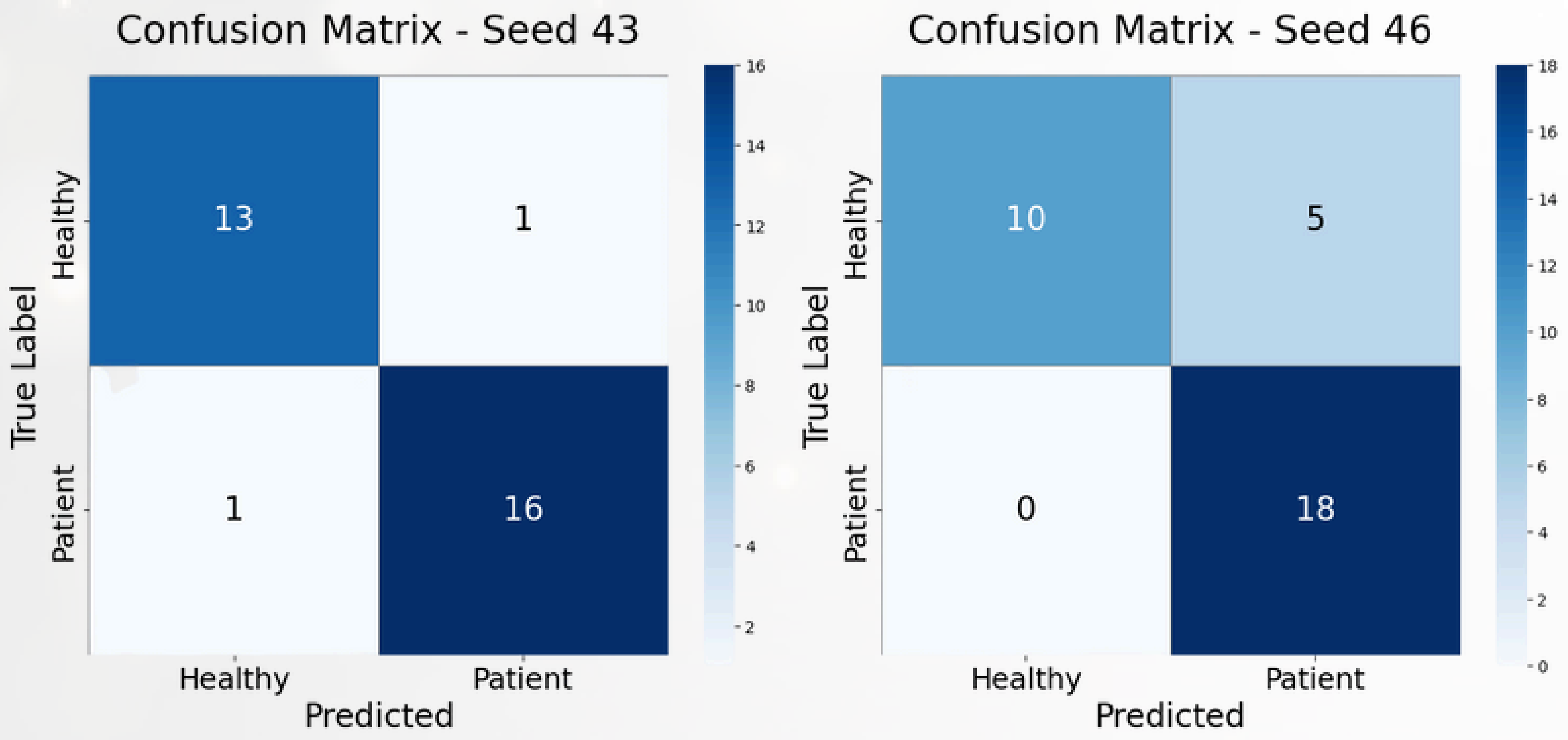



Table: Fusion model accuracy across five seeds (42–46)

# Error Analysis

## Confusion Matrices



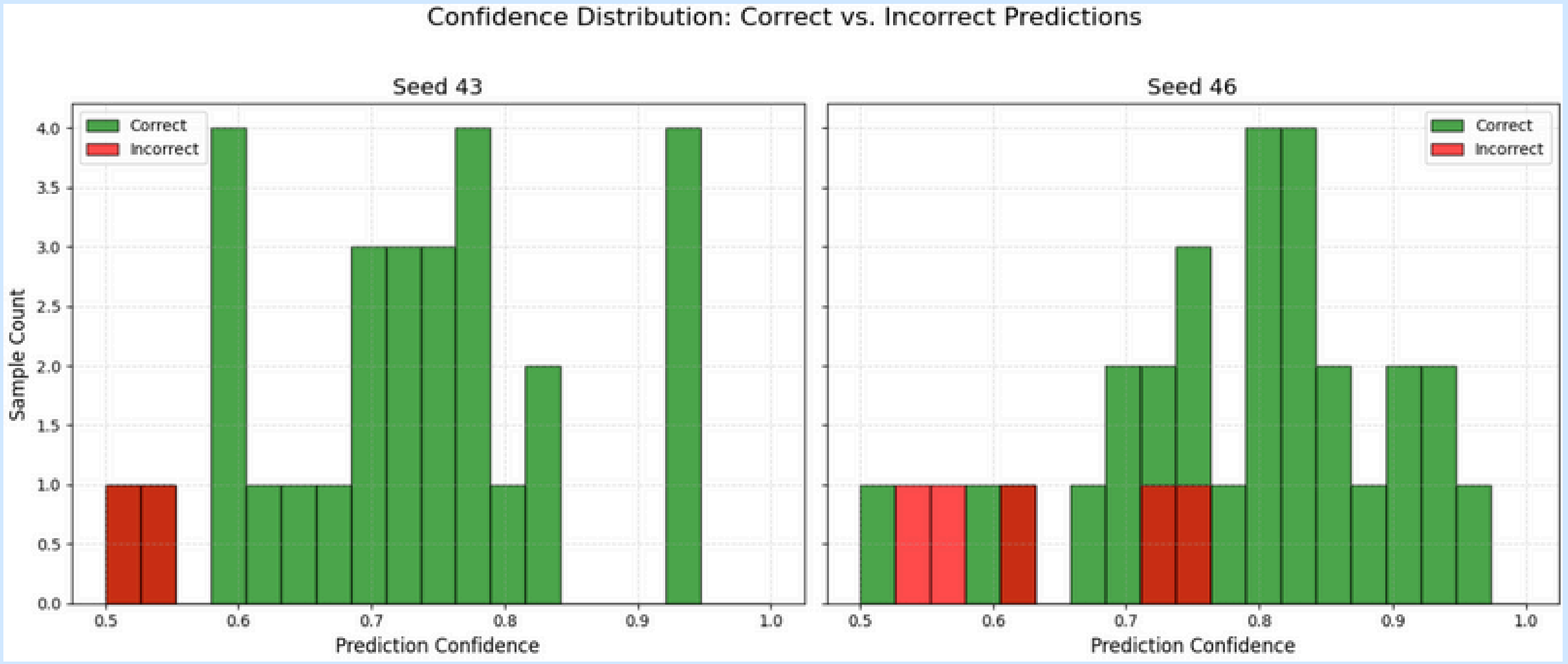

- **Seed 43:** Only 1 False Positive & 1 False Negative → balanced performance.
- **Seed 46:** No False Negatives, but 5 False Positives.

False negatives are riskier in screening → better to flag uncertain cases.

## Confidence Distribution Analysis

- **Seed 43**: Well-calibrated → correct predictions mostly within 0.65 – 0.95, errors only at low confidence.
- **Seed 46**: Overconfident → errors even above 0.7, showing poor calibration.

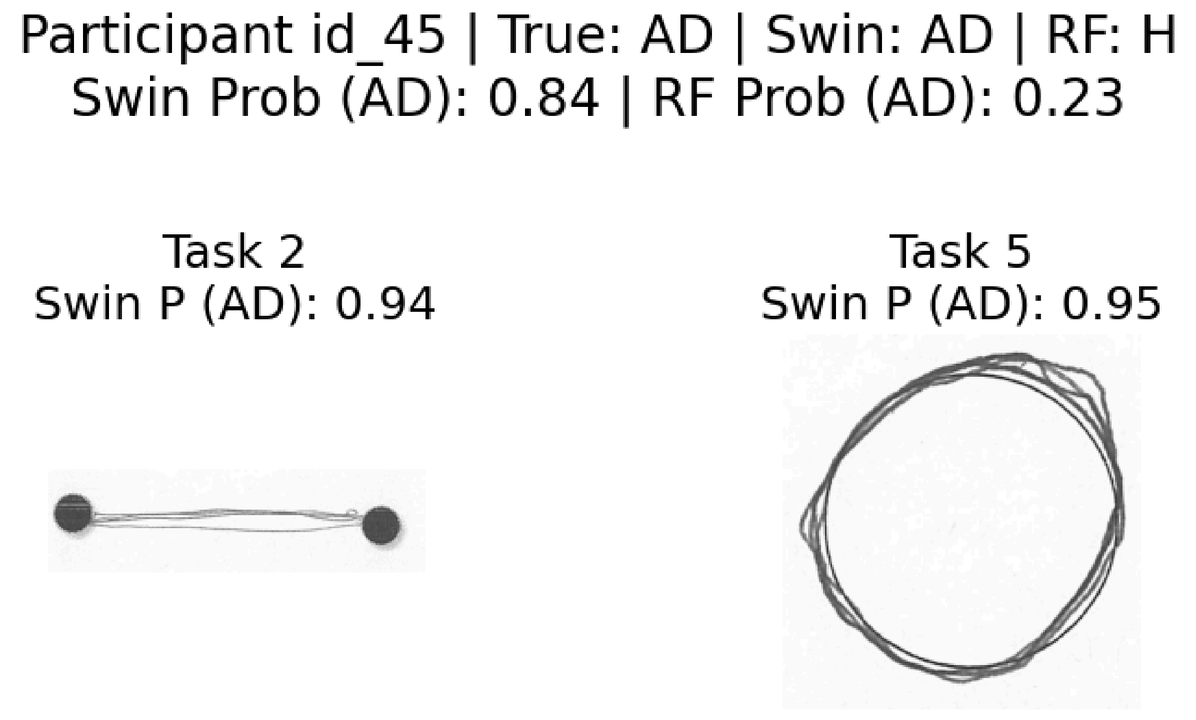
Setting a confidence threshold (≈ 0.7) can flag uncertain cases for expert review in clinical setting



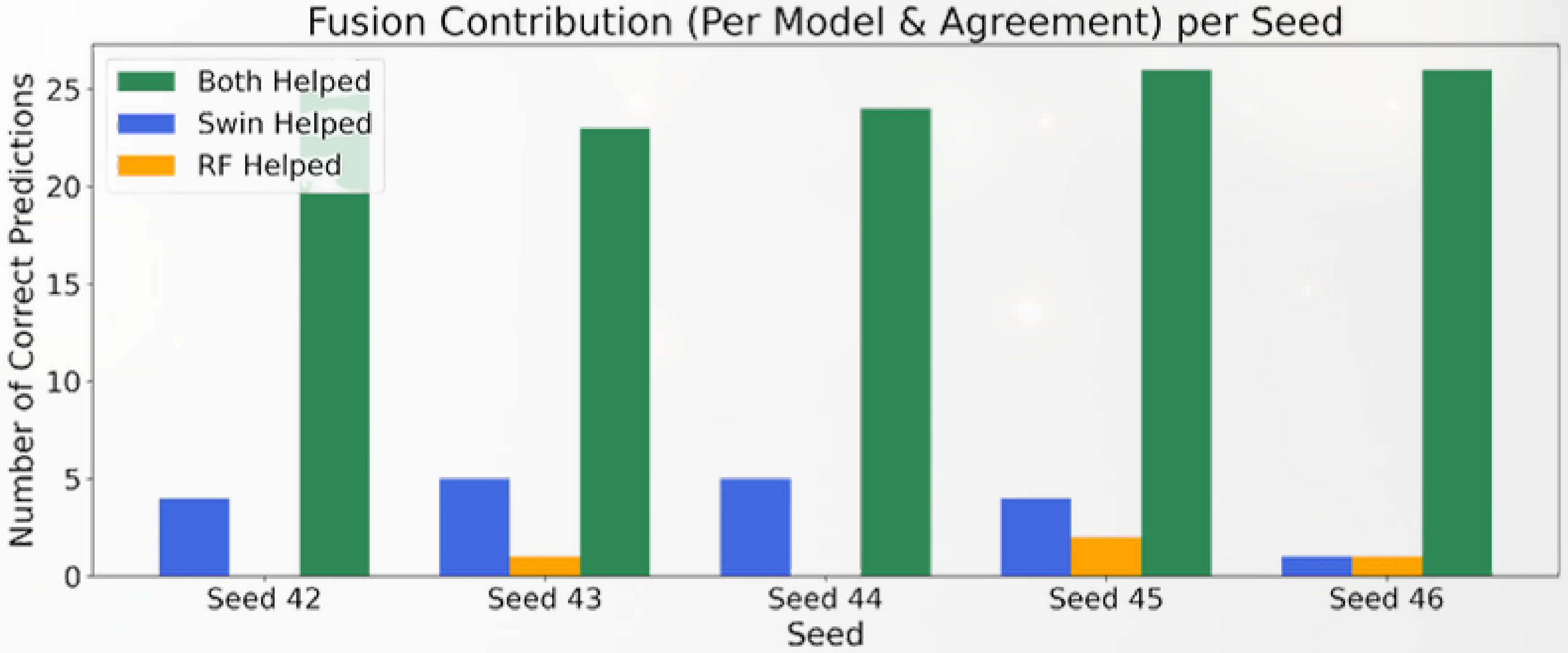
# Error Analysis

## Why the Image Modality Matters?

- **RF** misclassified the case as healthy due to normal-looking tabular features (e.g., low pressure variance, short completion time).
- **Swin** correctly detected AD from spatial distortions and unstable strokes




## Contribution Balance




- Swin contributed more to correct predictions in most runs (Seeds 43–45).
- **Seed 46**: Equal Swin/RF contributions → lowest accuracy (84.85%).
- **Seed 43**: Strong Swin dominance → highest accuracy (93.55%).

When models disagree: Swin's visual predictions are more reliable.

# Discussion

- **Dataset limitations:** The dataset included six simple drawing tasks (lines, spirals, shapes). It lacked linguistic or recall elements essential for cognitive assessment. Future work should include more cognitively demanding tasks (copying, recalling, dictation).

- **Comparative insight:** lower performance than prior studies (85–94%) due to smaller dataset size and reduced task diversity. The fusion model achieves a strong and competitive result with 89.15% accuracy.

- **Practical value:** Handwriting analysis is accessible, scalable, and cost-efficient, enabling remote early AD screening in low-resource settings.

# Conclusion

- **Approach:** Combined handwriting images and tabular features for early Alzheimer’s detection using the DARWIN dataset.
- **Key Result:** Late fusion achieved the best performance (89.15% ± 1.73), confirming the benefit of integrating visual and motion-based features.
- **Model Insight:** Handwriting images corrected errors from tabular data and proved robust across runs.
- **Practical Impact:** Handwriting is a low-cost, accessible, and scalable tool for early AD screening, even outperforming some traditional clinical data.

# THANK YOU