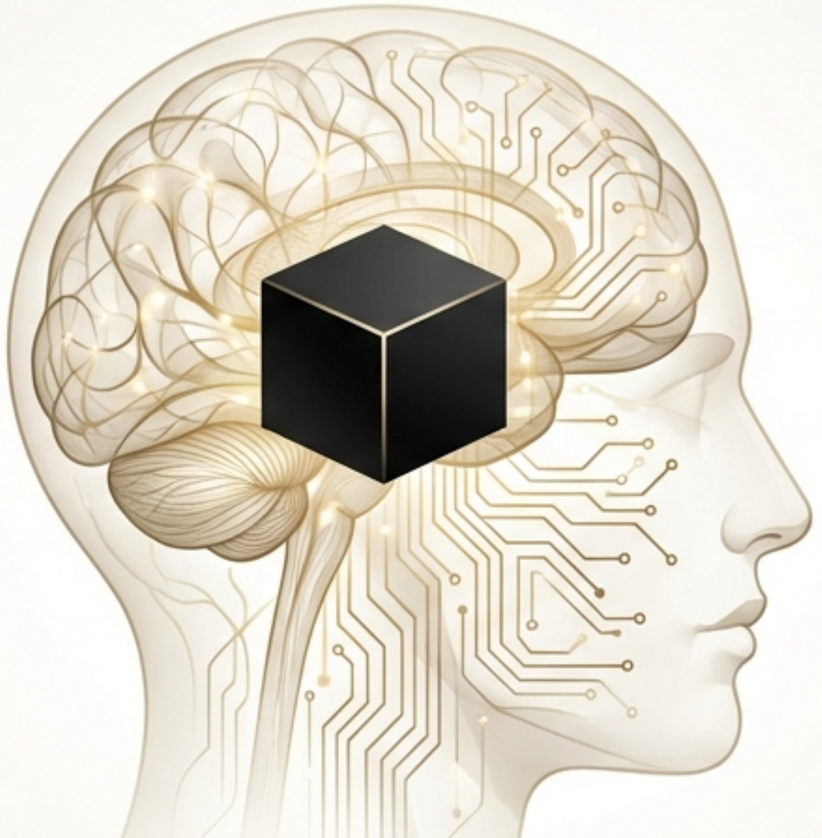
Cracking the Code of the Brain: A Blueprint for Interpretable AI in Neuroscience



A case study in classifying brain states from EEG signals during visual perception.

Source Serif Pro Regular

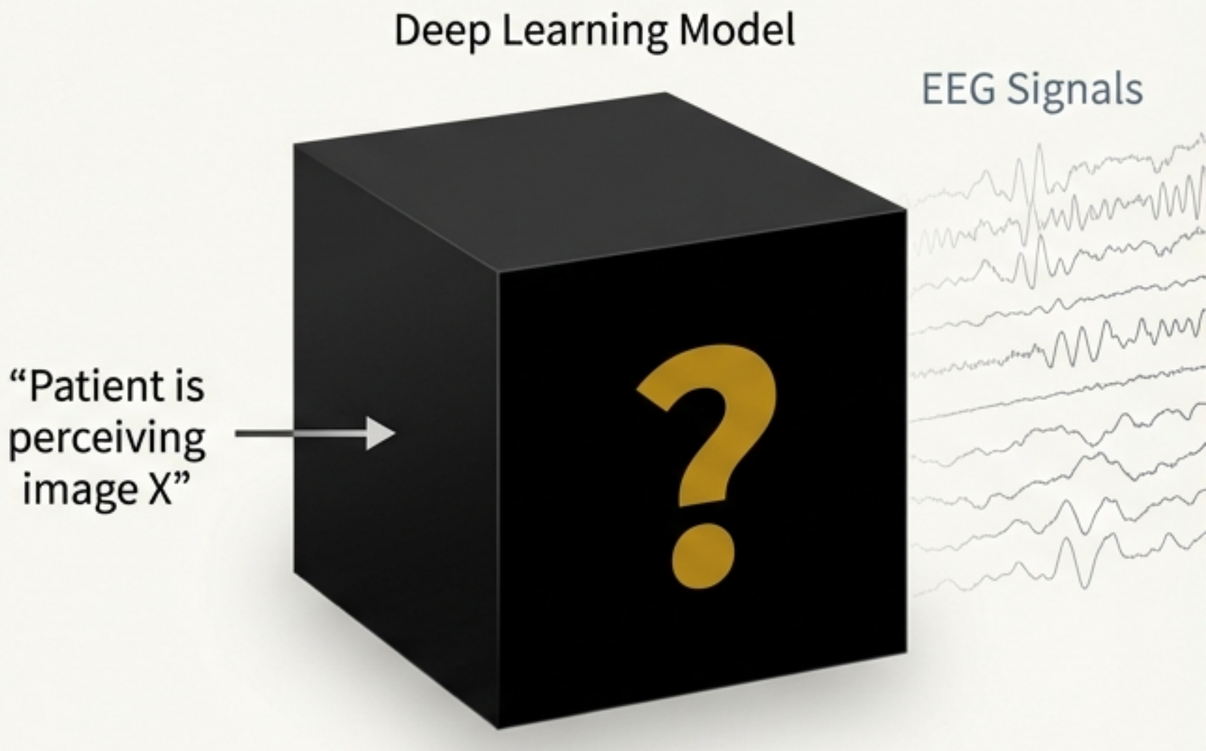
Based on Islam, R. et al. "Explainable Machine Learning Methods for Classification of Brain States during Visual Perception." *Mathematics* 2022, *10*, 2819.

The ‘Black Box’ Problem in Brain-Computer Interfaces

Modern AI can achieve high accuracy in classifying brain states, but often we don’t know *how* it makes its decisions. This lack of transparency is a major barrier to clinical adoption and true scientific understanding.

* For assistive medical systems, clinicians must understand and interpret the decisions made by AI.
* For BCI development, interpretability reveals which EEG features are truly driving the model, helping to build more robust and reliable systems.

* Our goal: To find a machine learning model that is not only accurate but also **explainable**.

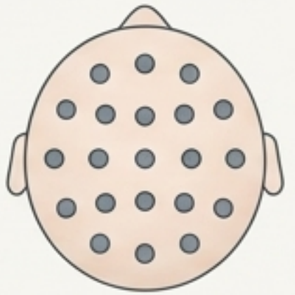


The Experimental Setup: Classifying Brain States from Visual Stimuli.

Core Task: Can we train a model to accurately identify the contrast level of an image a person is viewing, based solely on their EEG data?

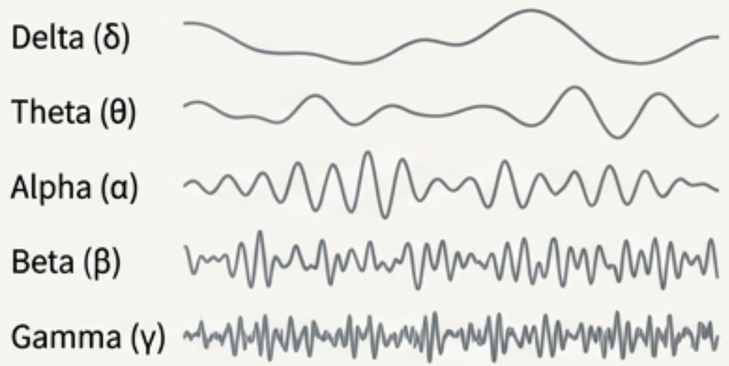
Data Source:

* **EEG**: 31-channel EEG recordings from 5 subjects.
* **Stimuli**: Subjects viewed two types of images (Necker Cube and Mona Lisa) at 10 different **contrast levels** (from 0.1 to 1.0).
* **Signal Processing**: Data was filtered into five standard frequency bands (Delta, Theta, Alpha, Beta, Gamma).



31 EEG Channels

0.1 10 Contrast Levels 1.0

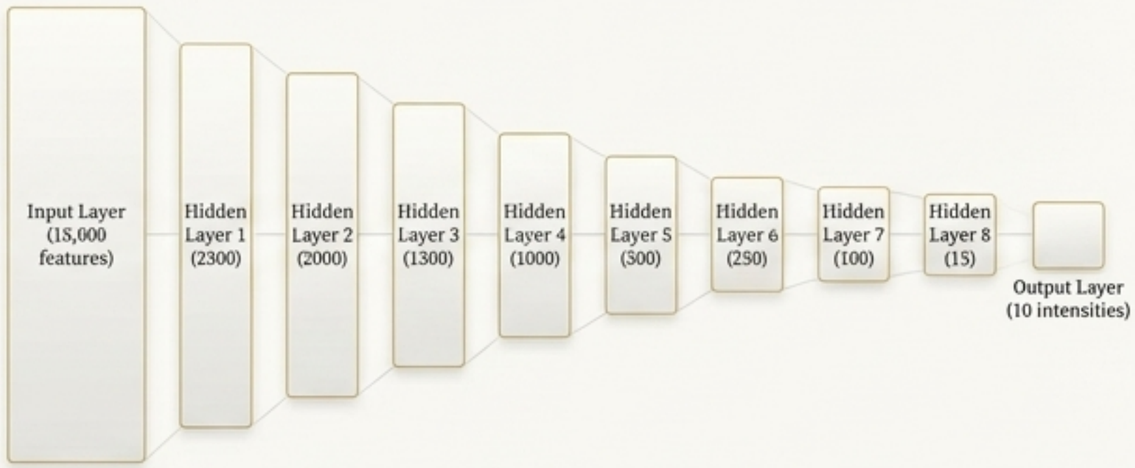


5 Frequency Bands

A First Attempt: A Standard Deep Learning Approach.

Methodology (Case Study I):

* A deep learning model with 8 hidden layers was constructed.
* **Activation Function:** Tanh.
* **Optimizer:** RMSprop.
* **Structure:** The data was split into training (80%) and testing (20%) sets. No validation set was used.



A Perfect Score...

...Or a Deceptive Trap?

Key Result: The initial model achieved **100% classification accuracy** on nearly every dataset.

100%

📈 **Mona Lisa Dataset (All Channels):** Accuracy = 100%* (δ, θ, α bands)

🧠 **Necker Cube Dataset (All Channels):** Accuracy = 100%* (δ, θ, α, β bands)

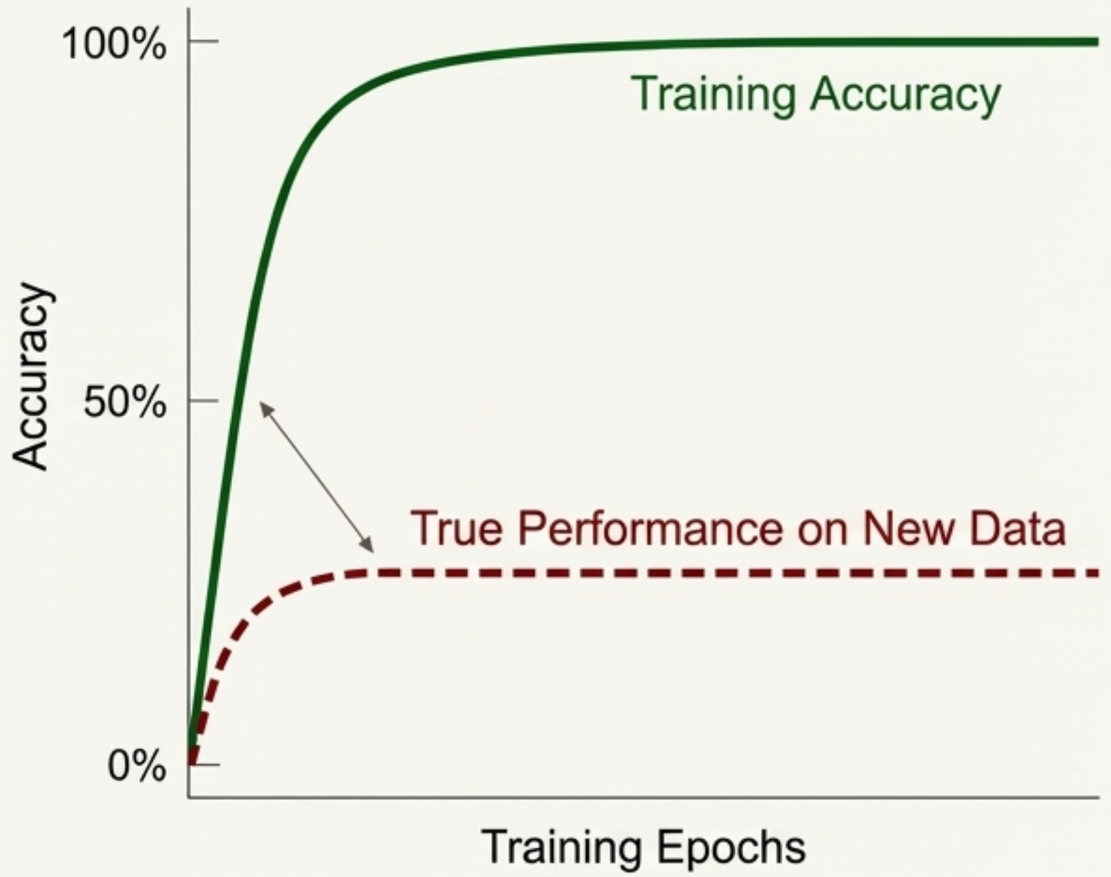
In machine learning, a perfect score is often a **red flag.** Did the model truly learn to classify brain states, or did it just memorize the test?

The Diagnosis: A Classic Case of Overfitting

Explanation: The 100% accuracy indicates the model didn't generalize. Instead, it learned the specific noise and artifacts of the training data perfectly. When faced with new, unseen data, its performance would be poor.

* The model achieved perfect accuracy on the test set because it was too similar to the training set.
* Without an independent **validation set** during training, there was no way to check if the model was generalizing or just memorizing.

The Insight: Achieving a high score isn't the real goal. The goal is to build a robust, generalizable model. This requires a more rigorous methodology.

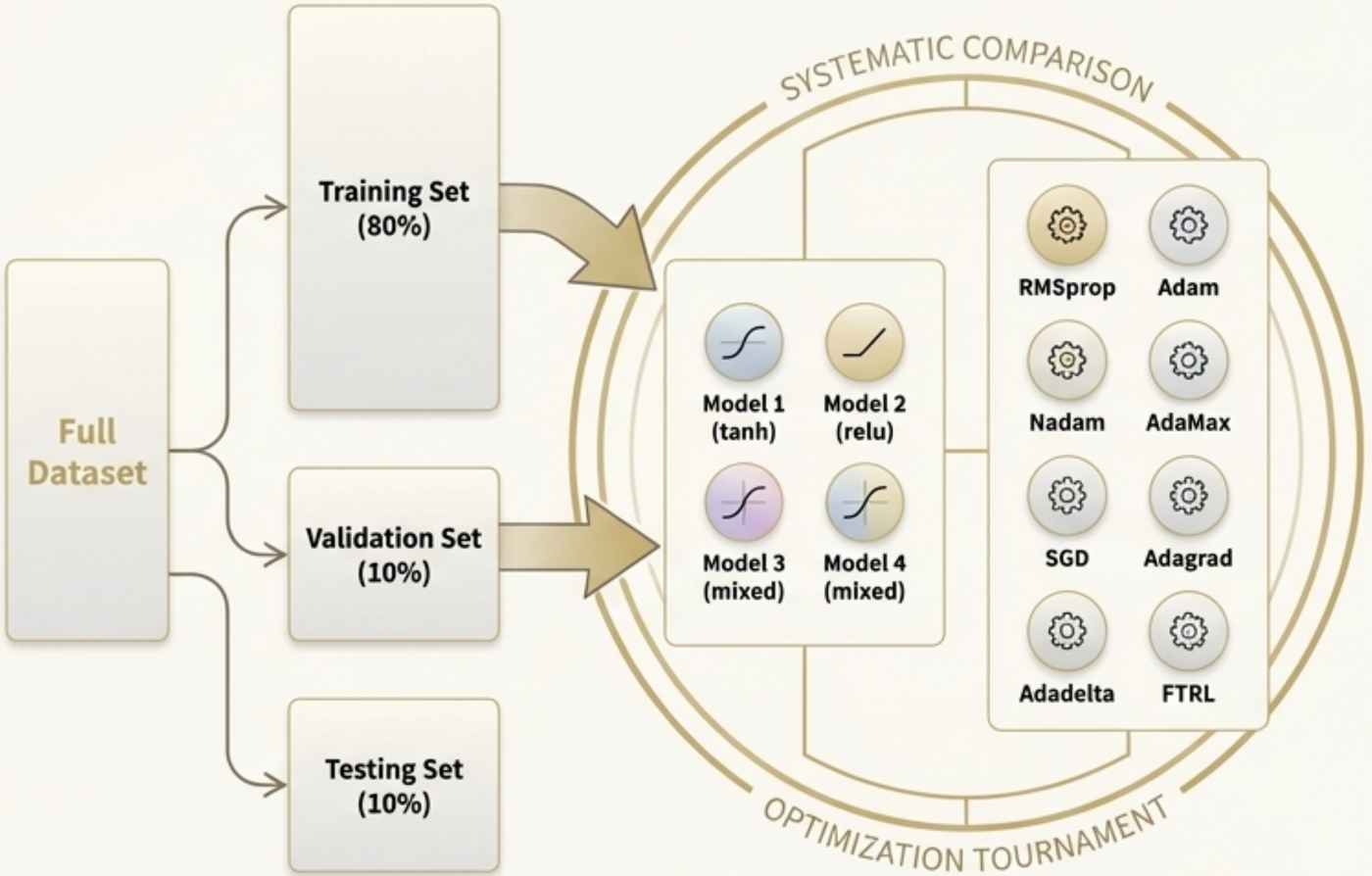


A More Rigorous Path: A Systematic Comparison of Models

New Methodology (Case Study II):

1. **Introducing a Validation Set:** Data was split into Training (80%), Validation (10%), and Testing (10%). The validation set is used during training to tune the model and prevent overfitting.

2. **Testing Multiple Architectures:** Four different models were built using combinations of `tanh` (tangent) and `relu` (linear) activation functions to see which worked better.
3. **A Head-to-Head Optimizer Showdown:** We systematically tested the performance of **eight different** optimization algorithms to find the most effective one for this specific task.



The Verdict: A Clear Winner Emerges from the Optimizers.

The choice of optimizer has a dramatic impact on performance. Some excel, while others fail completely.

Optimizer Performance Leaderboard

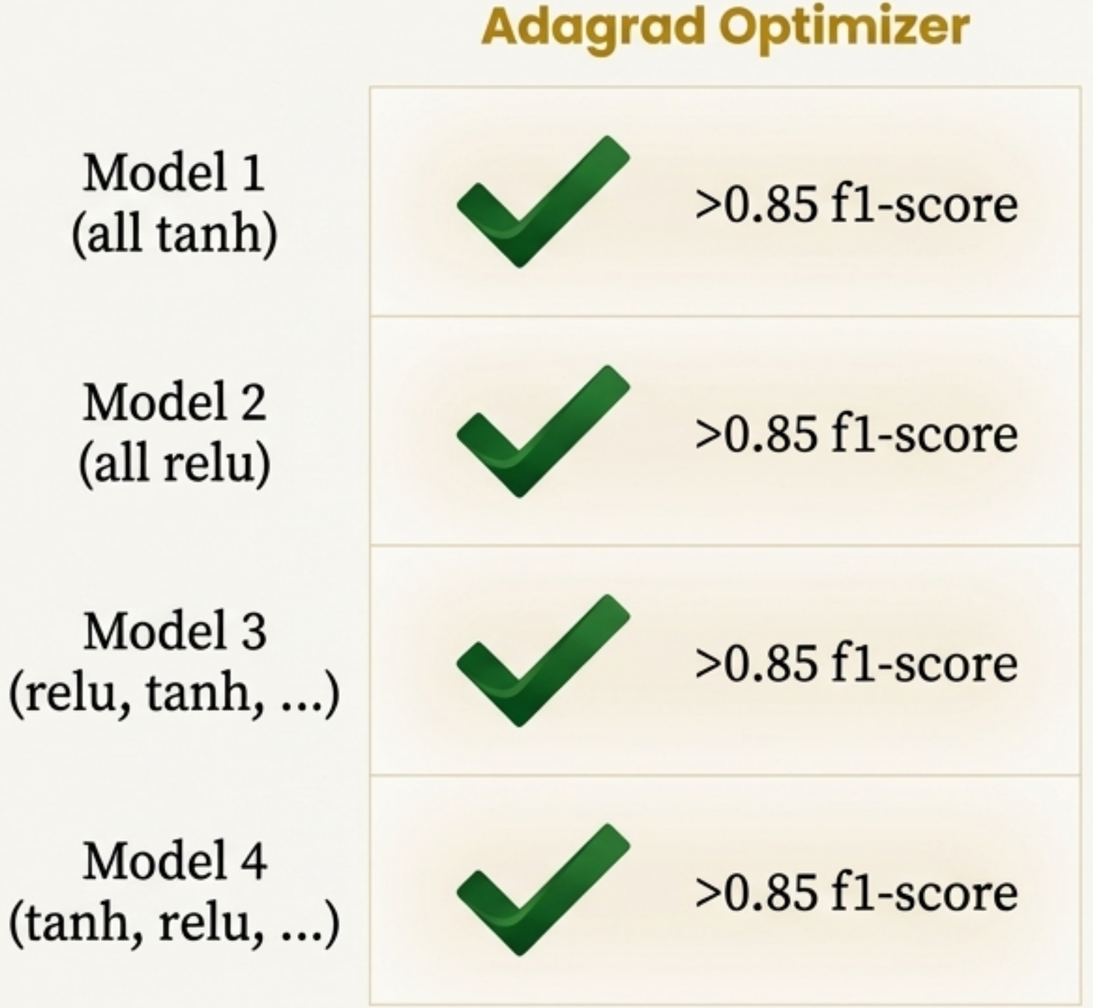
Why Adagrad Is the Top Choice for This Task

Key Insight:

Adagrad was the only optimizer that performed well for **both** linear (`relu`) and tangent (`tanh`) model architectures. This makes it a uniquely robust and reliable choice.

Supporting Evidence:

- **Behavior:** Consistently demonstrated ideal 'B1' behavior (normal accuracy and loss curves during training and validation).
- **Performance:** Achieved high f1-scores across both datasets and all frequency bands, including scores of 1.0 (100%) in some cases.



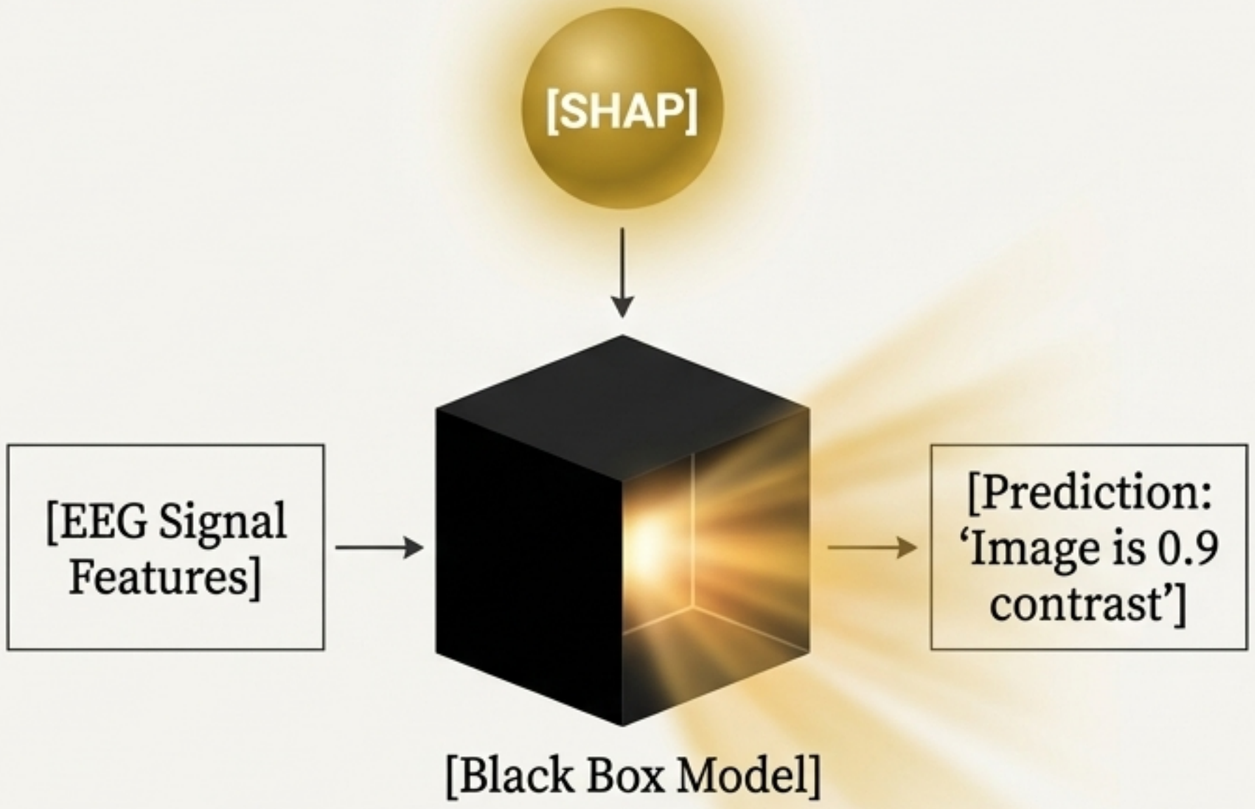
Beyond Accuracy: Shining a Light Inside the Black Box with SHAP.

The Challenge

We have a high-performing model (using the Adagrad optimizer), but we still don't know *why* it's making its decisions.

The Solution: SHAP (SHapley Additive exPlanations):

- SHAP is a technique from cooperative game theory that explains the output of any machine learning model.
- It calculates the contribution of each input feature to the final prediction. In essence, it shows how much each feature 'pushed' the prediction toward a certain outcome.



SHAP reveals how much each feature contributed to the final prediction.

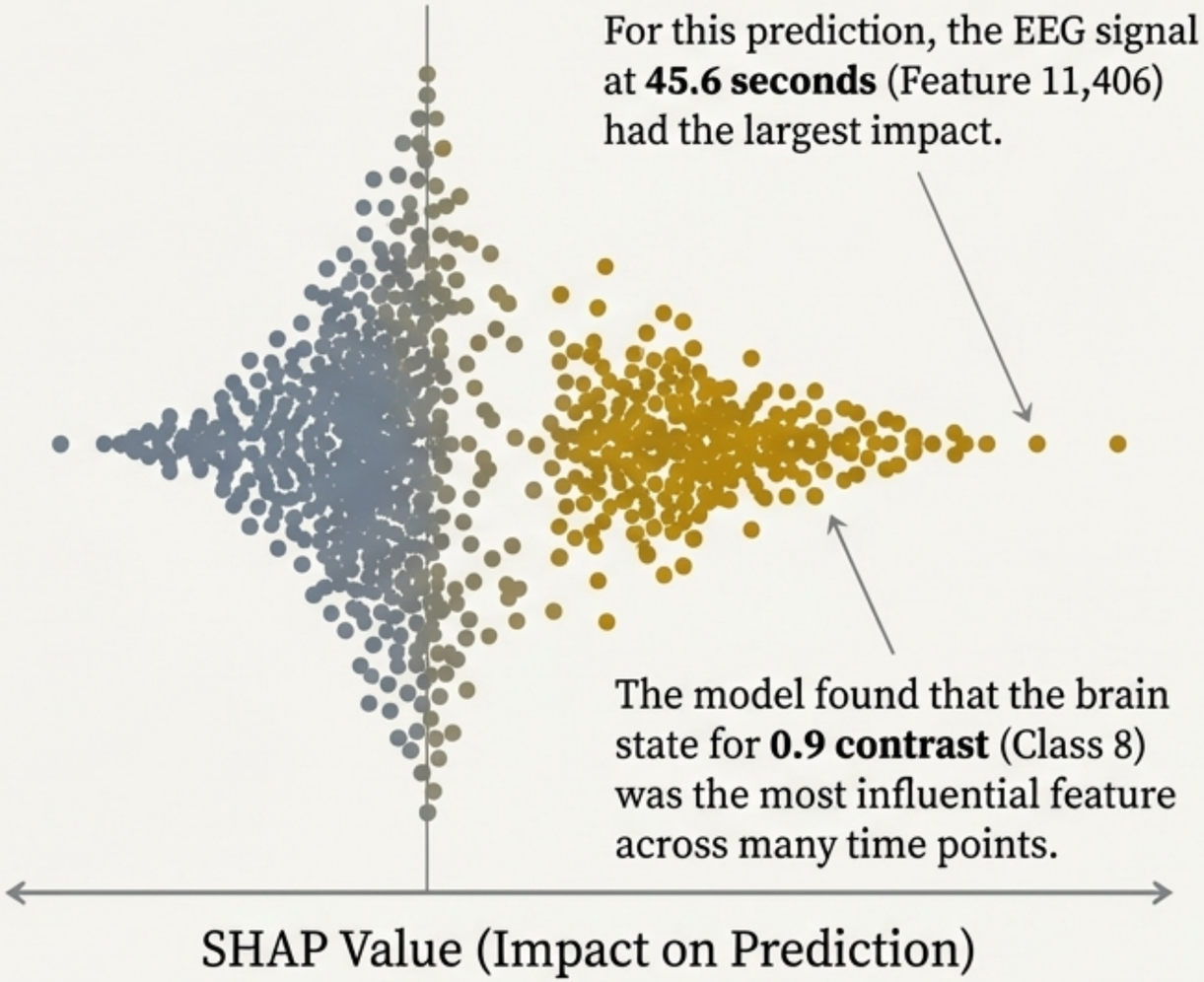
The Model’s Focus: What Features Drive Classification?

Key Insight from SHAP Analysis

- SHAP identifies the specific time points in the 60-second EEG recording that were most influential for the model's classification.
- It also reveals which image contrast levels had the strongest and most distinct neural signatures.

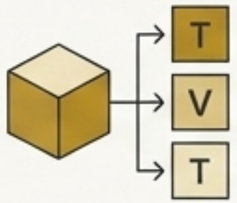
The Value

This level of detail allows researchers to connect the model's decision-making process back to known neurophysiological phenomena.



A Blueprint for Building Trustworthy and Interpretable BCIs

The Recipe for Success:

1 

Start with a Rigorous Framework: Always use a three-way data split (Train/Validation/Test) to prevent overfitting and ensure your model generalizes.

2 

Select the Right Tool for the Job: For EEG classification tasks of this nature, begin with the **Adagrad optimizer.** It is robust across different model architectures. Avoid FTRL.

3 

Validate, Then Interpret: Once you have an accurate model, use tools like **SHAP** to open the black box. Don't trust a model's accuracy until you understand the basis for its decisions.

True progress isn't just a high accuracy score; it's the rigorous journey to find a solution that is both accurate and trustworthy.