aival

Healthcare AI evaluation and monitoring

AI will transform healthcare

Aival enables an AI product to be properly analysed, to give clinicians the confidence to know that it will work at their local site and for their patients

Our solution allows rapid, scalable and repeatable independent assessment of AI products without requiring technical expertise

Our methodology is based on decades of experience in developing medical imaging AI algorithms and commercial products, understanding their failure modes and weaknesses and how to test for them

Our founder

has 15+ years' experience in research, development, and regulation of AI products for healthcare

Kanwal Bhatia, Ph.D.

Head of Data Science at Visulytix, leading a team of 6 data scientists. Developed IP that sold to big pharma / device manufacturers

AI Architect at Odin Vision

Technical Advisory Board at Ultromics

Ph.D. in Medical Image Computing (Imperial College London, 2007) with 1500+ citations

Imperial College London IXICO SIEMENS KING'S College LONDON ULTROMICS™

Clinical adoption of AI is slow

- > 500 AI devices cleared for clinical use (FDA)
  - $5bn investment into medical imaging AI since 2015
- Lack of standard pathways to adoption
  - AI products are hard to understand, operating as 'black boxes'
  - No standard pathways for validating products before adoption (current methods are expensive / biased)
  - Weak monitoring of AI performance once in use
  - Clinical staff do not have the time or skillset to evaluate technical performance and safety of AI

** Signify Research, May 2023*

Build trust through evaluation

Aival software evaluates AI products rapidly and at scale

- Comparison
  - Identify and compare products that provide greatest clinical benefit for a given site
- Evaluation
  - Rigorous *independent* assessment
  - Substantiate manufacturer performance claims
  - Accelerate time to sale and reduce cost of adoption
- Monitoring
  - Ensure products continue to perform as expected over time
  - Standardise post-market surveillance reporting

AI product assessment on local site data

Performance metrics

How well does it work?

Does it work across all acquisition devices and pathologies

Explainability

Did the algorithm make the right prediction for the right reasons?

Fairness & bias

Are all population subgroups treated in the same way?

Robustness

Will the algorithm perform just as well with unexpected / variable data?

Performance

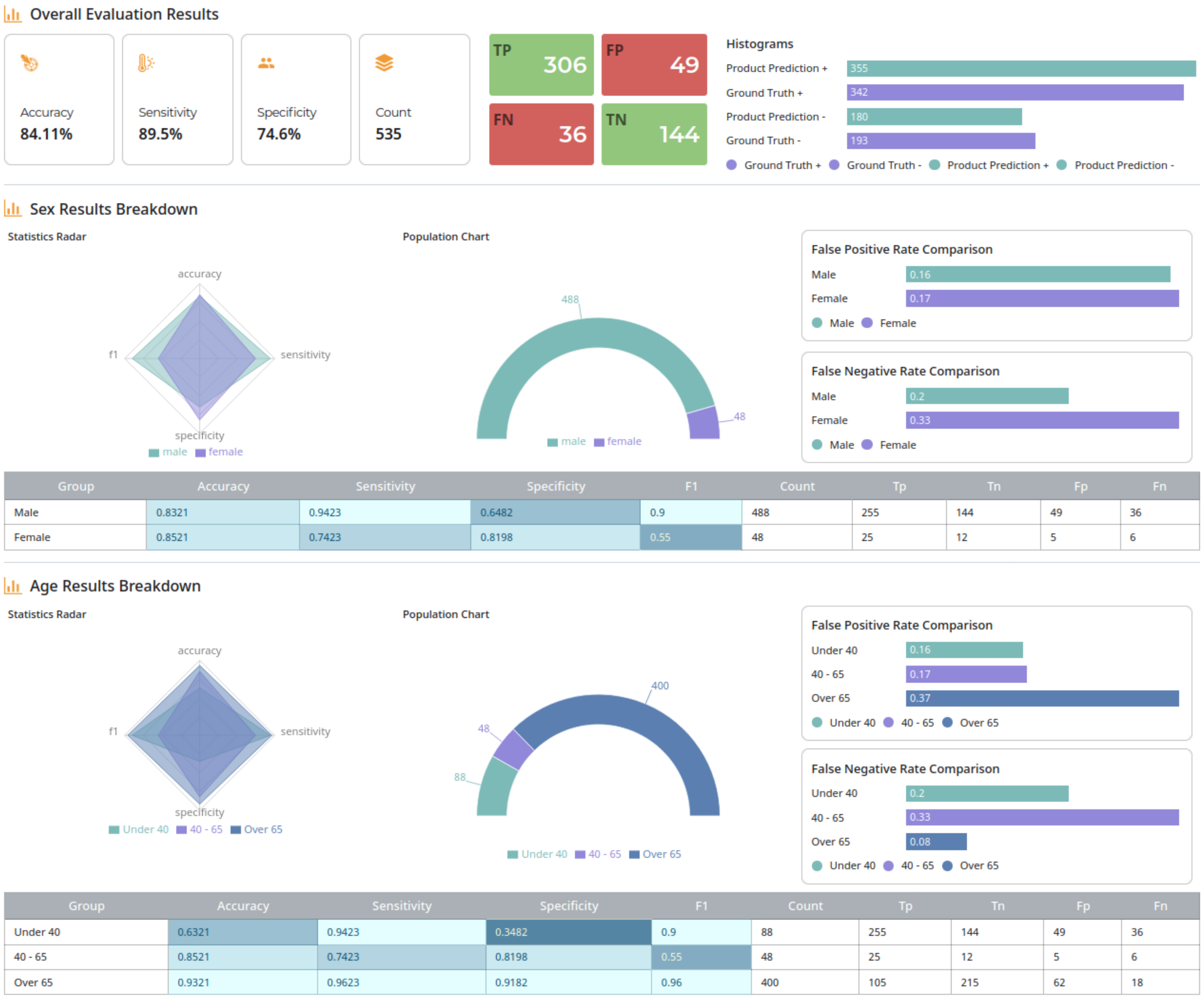
Accuracy

Sensitivity

Specificity

Errors

Subgroup analysis



Overall Evaluation Results

Accuracy 84.11%	Sensitivity 89.5%	Specificity 74.6%	Count 535

TP 306	FP 49
FN 36	TN 144

Sex Results Breakdown

Group	Accuracy	Sensitivity	Specificity	F1	Count	Tp	Tn	Fp	Fn
Male	0.8321	0.9423	0.6482	0.9	488	255	144	49	36
Female	0.8521	0.7423	0.8198	0.55	48	25	12	5	6

Age Results Breakdown

Group	Accuracy	Sensitivity	Specificity	F1	Count	Tp	Tn	Fp	Fn
Under 40	0.6321	0.9423	0.3482	0.9	88	255	144	49	36
40 - 65	0.8521	0.7423	0.8198	0.55	48	25	12	5	6
Over 65	0.9321	0.9623	0.9182	0.96	400	105	215	62	18

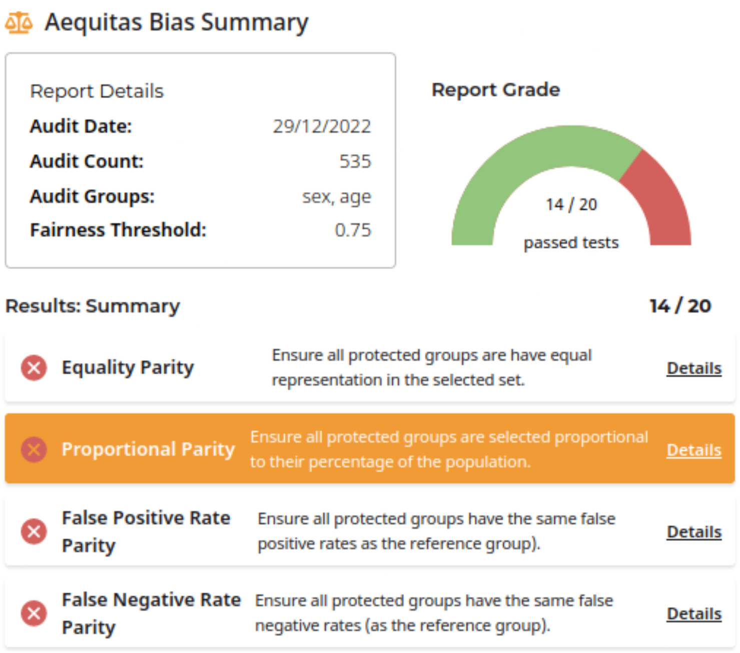
Fairness

Accuracy parity

Predictive value parities

False positive rate parity

False negative rate parity



Aequitas Bias Summary

Report Details

Audit Date:	29/12/2022
Audit Count:	535
Audit Groups:	sex, age
Fairness Threshold:	0.75

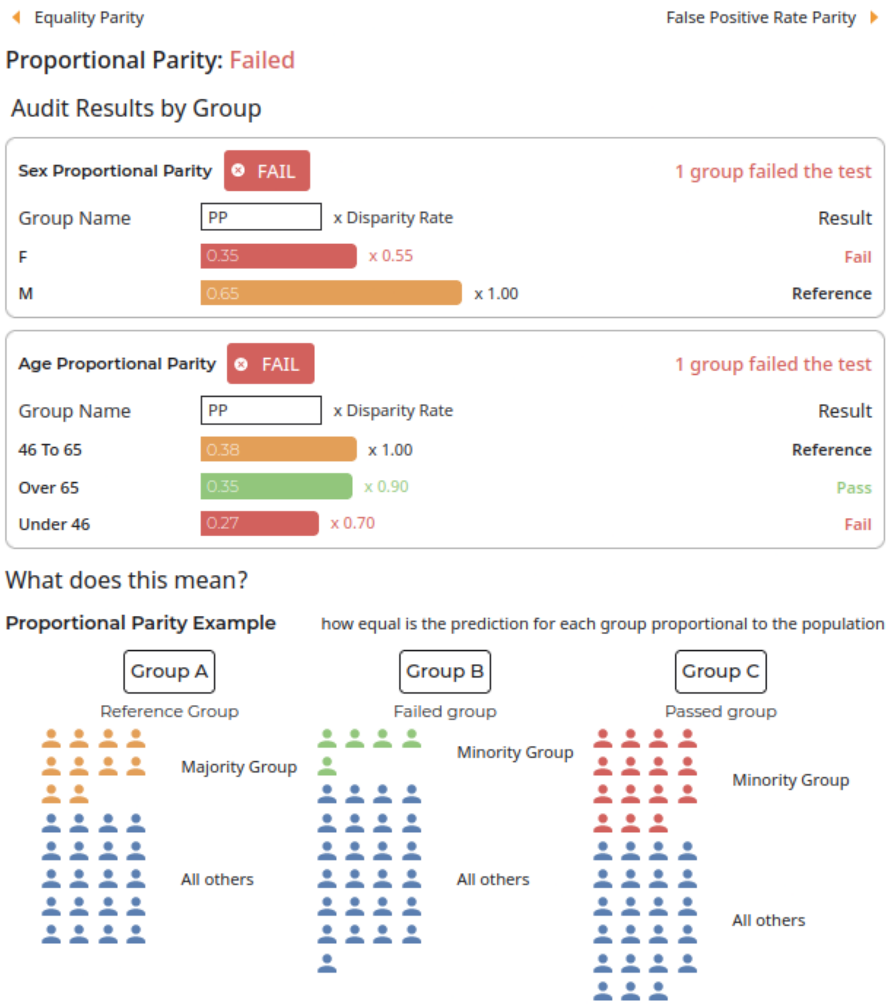
Report Grade

14 / 20

passed tests

Results: Summary **14 / 20**

☒ **Equality Parity**	Ensure all protected groups are have equal representation in the selected set.	Details
☒ **Proportional Parity**	Ensure all protected groups are selected proportional to their percentage of the population.	Details
☒ **False Positive Rate Parity**	Ensure all protected groups have the same false positive rates as the reference group).	Details
☒ **False Negative Rate Parity**	Ensure all protected groups have the same false negative rates (as the reference group).	Details



◂ Equality Parity   False Positive Rate Parity ▸

Proportional Parity: Failed

Audit Results by Group

Sex Proportional Parity ☒ FAIL   1 group failed the test

Group Name	PP	x Disparity Rate	Result
F	0.35	x 0.55	Fail
M	0.65	x 1.00	Reference

Age Proportional Parity ☒ FAIL   1 group failed the test

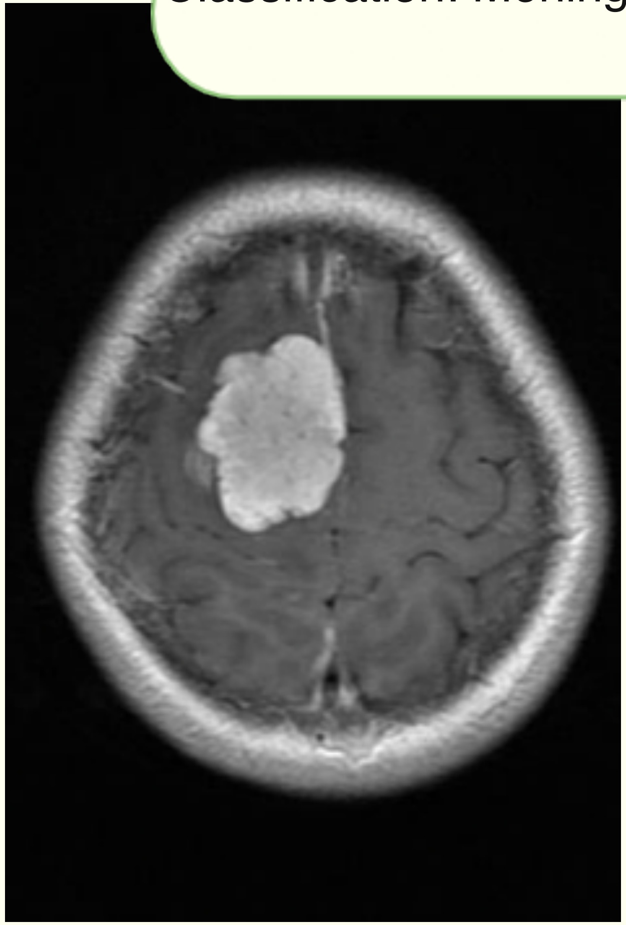
Group Name	PP	x Disparity Rate	Result
46 To 65	0.38	x 1.00	Reference
Over 65	0.35	x 0.90	Pass
Under 46	0.27	x 0.70	Fail

What does this mean?

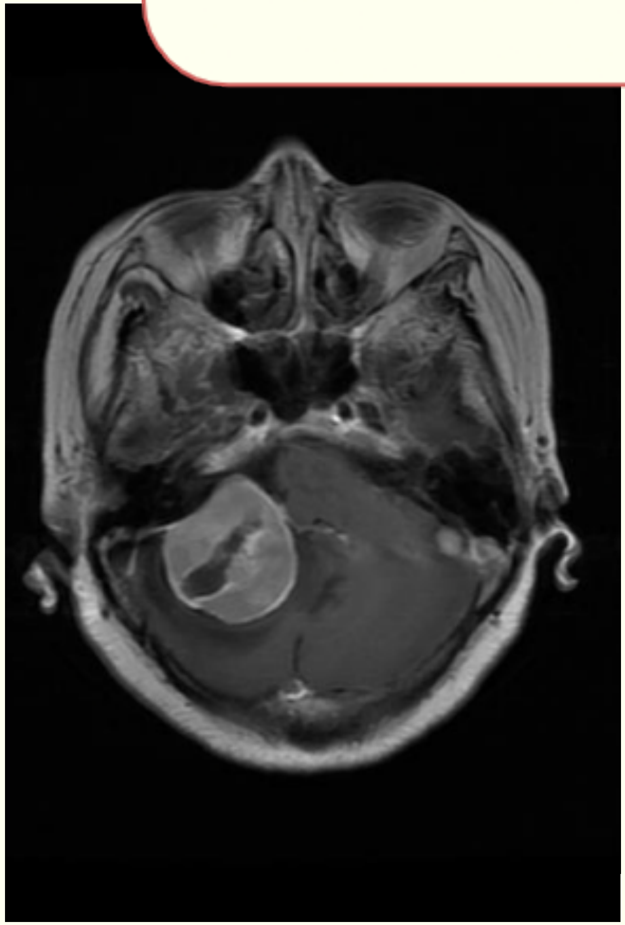
Proportional Parity Example how equal is the prediction for each group proportional to the population

Explainability (black-box)

Is the AI product making the right decisions for the right reasons? We test products as black-boxes without access to underlying model / architecture



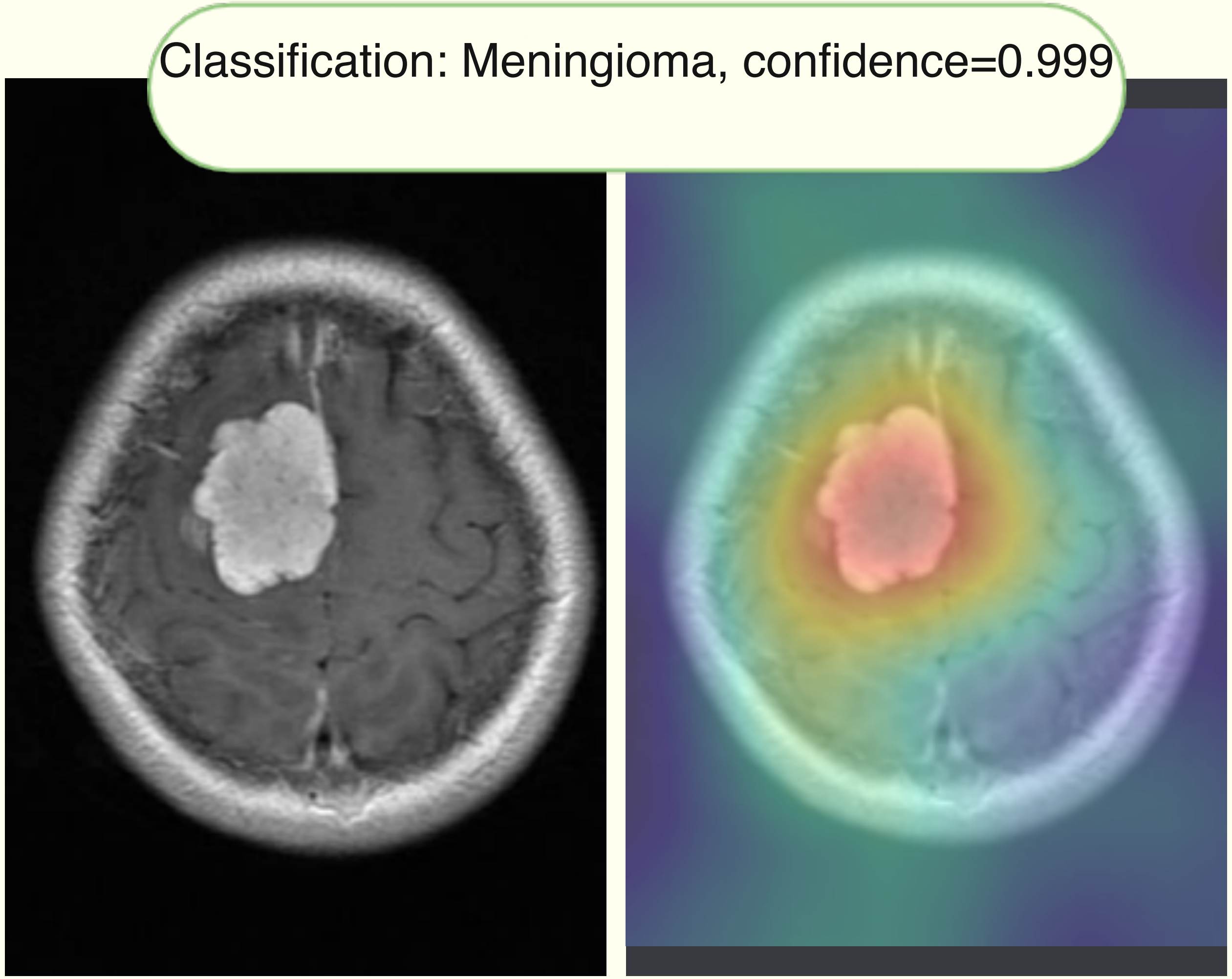
Classification: Meningioma, confidence=0.999



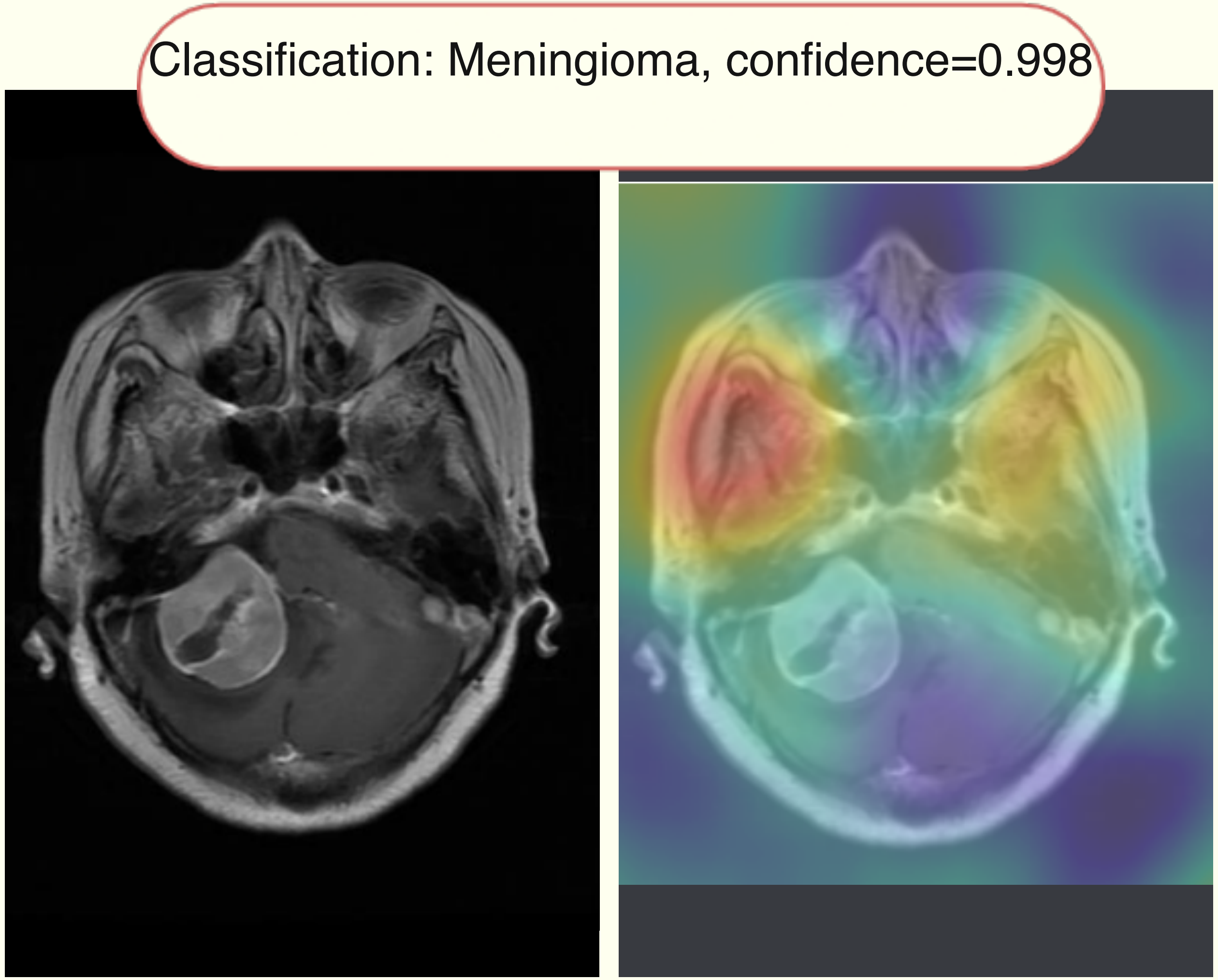
Classification: Meningioma, confidence=0.998

Explainability (black-box)

Is the AI product making the right decisions for the right reasons? We test products as black-boxes without access to underlying model / architecture

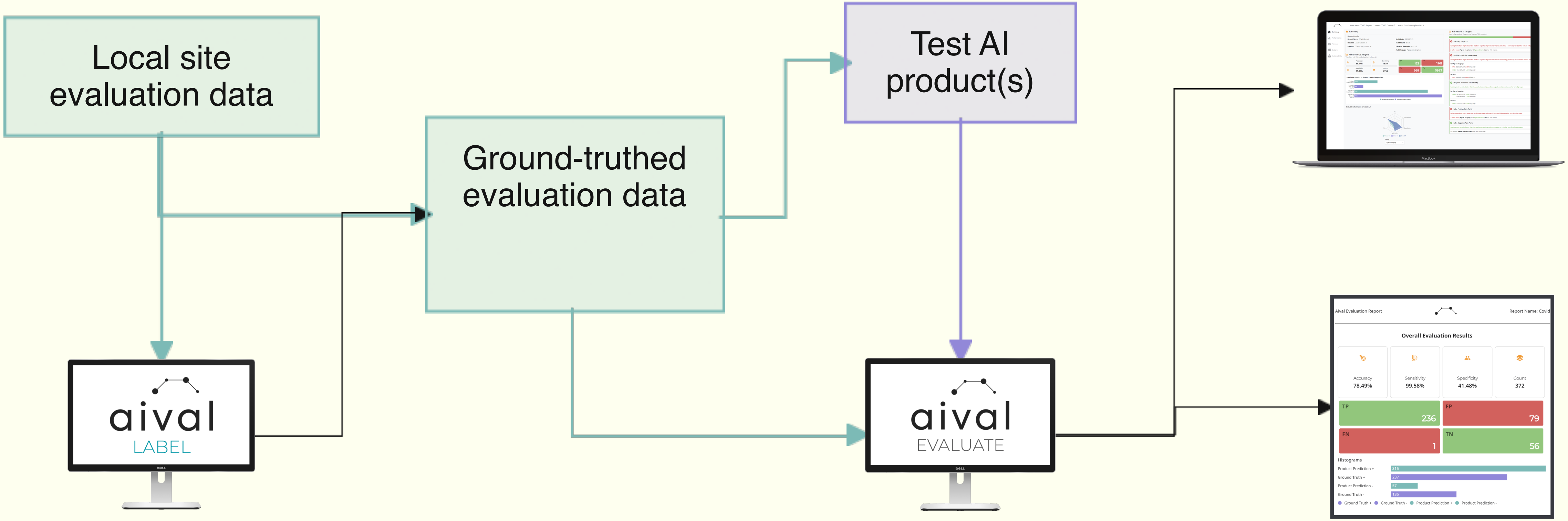


Classification: Meningioma, confidence=0.999



Classification: Meningioma, confidence=0.998

Aival evaluation workflow



Local site evaluation data

Ground-truthed evaluation data

Test AI product(s)

aival LABEL

aival EVALUATE

Aival Evaluation Report — Report Name: Covid

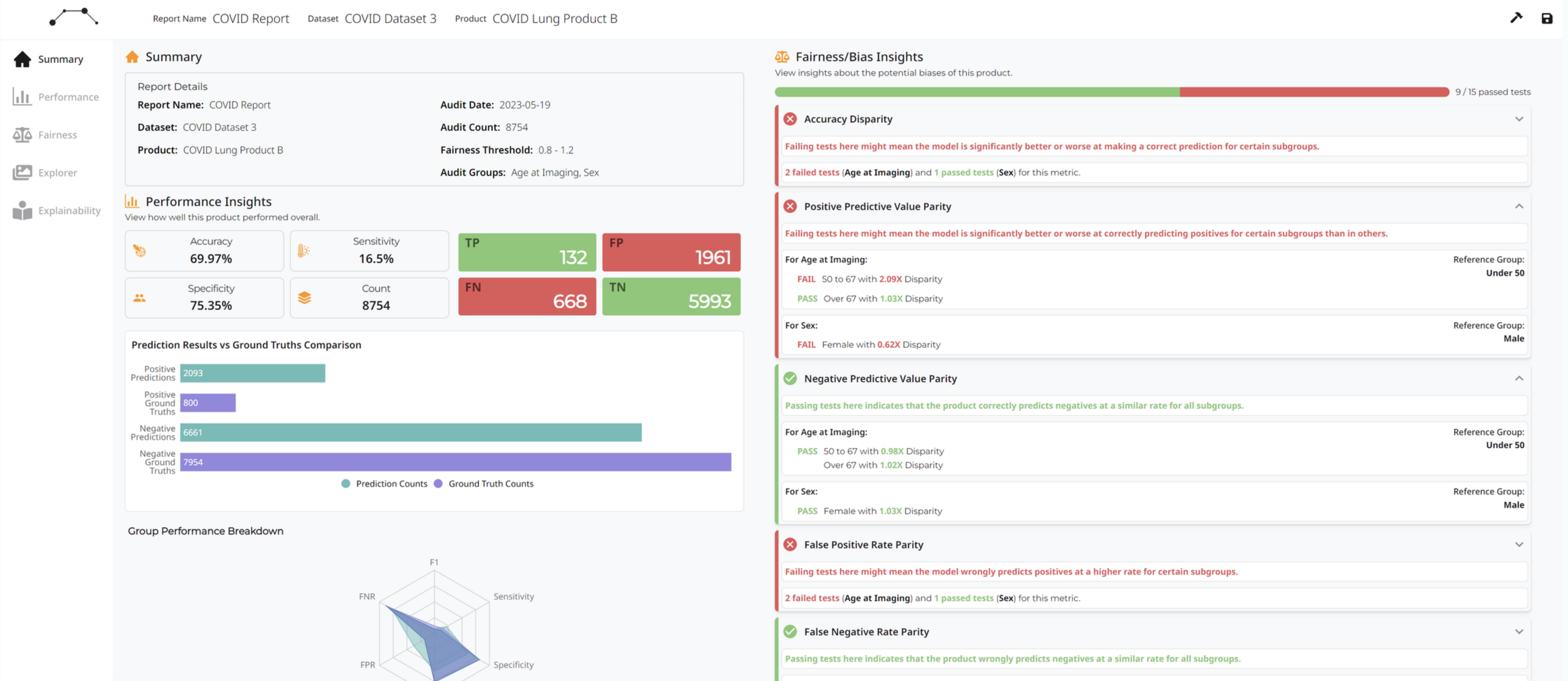
Overall Evaluation Results

Accuracy	Sensitivity	Specificity	Count
78.49%	99.58%	41.48%	372

TP 236	FP 79
FN 1	TN 56

Histograms

Sample analysis report



Report Name COVID Report Dataset COVID Dataset 3 Product COVID Lung Product B

Summary

Report Details

Report Name: COVID Report	**Audit Date:** 2023-05-19
Dataset: COVID Dataset 3	**Audit Count:** 8754
Product: COVID Lung Product B	**Fairness Threshold:** 0.8 - 1.2
	Audit Groups: Age at Imaging, Sex

Performance Insights

View how well this product performed overall.

Accuracy 69.97%	Sensitivity 16.5%	TP 132	FP 1961
Specificity 75.35%	Count 8754	FN 668	TN 5993

Prediction Results vs Ground Truths Comparison

Positive Predictions 2093
Positive Ground Truths 800
Negative Predictions 6661
Negative Ground Truths 7954
Prediction Counts
Ground Truth Counts

Group Performance Breakdown

Fairness/Bias Insights

View insights about the potential biases of this product.

9 / 15 passed tests

Accuracy Disparity

Failing tests here might mean the model is significantly better or worse at making a correct prediction for certain subgroups.

2 failed tests (**Age at Imaging**) and 1 passed tests (**Sex**) for this metric.

Positive Predictive Value Parity

Failing tests here might mean the model is significantly better or worse at correctly predicting positives for certain subgroups than in others.

For Age at Imaging: Reference Group: **Under 50**

FAIL 50 to 67 with 2.09X Disparity

PASS Over 67 with 1.03X Disparity

For Sex: Reference Group: **Male**

FAIL Female with 0.62X Disparity

Negative Predictive Value Parity

Passing tests here indicates that the product correctly predicts negatives at a similar rate for all subgroups.

For Age at Imaging: Reference Group: **Under 50**

PASS 50 to 67 with 0.98X Disparity
Over 67 with 1.02X Disparity

For Sex: Reference Group: **Male**

PASS Female with 1.03X Disparity

False Positive Rate Parity

Failing tests here might mean the model wrongly predicts positives at a higher rate for certain subgroups.

2 failed tests (**Age at Imaging**) and 1 passed tests (**Sex**) for this metric.

False Negative Rate Parity

Passing tests here indicates that the product wrongly predicts negatives at a similar rate for all subgroups.

Use cases

Healthcare Providers

- Validate manufacturer performance claims on local data
- Ensure fairness across demographics
- Compare different AI products

AI Vendors

- Gain trust with clinical users
- Internal self-assessment of failure modes
- Standardise reporting for regulatory submissions

AI Platforms

- Help your users to assess different products across your platform

Get in touch

@ kanwal@aival.io

☎ +447795975256

🌐 https://www.aival.io

aival