WiDS Datathon 2025

kaggle

Reading the 1st- and 2nd-place winning solutions

A professional review of the published winner writeups, with commentary on what is directly evidenced, what is implied, and what is transferable to future competitions.

Focus:
- Team Jenna (1st place)
- Team11 (2nd place)
- Transferable Kaggle lessons

Eric Benhamou

PSL UNIVERSITÉ PARIS
Dauphine UNIVERSITÉ PARIS
ENS ÉCOLE NORMALE SUPÉRIEURE
MINES ParisTech

Executive summary

1st place is representation-led

The clearest public clue from Team Jenna is that multidimensional functional MRI connectivity was decisive, especially for the Sex_F outcome. The signal story comes before the ensemble story.

2nd place is diversity-led

Team11 explicitly says validation improved by combining different models and different subsets of data. Their solution appears to win by controlled heterogeneity, not by a single dominant learner.

Transferable lesson

For similar competitions: build a strong representation first, pressure-test cross-validation second, and only then widen the search into model/subset ensembling.

How to read this deck

This review separates direct evidence from commentary. When the public writeup text is not fully machine-readable, the deck stays explicit about what is observed versus what is inferred.

Why this matters

The value is not only historical. These two solutions reveal the two most durable Kaggle levers for hard tabular-plus-structured-data problems: better representation and better diversity.

Bottom line: Team Jenna points to the feature frontier; Team11 points to the ensemble frontier.

What this deck covers

Direct evidence

- The competition asked teams to predict sex and ADHD diagnosis from functional brain imaging plus socio-demographic, emotional, and parenting information.
- The official WiDS workshop material states the evaluation used F1 score.
- Team Jenna highlighted multidimensional connectivity as key to Sex_F.
- Team11 highlighted ensembling across models and data subsets.

Commentary lens

- What kind of solution philosophy does each quote imply?
- Which decisions are likely representation choices versus ensemble choices?
- Which tactics are robust enough to transfer to a new Kaggle problem?
- Which conclusions would be overreach and should be avoided?

Output

A forensic, competition-oriented reading of the two winner writeups.

Constraint

Not every writeup detail is publicly machine-readable in this environment.

Approach

Use only what is evidenced, then comment carefully and transparently.

This is intentionally not a hallucinated reverse-engineering of exact code; it is an evidence-led solution review.

Competition recap

Task

Predict both an individual's sex and ADHD diagnosis.

Data

Functional brain imaging data of children and adolescents, plus socio-demographic, emotions, and parenting information.

Evaluation

The official WiDS workshop materials describe F1 score as the competition metric, making class balance and threshold behavior materially important.

Baseline workflow implied by official workshops

Merge the categorical columns with the functional connectivity matrices, encode categories, explore missingness, and clean before modeling.

Why the leaderboard is tricky

This is not a plain tabular contest. It blends high-dimensional brain connectivity information with lower-dimensional metadata, so representation and validation both matter more than brute-force model churn.

The official workshops already hint at the right order of operations: merge correctly, clean aggressively, evaluate under the real metric.

Why this was a hard leaderboard

Mixed modalities

High-dimensional connectivity features and low-dimensional metadata must be made to coexist without overwhelming one another.

Asymmetric targets

The public winner quote from Team Jenna suggests that the sex target exposed stronger structure in connectivity than the ADHD target did.

Metric sensitivity

Under F1, small threshold or class-balance mistakes can erase gains that looked convincing during training.

Validation quality dominates

When data are structured, noisy, and partially redundant, the split strategy becomes part of the model. Weak validation can reward the wrong experiments.

Late-stage gains are plural

Such competitions are rarely won by one magic trick. The signal usually arrives from a stack of disciplined improvements: representation, folds, model families, subsets, and blend control.

The two public winner messages fit this pattern perfectly: one points to representation, the other to ensemble diversity.

Public evidence base for the review

Winner quote: Team Jenna

“Our biggest takeaway was the knowledge of multidimensional functional MRI connectivity data – key to improving the prediction performance for the Sex_F outcome.”

Winner quote: Team11

“Ensemble from different models and subsets of data helps improve validation performance.”

Panel context

The official winner panel is presented as a discussion of results, model-building process, struggles, and key insights.

Official workshop context

Preprocessing guidance explicitly discusses encoding categorical variables, merging them with connectivity matrices, and handling NaN values before model development.

Review posture

These clues are enough to read each solution philosophy, even if they are not enough to claim the full exact code path. So the deck emphasizes decision logic over unsupported implementation detail.

The advantage of this approach is reliability: every strong claim stays anchored to public evidence.

01

How to read a winning solution

Separate observation, interpretation, and transfer.

A simple lens for forensic reading

Representation

How did they represent the data so that useful signal survived?

Validation

How did they decide an experiment was truly better?

Diversity

Where did they create non-redundant variation across models or subsets?

Blend

How did they convert several good ideas into one final submission?

This lens is more useful than chasing exact hyperparameters, because it exposes why a solution generalized.

What counts as evidence, and what does not

Strong evidence

Exact winner quotes, official workshop descriptions, and competition statements. These anchor the core claims in this deck.

Reasonable inference

If a winner highlights representation or ensembling, we can comment on the solution philosophy that such a statement implies.

What to avoid

Claiming exact model classes, exact fold recipes, or exact blending weights without direct evidence would be overreach.

Professional standard

A good postmortem is still useful even when it declines to invent detail. Precision beats false certainty.

Official workshop baseline: the default pipeline everyone started from

Merge

Merge participant-level metadata with the functional connectivity matrices.

Encode

Encode categorical variables before model training.

Clean

Address NaN values and other quality issues during exploratory analysis.

Why this matters for reading the winners

The top teams were almost certainly not winning because they skipped fundamentals; they won by extending this baseline more intelligently.

Implication

When the public winner clues diverge, they are showing where they pushed beyond the common starting point: richer representation versus richer ensembling.

Public winner solutions should be read as improvements on the official workflow, not as replacements for it.

02

1st place: Team Jenna

A representation-first reading of the winning solution.

Team Jenna in one sentence

“Our biggest takeaway was the knowledge of multidimensional functional MRI connectivity data – key to improving the prediction performance for the Sex_F outcome.”

Team Jenna, WiDS Global Datathon 2025 Winners

Plain-language reading

Their winning edge appears to come from how they represented the connectome, not from a generic metadata-only pipeline.

Why this matters

The quote is unusually specific. It names both the decisive information source and the target where that source paid off.

If you only keep one line from the 1st-place solution, keep this one.

Reading Team Jenna as a representation-first solution

Preserve structure

“Multidimensional connectivity” suggests they resisted collapsing the brain graph into one crude flat summary too early.

Use multiple views

A winning representation often combines raw blocks, aggregated summaries, and target-aware transformations of the same source.

Make metadata complementary

Metadata likely helped, but the quote implies that it did not substitute for the connectivity signal.

Commentary

This is the fingerprint of many first-place solutions in hard structured-data competitions: better information geometry before better model complexity.

Why that approach fits this competition

Connectivity is native signal

The competition itself is framed around functional brain imaging, so the winner's emphasis on connectivity is conceptually aligned with the data-generating process.

Sex_F stood out

Team Jenna names the Sex_F outcome specifically, which implies target-level diagnostics rather than a one-size-fits-all treatment of both outputs.

Representation can beat model swapping

If the signal is trapped in a poor view of the connectivity data, trying ten more learners usually does less than fixing the view.

Comment on ADHD

The absence of an equally strong public quote about ADHD is itself informative: the noisier target may have benefited less from pure connectivity representation.

Practical lesson

In multi-target competitions, do not assume every target wants the same features, thresholds, or calibration logic.

Team Jenna's public statement is a reminder to diagnose the problem target by target, not only submission by submission.

What the Team Jenna experiment loop likely looked like

Connectivity views

Engineer several non-equivalent views of the connectome rather than one monolithic table.

Merge metadata

Join demographic and behavioral variables after the connectivity representation is credible.

Diagnose by target

Measure which output improves, which plateaus, and where F1 is being won or lost.

Choose robust learners

Only after the feature space behaves well does model choice become decisive.

This slide is commentary, not a claim of exact implementation detail.

What I would borrow from Team Jenna immediately

Feature view discipline

Create several principled views of the connectivity data and compare them under the same folds.

Target-wise analytics

Inspect the two outputs separately instead of celebrating only the aggregate score.

Representation before tuning

Do not spend your best hours on hyperparameters while the data representation is still weak.

Experiment hygiene

Lock the folds, log every representation change, and evaluate whether gains persist across all folds—not only in one lucky split.

Exit criterion

Move on to larger ensembles only after you have one base pipeline whose improvements feel causally understood.

This is the strongest anti-chaos lesson from a representation-led winner.

What can go wrong if you copy Team Jenna too literally

Feature sprawl

More views of the connectome can become a search explosion if they are not organized and validated cleanly.

False structure

High-dimensional brain features can create persuasive but unstable correlations if the fold design is weak.

Target imbalance

A representation that helps one target substantially can still under-serve the other if the loss, threshold, or blend is not checked carefully.

Comment

Representation-first is powerful, but only if you can distinguish true structural signal from fancy feature engineering theater.

03

2nd place: Team11

A diversity-first reading of the runner-up solution.

Team11 in one sentence

“Ensemble from different models and subsets of data helps improve validation performance.”

Team11, WiDS Global Datathon 2025 Winners

Plain-language reading

Their edge appears to come from controlled diversity: several competent models, trained on meaningfully different views of the problem.

Why this matters

This is a classic late-stage competition move. Once single models saturate, diversity becomes the remaining source of lift.

The key phrase here is not just “ensemble”; it is “different models and subsets of data.”

Reading Team11 as a diversity-first solution

Model diversity

Different learners make different mistakes. An ensemble only helps when those mistakes are not too correlated.

Subset diversity

Different subsets of data can force models to specialize, stabilize, or surface signal that a single global fit smooths away.

Validation-led selection

Because Team11 mentions validation explicitly, their blending decisions were likely governed by fold behavior rather than public leaderboard excitement.

Commentary

This is what strong runner-up and winning solutions often look like once the single-model frontier has narrowed: not deeper tuning, but wider controlled diversity.

What “subsets of data” can mean in practice

Row subsets

Different training slices by fold, resampling strategy, or cohort composition.

Feature subsets

Separate models for connectivity blocks, metadata blocks, or carefully filtered feature families.

Target subsets

Different modeling emphasis for sex versus ADHD, or different calibration per output.

Hybrid

A mix of the above.

Important caveat

The exact Team11 subset recipe is not public in the materials we could reliably inspect, so this slide lists the credible interpretations rather than pretending certainty.

Why this strategy works well in a hidden-leaderboard competition

Variance reduction

Averaging across different models usually reduces the chance that one brittle fit dominates the final score.

Coverage of mixed signals

Different learners can exploit different parts of the problem: dense metadata patterns, sparse connectivity structure, or target-specific behavior.

Protection from over-reading the LB

If model inclusion is controlled by cross-validation, the blend is less likely to chase public-board noise.

Subtle advantage

Ensembling also exposes whether an apparent breakthrough is real. If a model only shines alone but damages the blended OOF behavior, it is probably not robust.

Reading Team11

Their quote reads like the voice of a team that trusted validation discipline more than single-model charisma.

This is the runner-up pattern you should respect: deliberate diversity, selected by folds.

What I would borrow from Team11 immediately

Track OOF complementarity

Before blending, check whether candidate models make genuinely different out-of-fold errors.

Build from a shortlist

Blend only models that are individually credible. Weak models rarely become strong because they are averaged.

Create purposeful subsets

Subset design should reflect a hypothesis—feature family, sample slice, or target behavior—not randomness for its own sake.

Use validation as the gatekeeper

Decide blend membership and weights using OOF logic first; use the public leaderboard only as a weak secondary signal.

Stop rule

If every added model makes the ensemble more complex but not more stable, the blend has gone too far.

> Ensembling is not a bag of models; it is a controlled portfolio of non-redundant bets.

What can go wrong if you copy Team11 too literally

Correlated clones

Five versions of the same learner trained on nearly the same data may look diverse in a spreadsheet but not in reality.

Blend-before-understanding

If the base pipelines are weak or leaky, ensembling them only hides the diagnosis problem.

Leaderboard chasing

A complex ensemble can be overfit to public-board feedback even while appearing “sophisticated.”

Comment

The lesson from Team11 is not “add more models”; it is “add more independent information, then keep only what validates.”

04

1st place vs 2nd place

Where the two solutions converge, and where they clearly differ.

Comparative matrix

Dimension	Team Jenna	Team11	Implication
Winning clue	Connectivity representation	Model/subset ensembling	Different frontier of advantage
Primary emphasis	Signal extraction	Controlled diversity	Both matter; order matters more
Validation role	Diagnose target-specific gains	Select ensemble membership	OOF quality is central in both
Transferability	Very high for feature engineering	Very high for late-stage blending	Best combined sequentially
Copy risk	Feature sprawl	Ensemble clutter	Need discipline either way

> The two winners are not contradictory; they illuminate different layers of the same winning stack.

Where the two solutions converge

They respect the real signal

Neither public quote sounds like blind model shopping. Both point to the structure of the data itself.

They are validation-conscious

Team11 says it directly. Team Jenna implies it by naming the target where gains materialized.

They are multi-view thinkers

One through multiple connectivity views; the other through multiple models/subsets.

Common principle

Top Kaggle solutions on difficult data almost always win by adding orthogonal information, not just by searching harder inside one narrow pipeline.

Practical reading

If your own solution has only one view of the data and one family of model, it is probably under-developed.

The common denominator is disciplined pluralism.

Where the two solutions diverge

Team Jenna

The center of gravity is earlier in the pipeline:

- represent the connectome better
- surface the strongest target-specific signal
- let model choice capitalize on that representation

This reads like a solution that wins by seeing the data correctly.

Team11

The center of gravity is later in the pipeline:

- cultivate several competent models
- vary training subsets or feature subsets
- select the blend by validation behavior

This reads like a solution that wins by combining strengths cleanly.

One winner pushes the feature frontier; the other pushes the ensemble frontier.

The combined recipe the leaderboard seems to reward

1. Build signal

Engineer the best representation you can, especially for the highest-signal modalities.

2. Validate honestly

Freeze folds and use them as the primary truth source.

3. Diversify

Add model or subset diversity only when it is truly non-redundant.

4. Blend conservatively

Keep only additions that improve OOF reliability, not only public LB mood.

This is the most actionable synthesis of 1st place plus 2nd place.

What these two solutions say about leaderboard dynamics

The final gains were probably narrow

When both representation and ensembling matter, the leaderboard is usually decided by several small compounding edges rather than one giant leap.

Public LB was not enough

Team11's wording centers validation, which is exactly what strong teams say when they do not trust the public board blindly.

Single-model heroics were unlikely

The winning clues point away from a one-model fairy tale and toward layered, portfolio-like construction.

Competition lesson

On difficult Kaggle problems, the winning question is often not "Which model wins?" but "Which sequence of disciplined decisions compounds most reliably?"

05

Transfer to future competitions

What to copy, what to adapt, and what not to over-generalize.

What transfers directly to the next Kaggle competition

Representation-first thinking

Before tuning, ask which view of the data preserves the strongest domain signal.

Fold discipline

Freeze a credible CV design early and use it as the decision engine for the project.

Target-wise diagnostics

Do not assume all targets respond equally to the same features, thresholds, or blends.

Purposeful diversity

Only ensemble models or subsets that bring genuinely different information or error patterns.

Sequential build order

Strong base pipeline first; broad blend second. That ordering transfers almost everywhere.

These are the durable parts of the winners' playbook.

What does not transfer automatically

Domain-specific representations

A connectivity feature engineering trick for fMRI does not automatically map to wildfire risk, retail demand, or satellite imagery.

Subset definitions

The right training subsets depend on the data topology and leakage risks of the specific competition.

Target asymmetry

Some competitions have one dominant target, others have balanced outputs, and others have ranking or regression objectives.

Rule of thumb

Copy the decision principles, not the surface tactics. The winners teach a way to think, not a universal code template.

Implementation checklist if you were to reproduce the spirit of these solutions

Phase 1 — representation

Build 3–5 serious feature views. Compare them on fixed folds. Keep a simple benchmark alive.

Phase 2 — validation

Audit split stability, target-wise behavior, and threshold sensitivity under the real metric.

Phase 3 — diversity

Introduce non-redundant learners or subsets only after one base pipeline is trustworthy.

Phase 4 — ensemble

Blend only the candidates whose OOF predictions improve the portfolio, not just their own standalone score.

Phase 5 — governance

Log every change, keep a reproducible experiment table, and protect yourself from public-board whiplash.

If you execute this checklist well, you are already working much closer to winner behavior.

Source map

Primary sources used in this review

- Kaggle competition overview: WiDS Datathon 2025
- Kaggle writeup URL: Jenna — 1st place winner
- Kaggle writeup URL: Team11 — 2nd place winner
- WiDS Worldwide winners blog
- WiDS Datathon 2025 winners panel page
- WiDS 2025 Workshop #2 description (preprocessing)
- WiDS 2025 Workshop #3 description (model building and F1 metric)

How to read the claims

- Exact winner quotes are treated as direct evidence.
- Competition framing and workshop process notes provide official context.
- All deeper solution commentary is marked as interpretation, not as hidden access to the exact winner notebooks.

This slide is here for auditability: the point is to make the review easy to trust.

Final takeaways

From Team Jenna

The best feature space can matter more than the fanciest model.

From Team11

The best final score often comes from a portfolio of good models, not one perfect model.

For future Kaggle work

Representation first. Validation second. Diversity third. Blend last.

One-line summary

1st place teaches how to see the data better. 2nd place teaches how to combine models better. Together they describe a very strong competitive workflow.

PSL UNIVERSITÉ PARIS Dauphine UNIVERSITÉ PARIS ENS MINES ParisTech kaggle

Thank you.