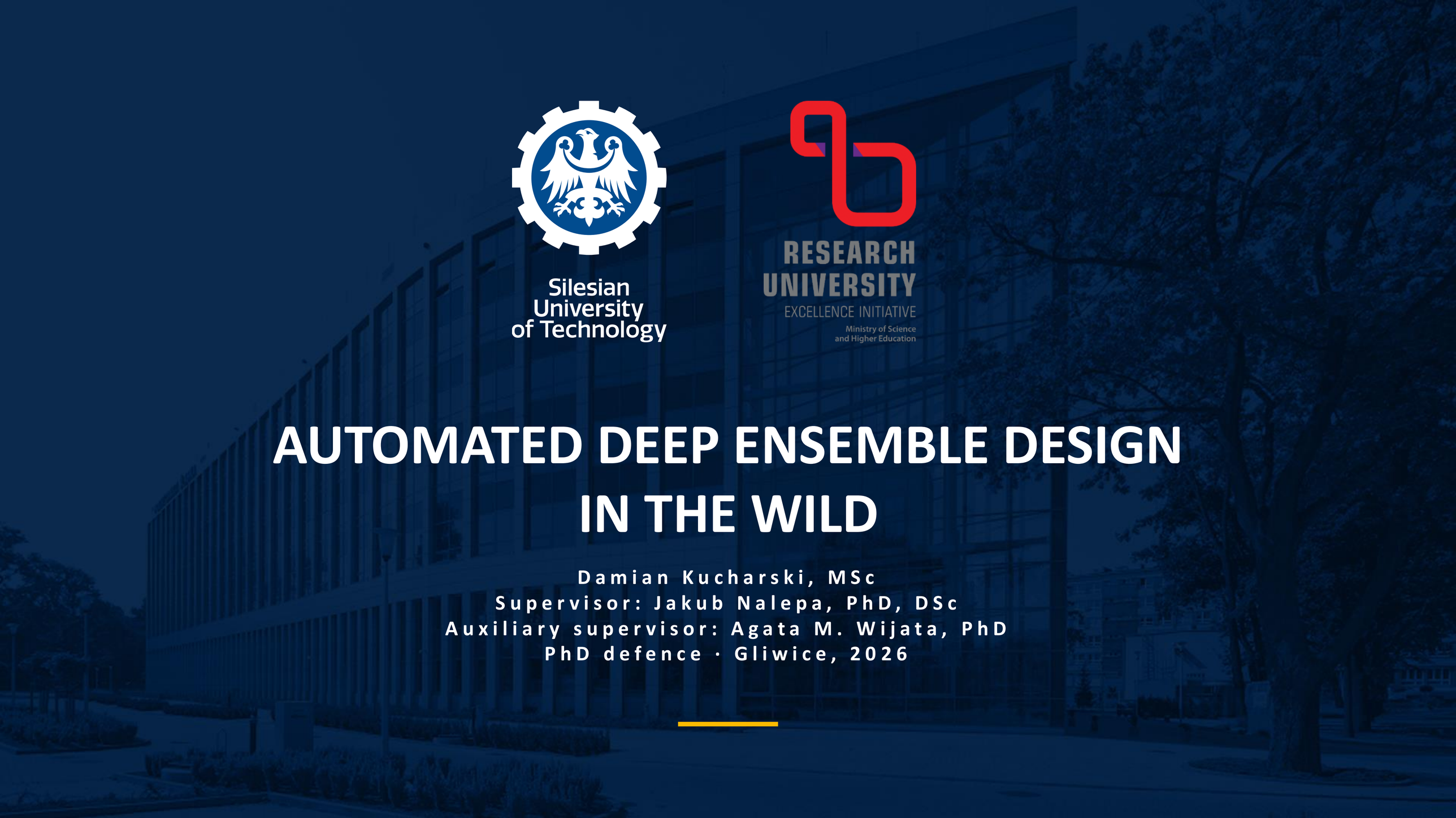


Silesian University of Technology

RESEARCH UNIVERSITY
EXCELLENCE INITIATIVE
Ministry of Science and Higher Education

AUTOMATED DEEP ENSEMBLE DESIGN IN THE WILD

Damian Kucharski, MSc
Supervisor: Jakub Nalepa, PhD, DSc
Auxiliary supervisor: Agata M. Wijata, PhD

PhD defence · Gliwice, 2026

Research programme and outputs

5	3	3	6
Journal publications	Conference papers	Patent applications	Conference abstracts

Clinical imaging	Diagnostic applications of MRI and CT
ECG analysis	Classification and printout variability
Ensemble design	 GIRAFFE to OKAPI  · Complexity and transfer

Journal publications

J1

IEEE TRANSACTIONS ON
EVOLUTIONARY COMPUTATION

OKAPI: Automated Deep Ensemble Design Through Genetic Programming

Damian Kucharski†, Jakub Nalepa†, *Senior Member, IEEE*, Arkadiusz Czerwinski, Yalin Zheng, Gregory Y. H. Lip, Agata M. Wijata, *Senior Member, IEEE*

J2

Computer Methods and Programs in Biomedicine

Detecting Cardiovascular Diseases Using ECG Scans and Explainable Artificial Intelligence

Arkadiusz Czerwinski[1], Damian Kucharski[1], Jacek Kawa[2], Yalin Zheng[3,4], Gregory Y. H. Lip[3,5,6], Jakub Nalepa[1,†], Agata M. Wijata[2,†*]

J3

npj | digital medicine

Interpretable arrhythmia detection in ECG scans using deep learning ensembles: a genetic programming approach

Check for updates

Arkadiusz Czerwinski[1,2], Damian Kucharski[1,2], Agata M. Wijata[2]✉, Hanadi Aldosari[2], Frans Coenen[3], Dhiraj Gupta[4], Lu Fu[4], Weidong Lin[4], Yumei Xue[5], Jacek Kawa[2], Yalin Zheng[3,6], Gregory Y. H. Lip[1,7,8] & Jakub Nalepa[1]✉

J4

Computerized Medical Imaging and Graphics

Deep learning ensembles for detecting brain metastases in longitudinal multi-modal MRI studies

Bartosz Machura[a], Damian Kucharski[a,d], Oskar Bozek[c], Bartosz Eksner[f], Bartosz Kokoszka[a], Tomasz Pekala[a], Mateusz Radom[b], Marek Strzelczak[b], Lukasz Zarudzki[b], Benjamín Gutiérrez-Becker[b], Agata Krason[c], Jean Tessier[c], Jakub Nalepa[d,a,*]

J5

Computers in Biology and Medicine

Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features

Krzysztof Kotowski[1], Damian Kucharski[1], Bartosz Machura[1], Szymon Adamski[1], Benjamín Gutierrez Becker[2], Agata Krason[3], Lukasz Zarudzki[4], Jean Tessier[3], Jakub Nalepa[1,5]

OKAPI

IEEE Transactions on Evolutionary Computation

2026

IF 15.9 / 200 pts

ECG robustness and XAI

Computer Methods and Programs in Biomedicine

2026

IF 6.4 / 100 pts

Clinical GIRAFFE

npj Digital Medicine

2025

IF 18.0 / 20 pts

Brain metastases

Computerized Medical Imaging and Graphics

2024

IF 4.9 / 100 pts

Liver cirrhosis

Computers in Biology and Medicine

2023

IF 7.0 / 100 pts

Conference publications

C1



Crafting Deep Learning Models for Classifying ECG Paper Printouts

Damian Kucharski[1,*], Arkadiusz Czerwinski[1,*], Agata M Wijata[2], Jacek Kawa[2], Yalin Zheng[3,4], Gregory Y H Lip[3,5], Jakub Nalepa[1]

PhysioNet Challenge

Computing in Cardiology

2024

C2



GIRAFFE: A GENETIC PROGRAMMING ALGORITHM TO BUILD DEEP LEARNING ENSEMBLES FOR ECG ARRHYTHMIA CLASSIFICATION

Damian Kucharski[1], Agata M. Wijata[1], Lu Fu[2], Weidong Lin[2], Yumei Xue[2], Jacek Kawa[1], Yalin Zheng[3,4], Gregory Y. H. Lip[3,5] and Jakub Nalepa[1]

GIRAFFE for ECG classification

IEEE International Conference on Image Processing

2024

70 points · CORE B

C3



Automated Hepatocellular Carcinoma Analysis in Multi-Phase CT with Deep Learning

Krzysztof Kotowski[1], Bartosz Machura[1], Damian Kucharski[1,2], Benjamín Gutiérrez-Becker[3], Agata Krason[4], Jean Tessier[4], and Jakub Nalepa[2,1]

HCC analysis in multiphase CT

CaPTion workshop at MICCAI

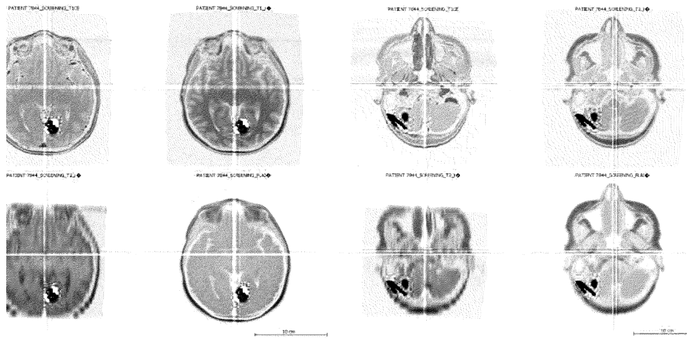
2024

140 points · CORE A

Patent applications

I am a co-inventor on three patent applications filed by Roche.

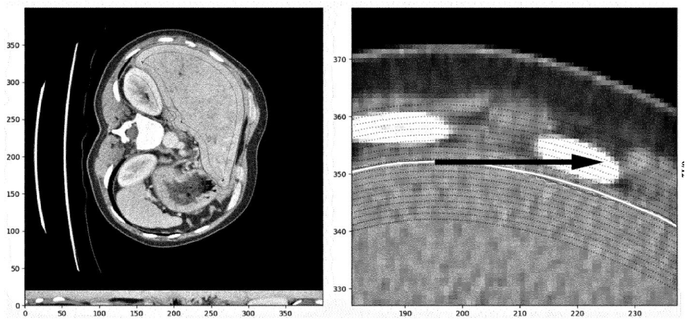
Brain metastases



WO 2025/082891 A1

Source: WO 2025/082891 A1 · Fig. 7

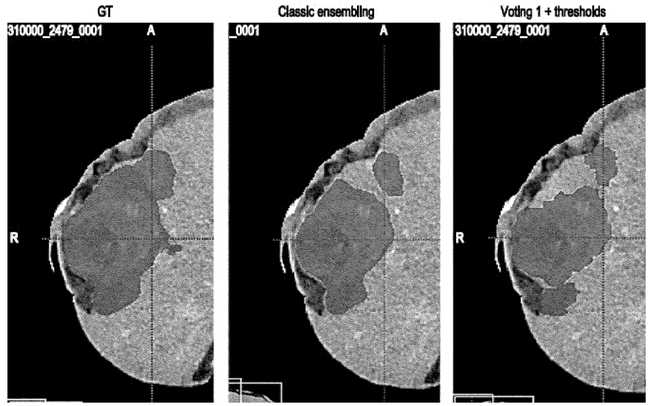
Liver cirrhosis



EP 4 476 684 A1

Source: WO 2023/152096 A1 · Fig. 5b

Liver tumours



WO 2024/184335 A1

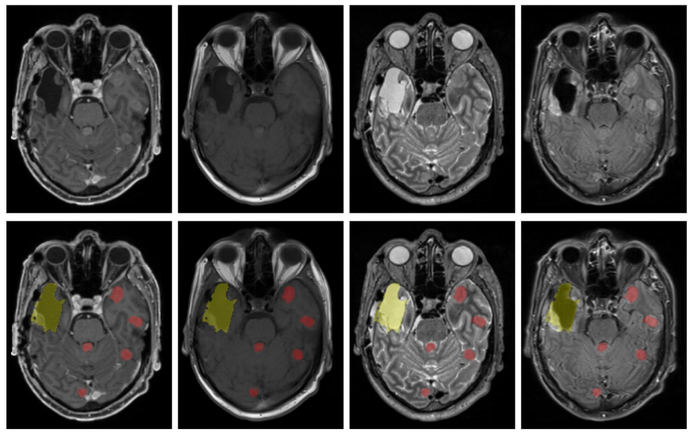
Source: WO 2024/184335 A1 · Fig. 10

Conference abstracts

CinC 2026	Reduced LVEF screening with PPG
EuroPCR 2026	Aortic valve disease screening
ESC Digital & AI 2026	AF detection across devices
ESC Digital & AI 2026	Clinical sources of AF detection errors
Kardiologia Prew. 2024	Explaining ECG ensembles
Kardiologia Prew. 2024	Learning from ECG printouts

Related work at Graylight Imaging

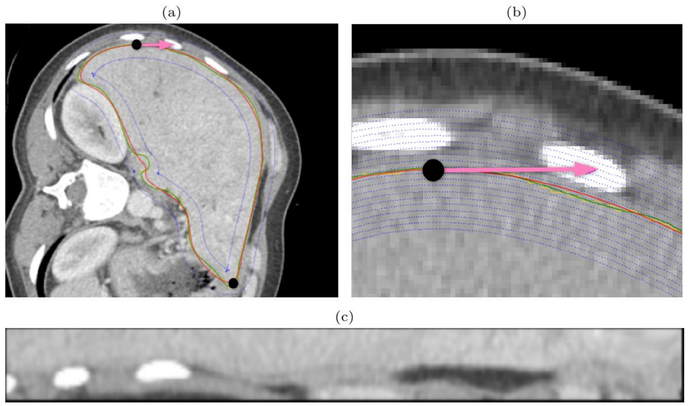
Brain metastases



Detect and track lesions on MRI

CMIG (J4) · Fig. 1

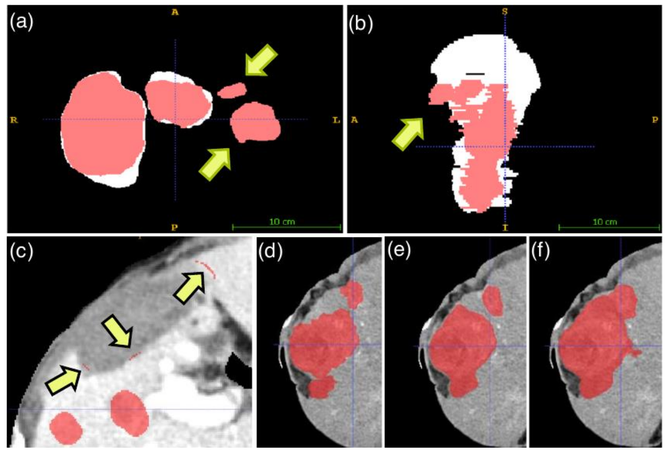
Liver cirrhosis



Detect cirrhosis from CT features

CBM (J5) · Fig. 2

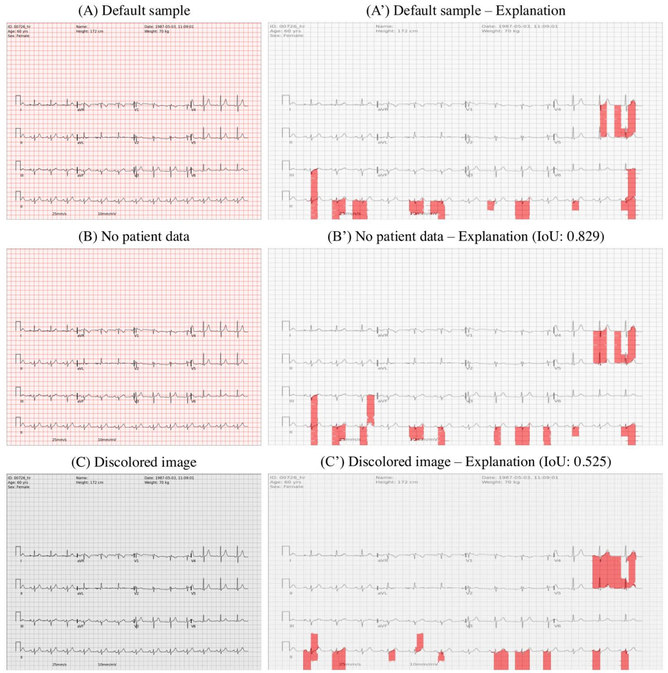
Hepatocellular carcinoma



Quantify lesions and treatment response

MICCAI CaPTion (C3) · Fig. 1

ECG printouts: classification and robustness



Perturbed ECG printouts and LIME explanations

CinC 24 Karlsruhe

PhysioNet: classify printouts with InceptionV3

Computer Methods and Programs in Biomedicine

Test how controlled artifacts affect explanations

Source · Computer Methods and Programs in Biomedicine, Fig. 4

Automated ensemble design

Can we learn how to combine model predictions?

The ensemble design space

Approach	What it determines
Simple averaging	A fixed rule
Subset and weight search	Which models and how much weight
Stacking	A fitted combiner
Open question	Can we search for the fusion rule itself?

The ensemble design space

Approach	What it determines
Simple averaging	A fixed rule
Subset and weight search	Which models and how much weight
Stacking	A fitted combiner
GIRAFFE → OKAPI	An executable fusion procedure

Research questions

RQ1 · Performance	Can evolving fusion rules outperform other ensemble methods? How prone is the search to overfitting?
RQ2 · Complexity	Can we favour smaller fusion trees while keeping similar test performance?
RQ3 · Interpretation	Which models and operations do evolved trees use, and where do the models complement each other?

Research objectives

RO1 · Representation	Build a stateful GP representation for joint model and fusion search
RO2 · Optimisation	Optimise predictive quality and program size as separate objectives
RO3 · Evaluation	Benchmark classification and segmentation against ensemble baselines
RO4 · Analysis	Study transfer, room for improvement and what evolved trees reveal

Research hypotheses

Stateful multi-objective GP can jointly select models and fusion structure while controlling complexity. Test gains depend on reliable validation rankings.

H1 · Representation	The same tree semantics work for class scores and segmentation maps
H2 · Parsimony	Optimising node count yields smaller trees at similar test performance
H3 · Validation size	Less validation data weakens selection and test performance
H4 · Anchoring	An averaging prior improves transfer while retaining an inspectable correction

RESEARCH QUESTION

Ensemble design after training

01

Trained models

Fixed pool

→

02

Predictions

Compatible scores or maps

→

03

Fusion rule

Select and combine

→

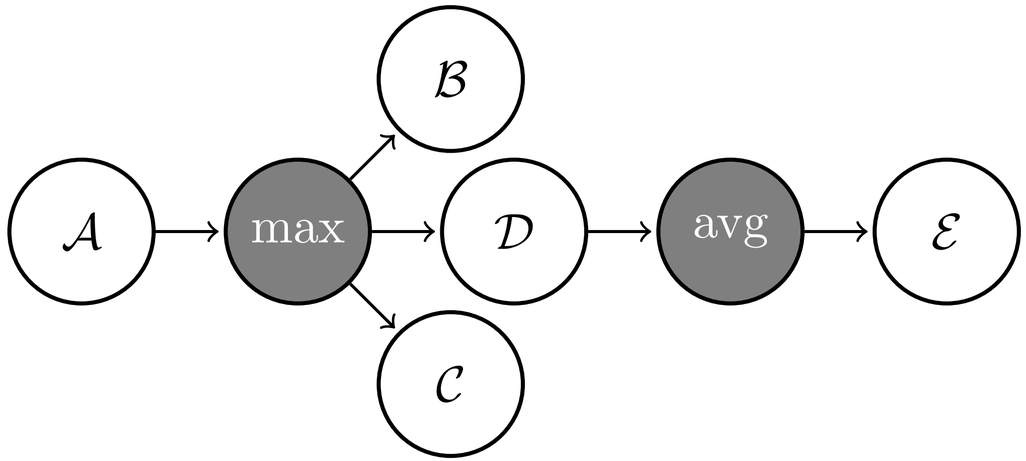
04

Output

Final prediction

The search acts on predictions from already trained models.

GIRAFFE introduces evolved fusion



Example evolved classifier

Models + operators

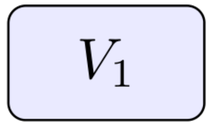
Evolve membership and the fusion procedure

First study · binary ECG classification

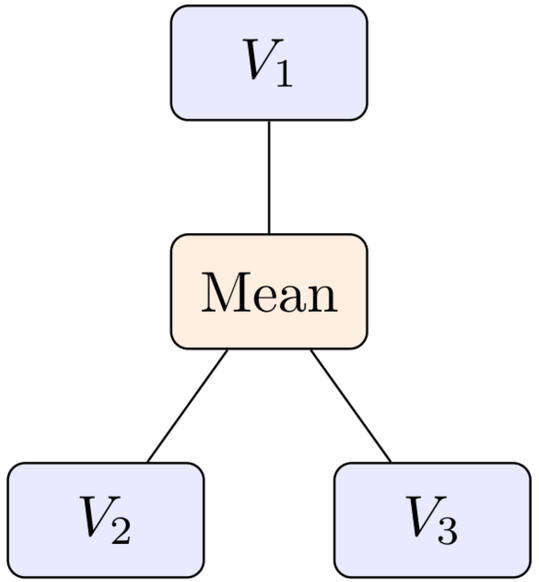
Source · IEEE ICIP paper (C2), Fig. 2a

Stateful fusion trees

(a) Minimal tree



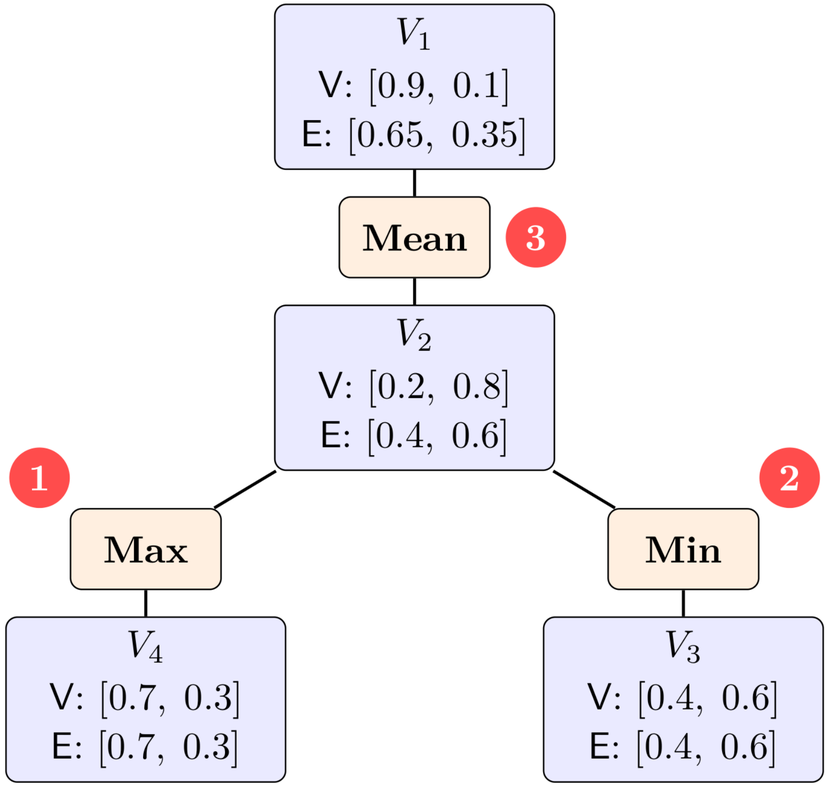
(b) Ensemble of three models



Value Nodes hold predictions. Operator Nodes update the accumulator.

Source · Thesis, Fig. 4.1

Stateful execution



Step 1 (Max):
$\max([0.2, 0.8],\ [0.7, 0.3])$
$= [0.7, 0.8]$
$\rightarrow V_2.\mathsf{E} \leftarrow [0.7, 0.8]$

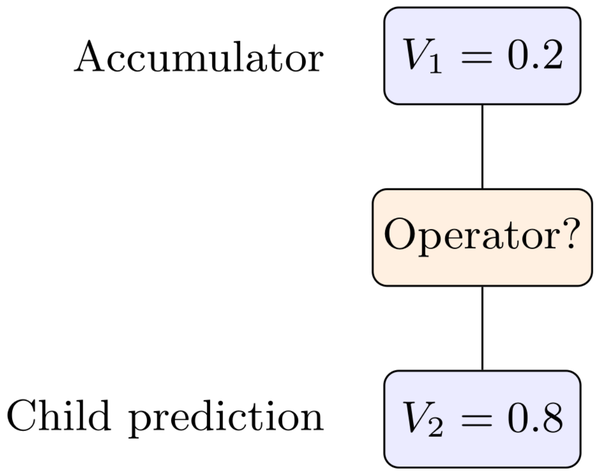
Step 2 (Min):
$\min([0.7, 0.8],\ [0.4, 0.6])$
$= [0.4, 0.6]$
$\rightarrow V_2.\mathsf{E} \leftarrow [0.4, 0.6]$

Step 3 (Mean):
$\mathrm{mean}([0.9, 0.1],\ [0.4, 0.6])$
$= [0.65, 0.35]$
$\rightarrow V_1.\mathsf{E} \leftarrow [0.65, 0.35]$

Each operator reads the current accumulator.

Source · Thesis, Fig. 4.6 · IEEE Transactions on Evolutionary Computation, Fig. 1

GIRAFFE fusion operators



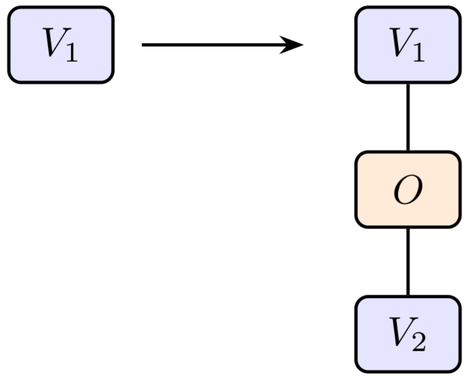
OPERATOR	OUTPUT AT V_1
Mean	0.5
Weighted Mean	0.65
Min	0.2
Max	0.8

Weighted Mean uses 25% of V_1 and 75% of V_2.

Source: Illustrative example prepared for this presentation

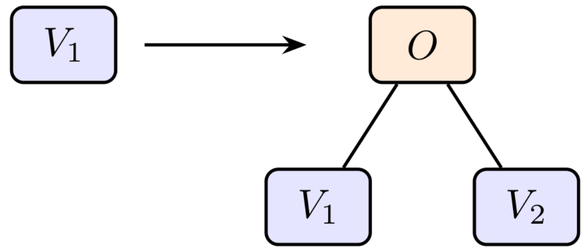
Constructive growth

(a) OKAPI in-place growth



The base prediction
remains the root accumulator;
growth appends a
modifying branch.

(b) Standard GP push-down growth

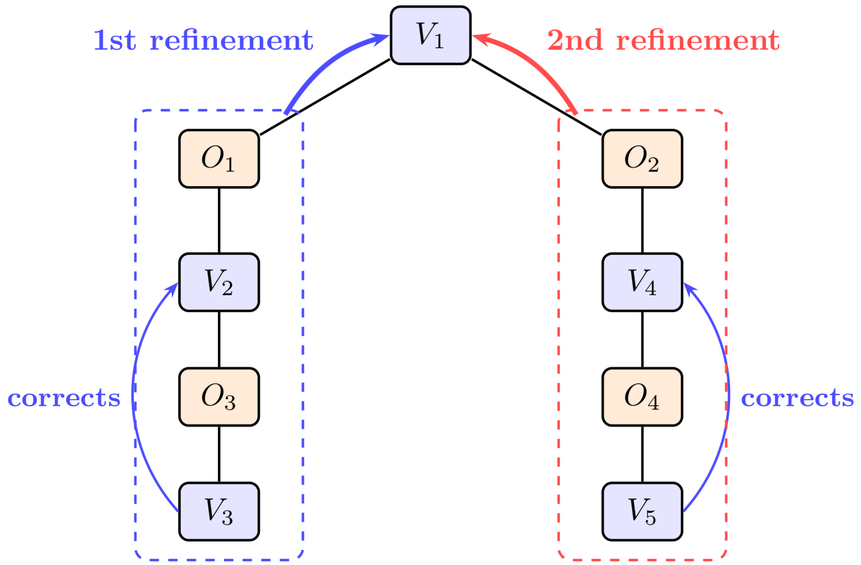


The original prediction
becomes a child
of a newly inserted operator.

New branches extend the fusion procedure.

Source · Thesis, Fig. 4.2 · IEEE Transactions on Evolutionary Computation, Supplementary Fig. S.1

Each branch refines the current prediction

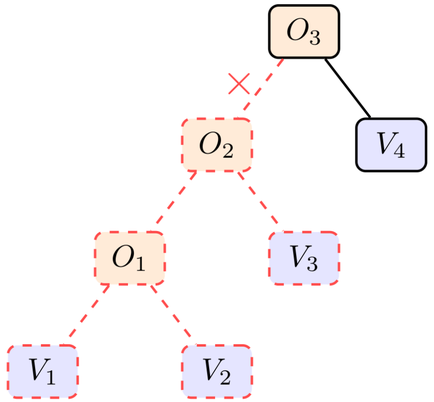


Each subtree rooted at a Value Node is a valid ensemble.
Execution from left to right lets later branches refine
the Evaluation field produced by earlier branches.

Source · Thesis, Fig. 4.3 · IEEE Transactions on Evolutionary Computation, Supplementary Fig. S.2

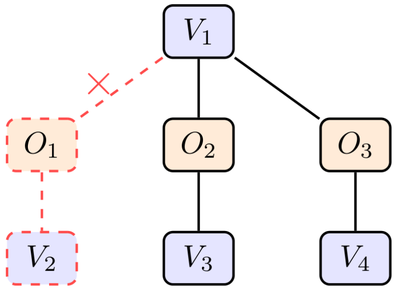
A local edit can preserve the cascade

(a) Standard GP: skewed correction cascade



Encoding "V_1 corrected by V_2, then V_3, then V_4" nests every step under a new root. One variation event near the root (×) severs V_1, V_2, and V_3 at once.

(b) OKAPI: radiating corrections

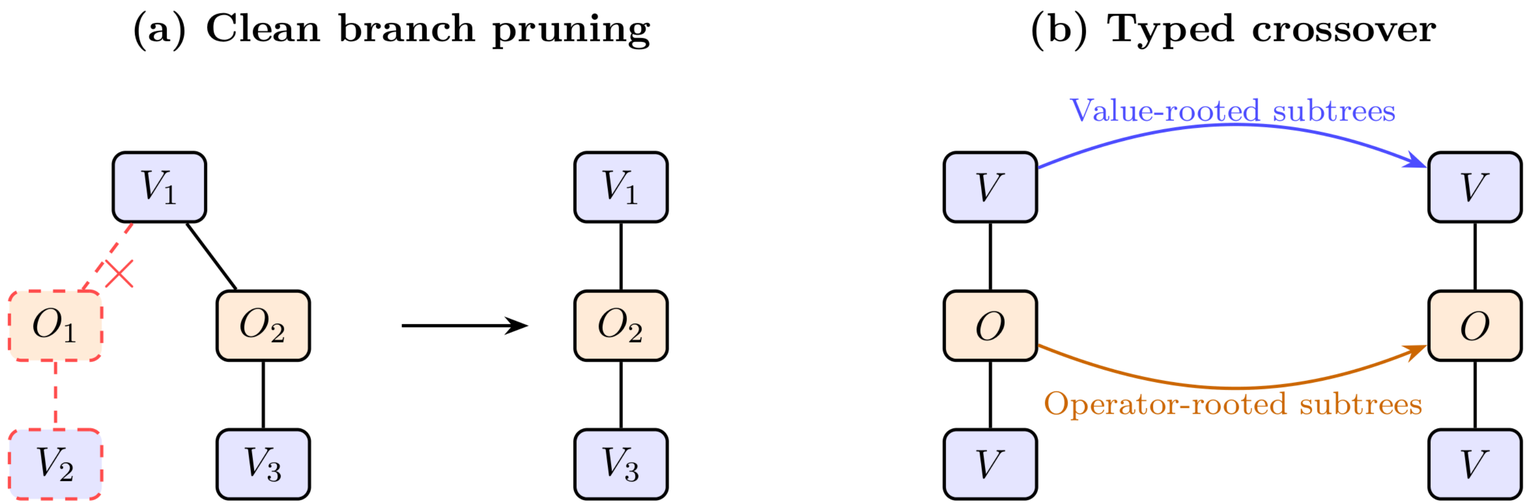


The same correction chain radiates from the root accumulator V_1. A variation event on one branch (×) removes a single correction; the remaining branch structure stays intact.

Sequential corrections

Source · Thesis, Fig. 4.5

Pruning removes the connecting operator



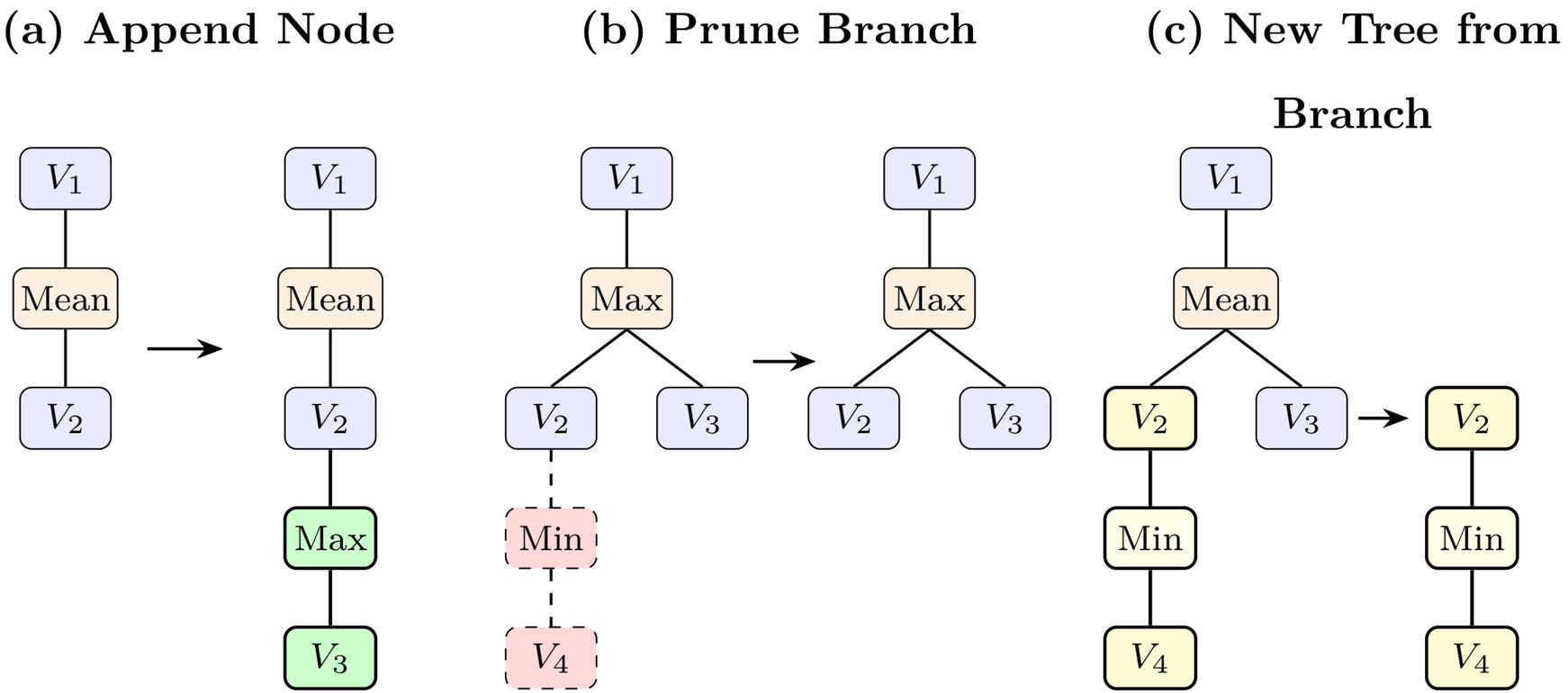
Removing a branch removes its connecting operator with it, keeping the tree compact.

Crossover swaps subtrees of matching node type, preserving the alternating structure.

Pruning and typed crossover

Source · Thesis, Fig. 4.4 · IEEE Transactions on Evolutionary Computation, Supplementary Fig. S.3

Mutation operators



Append a node · prune a branch · start a tree from a branch

Source · Thesis, Fig. 4.8

How fusion programs evolve

01

Inputs

Predictions and validation labels

→

02

Search

Evolve executable fusion trees

→

03

Selection

Choose from the Pareto set

→

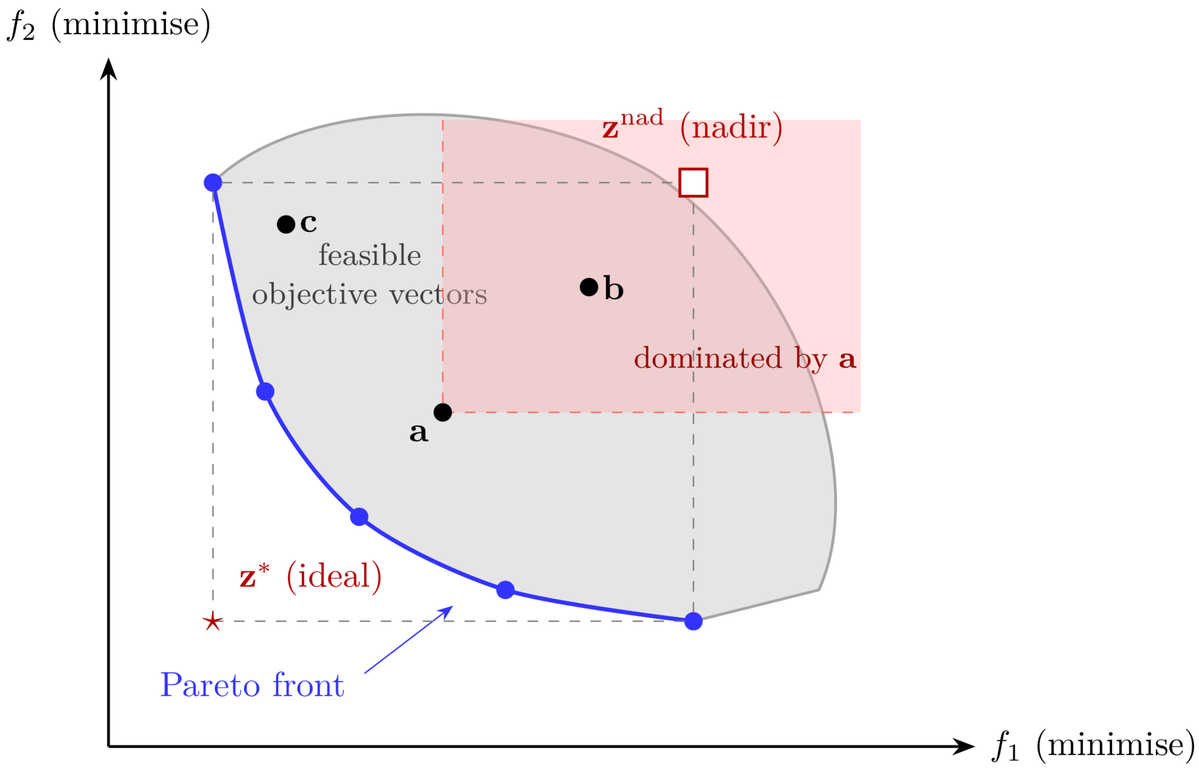
04

Test

Assess the selected tree

GIRAFFE already balances predictive quality and ensemble size.

Predictive quality and complexity



GIRAFFE already searches for accurate, compact ensembles.

Source · Thesis, Fig. 2.5

GIRAFFE: ECG classification results

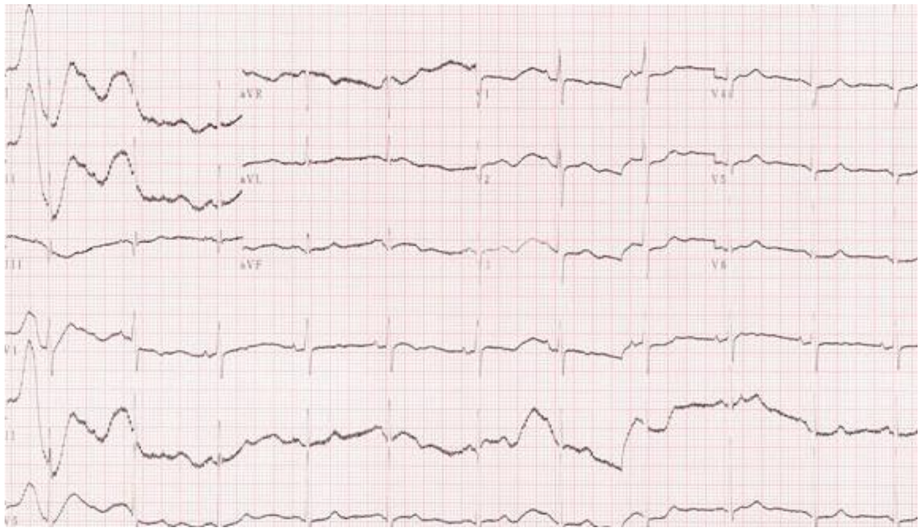


Method	ROC-AUC	F1	MCC
Base model	0.938	0.771	0.718
Full ensemble	0.965	0.872	0.835
GIRAFFE	0.960	0.909	0.883

Test results · Ranked on validation MCC, then average precision · ICIP Table 2

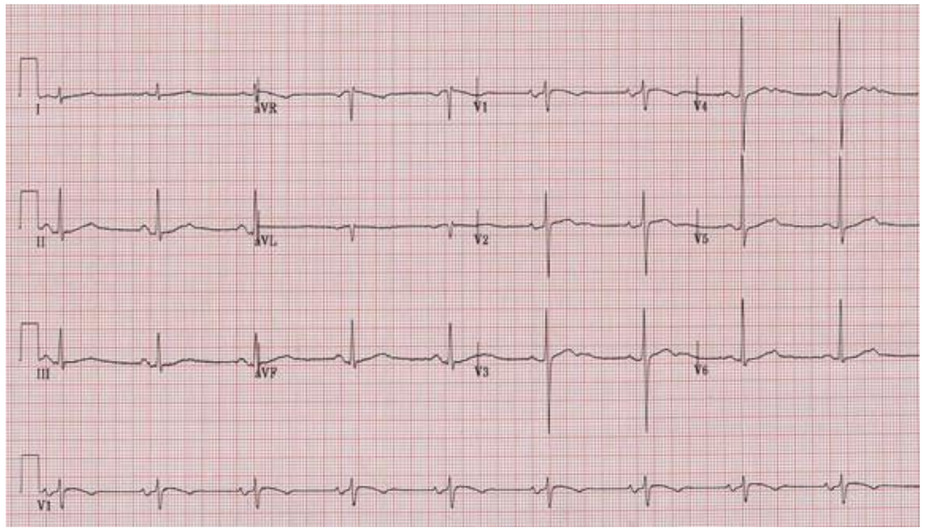
Clinical ECG studies with GIRAFFE

npj | digital medicine



G · Current arrhythmia detection

Guangzhou Heart Study (Also used in ICIP paper)



L · AF recurrence after ablation

Liverpool Heart and Chest Hospital

Example ECG scans

Source · npj Digital Medicine, Fig. 3

GIRAFFE: clinical ECG results

npj | digital medicine

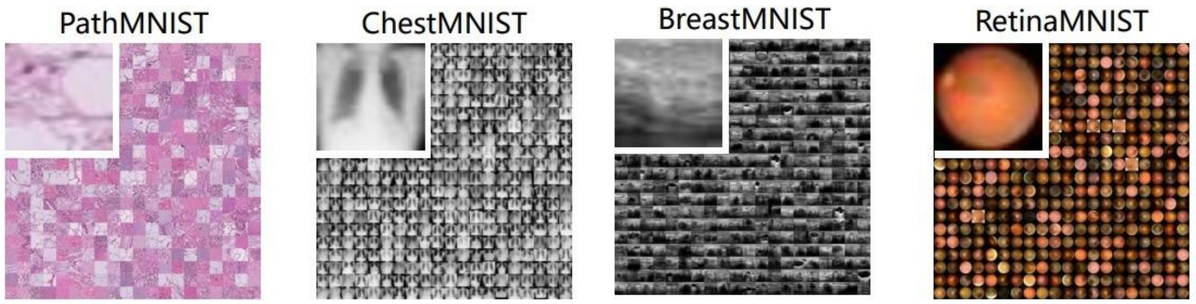
Cohort and task	Individual	Average	GIRAFFE selected	GIRAFFE best on test
G · Current arrhythmia	0.961	0.967	0.969	0.980
L · AF recurrence	0.773	0.770	0.760	0.799

Test ROC-AUC · Ceiling uses test outcomes · npj Digital Medicine, Table 2

Beyond ECG: more datasets and tasks

Broader experiments motivated a more general fusion method.

MedMNISTv2 · 12 datasets



Biomedical classification

Image source: Yang et al. · MedMNISTv2 · CC BY 4.0

STARCOP



Methane plume segmentation

Image source: Růžička et al. · STARCOP · CC BY-NC 4.0



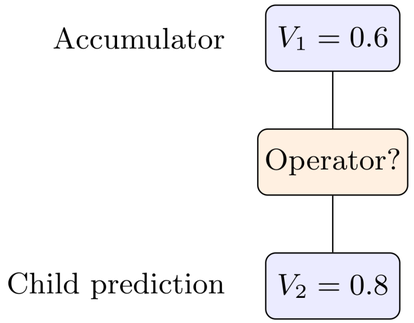
OKAPI: a broader fusion method

Tasks	Compatible prediction tensors, including segmentation maps
Objectives	Configurable performance criteria and complexity control
Operators	Threshold rules and fusion in logit space
Evaluation	Validation reliability, generalisation and prior anchoring



OKAPI extends the fusion operator set

Same two inputs



ℓ_1 = logit(0.60)

ℓ_2 = logit(0.80)

OPERATOR	RULE	OUTPUT
Close Threshold	closest to τ τ = 0.65	0.60
Far Threshold	farthest from τ τ = 0.65	0.80
Logit Mean	$\sigma(\kappa(\ell_1 + \ell_2)/2 + b)$ κ = 2, b = 0	0.86
Weighted Logit Mean	$\sigma(w_1\ell_1 + w_2\ell_2)$ $w_1 = 1$, $w_2 = 0.5$	0.75

Logit operators enable calibration and local logistic stacking.



Benchmarks and available base models

Benchmark	Task	Base models	Metric used for ranking
MedMNISTv2	Biomedical image classification	30 per dataset	Average precision (PR-AUC)
STARCOP	Methane plume segmentation	13 models	F1 for plume pixels

Multiobjective optimisation · The listed metrics are used for ranking



Evaluation protocol

01

Training

Fit the base models

→

02

Validation

Search and select fusion

→

03

Test

Measure performance on new data

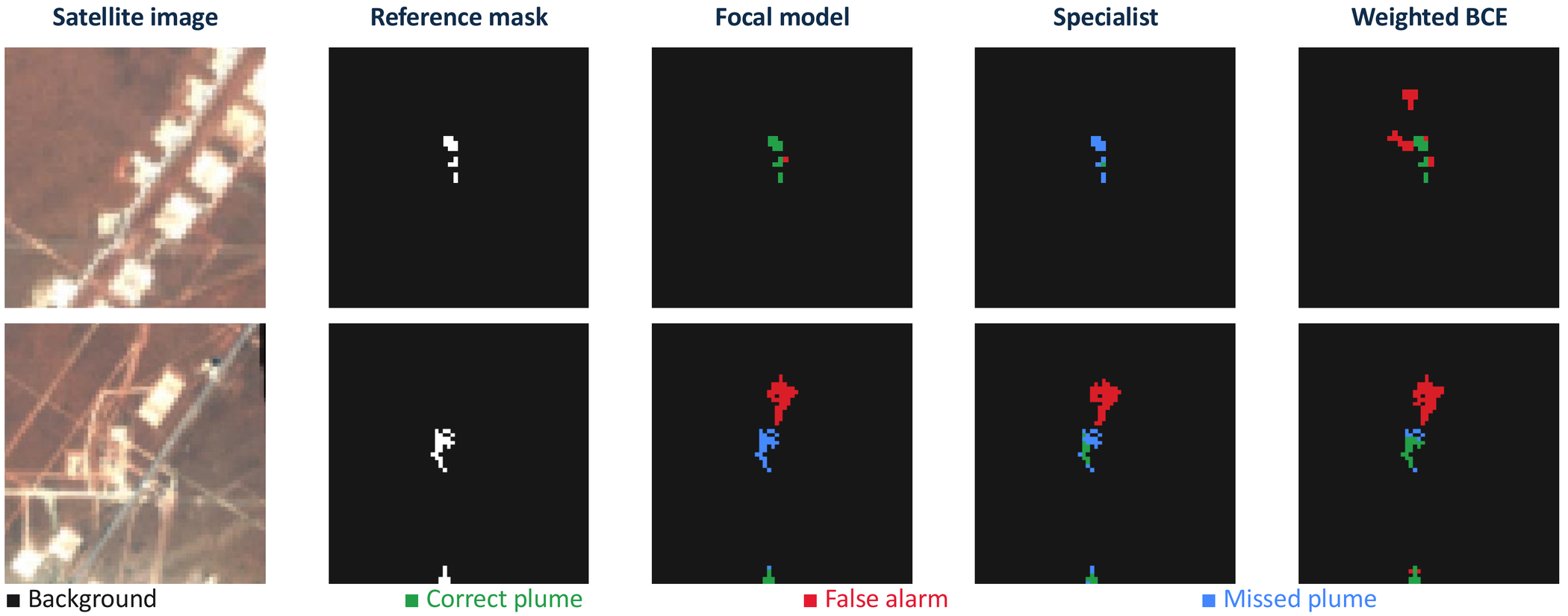
Maximum tree depth of 3 or 5 · with or without parsimony objective · 10 runs per setting



MedMNISTv2: averaging all models performs best

Method	Mean test PR-AUC
Simple Average	0.8047
NSGA-III	0.8020
Best-5	0.8003
OKAPI	0.7996
Logistic stacking	0.7886

STARCOP: examples of base model predictions



Source · Thesis, Fig. 5.1 · Enlarged plume regions from two test tiles

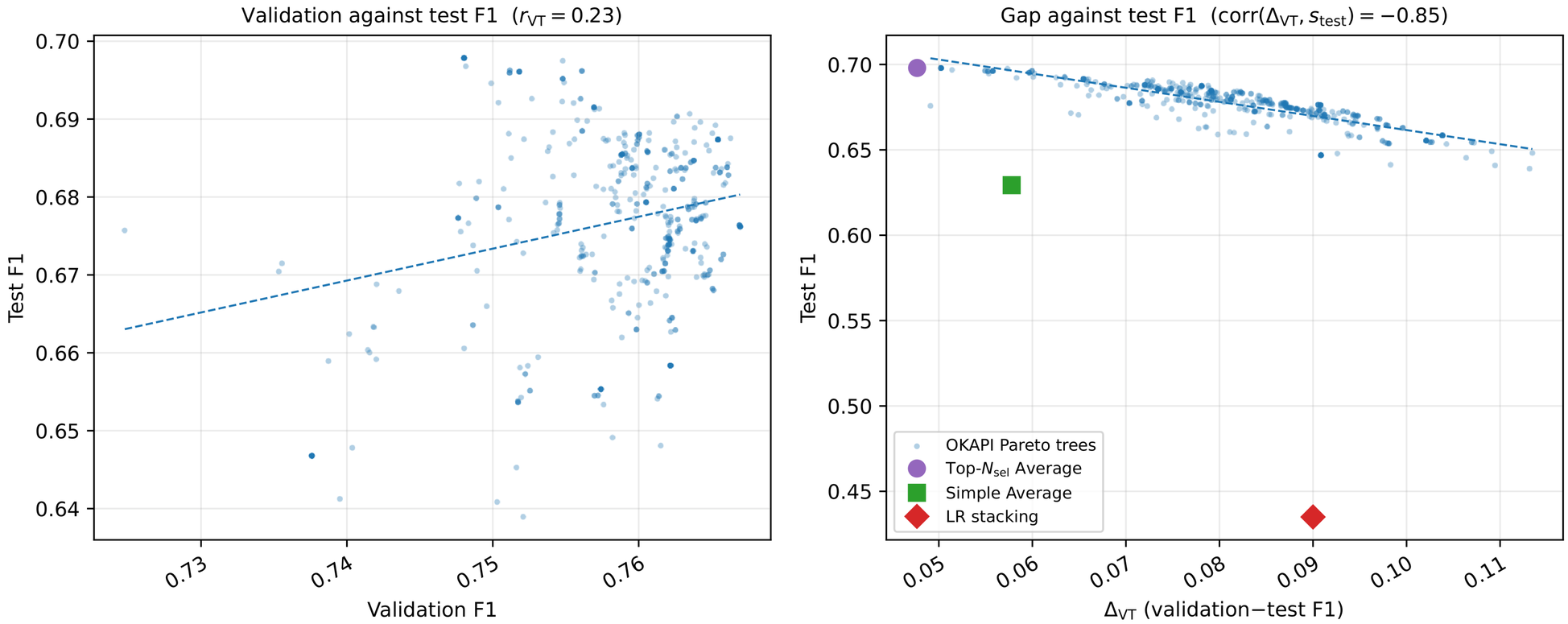


STARCOP: averaging a subset of models works best

Method	F1 on the test set
Average of a subset of models	0.698
OKAPI	0.679 ± 0.007
Selected single model	0.665
NSGA-III	0.657 ± 0.026
Simple Average	0.629
Logistic stacking	0.435

Models are ranked by the F1 score achieved on the validation set; this score also determines how many models to average.

STARCOP: validation gains do not always transfer

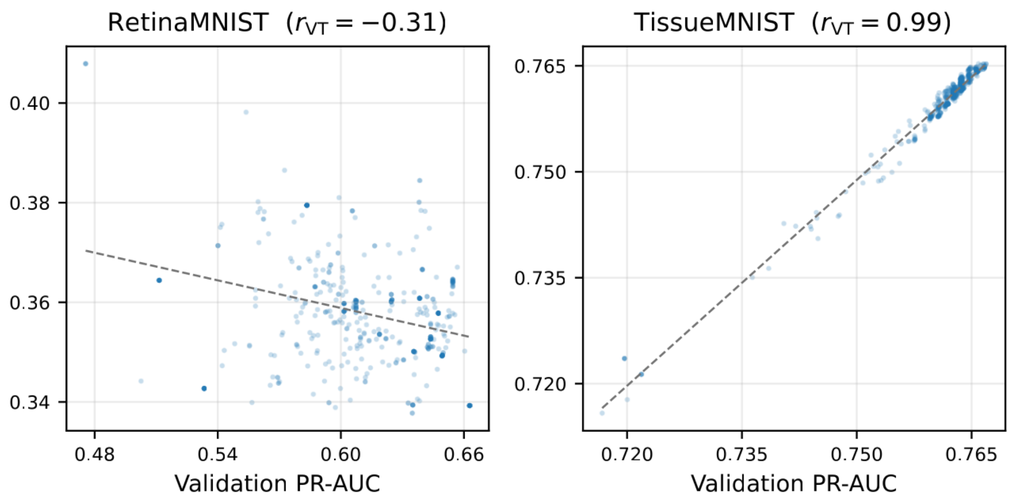


Pearson r = 0.23 for validation vs test F1 · Gap = validation − test

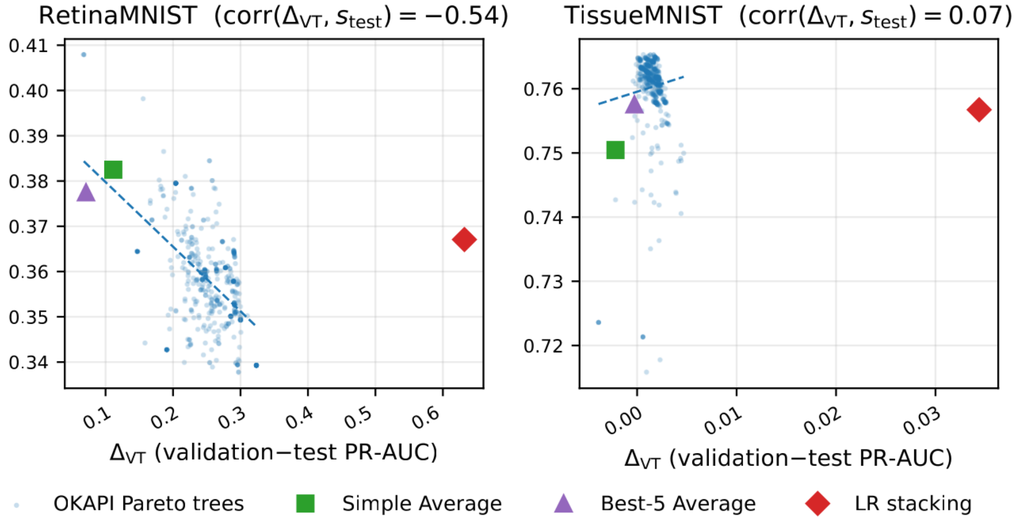
Source · Thesis, Fig. 5.4

MedMNISTv2: validation reliability varies

VALIDATION VS TEST



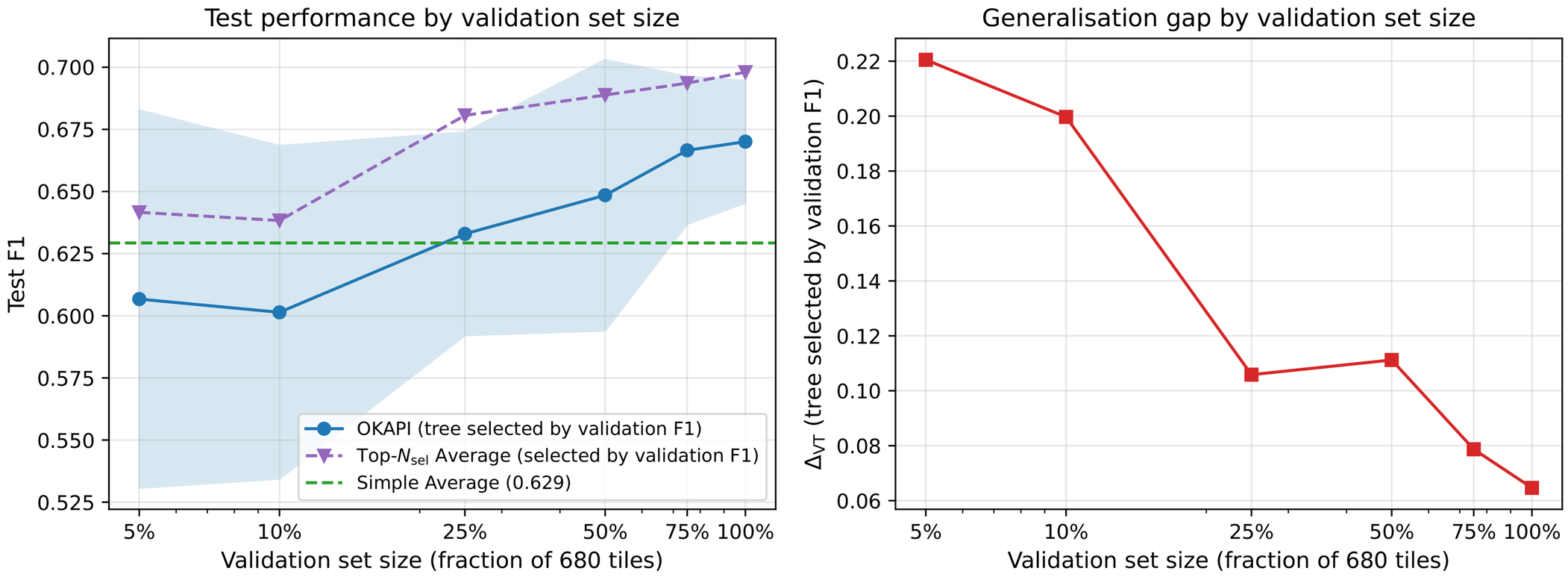
VALIDATION–TEST GAP VS TEST



RetinaMNIST and TissueMNIST · Gap = validation − test

Source · Thesis, Figs. 5.2 and 5.3

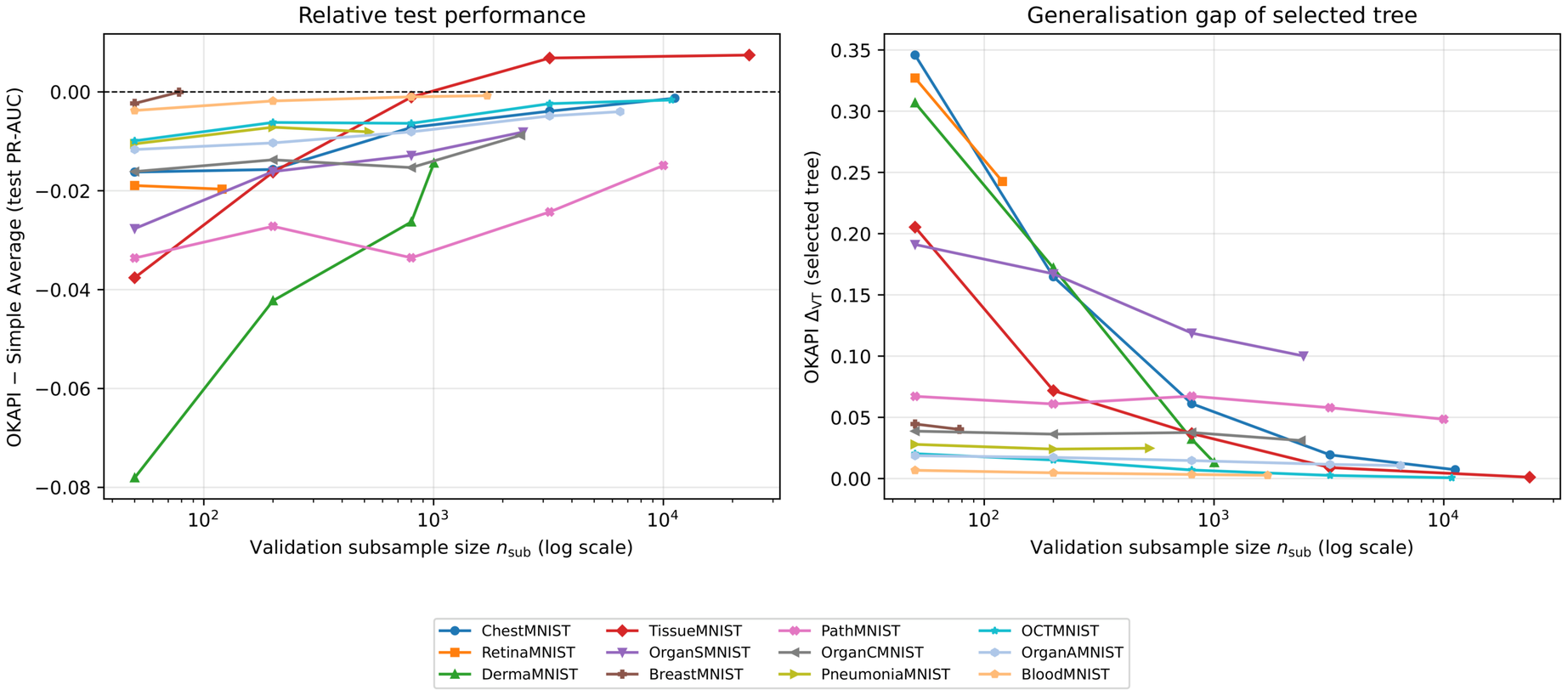
Validation set size and selection reliability for STARCOP



STARCOP: test F1 falls from 0.670 to 0.607 with 5% of validation tiles.

Source · Thesis, Fig. 5.5

Validation set size and selection reliability for MedMNISTv2



With 50 validation examples, the gap widens on all 12 datasets.

Source · Thesis, Fig. 5.6



Prior anchoring

$$Y\alpha = (1 - \alpha)\, T + \alpha\, P$$



Illustrated at α = 0.85

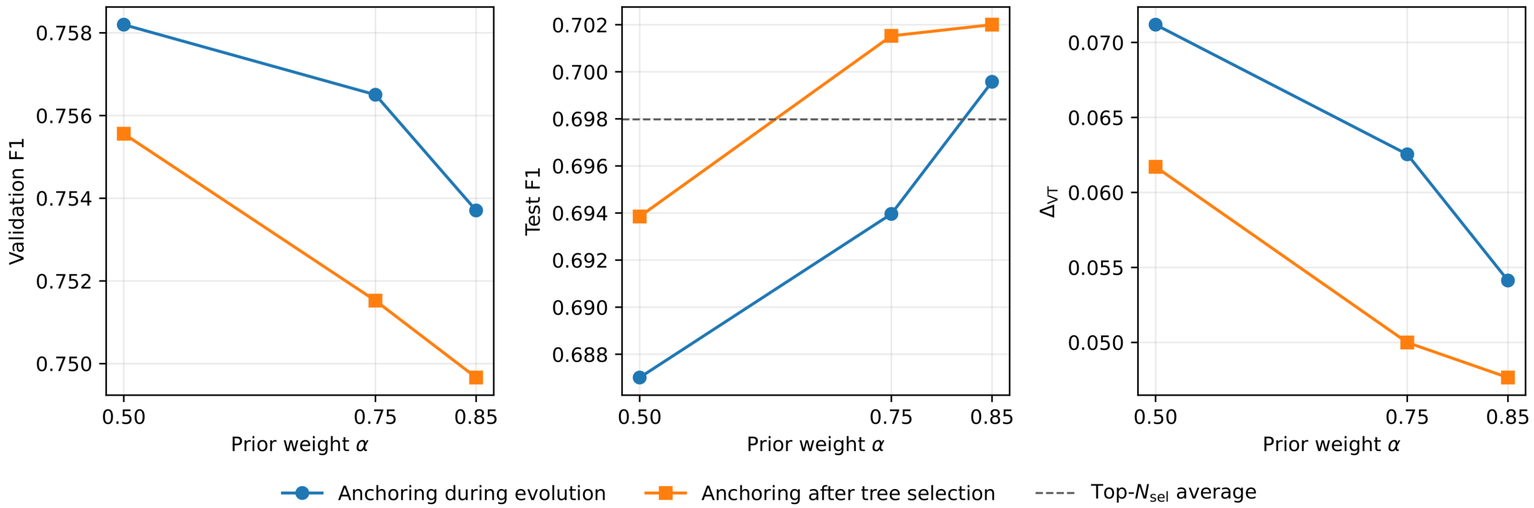
T = tree prediction · P = averaging prior · α = prior weight



Prior anchoring improves image benchmark results

STARCOP · validation, test and generalisation gap

Source · Thesis, Fig. 5.7



Test result at α = 0.85	Reference	During evolution	After selection
MedMNISTv2 PR-AUC	0.8047 prior	0.8054	0.8056
STARCOP F1	0.6980 prior	0.6996	0.7020



PMLB: evaluating OKAPI on tabular data

Penn Machine Learning Benchmarks

Data	136 tabular classification datasets common to all methods
Model pool	19 heterogeneous classifiers per dataset
Protocol	Stratified 60/20/20 split for training, validation and test
Metric	Mean test macro F1 across datasets

Does prior anchoring also help on tabular data?



PMLB results: moderate anchoring helped

Method	Mean test macro F1
TPOT	0.798
OKAPI, α = 0.50	0.795
Logistic stacking	0.795
OKAPI	0.794
Selected model average (prior)	0.792
Full Average	0.776

PRIOR · SELECTED MODELS

Average the top N models. For each dataset, choose N that gives the highest macro F1 on the validation set.



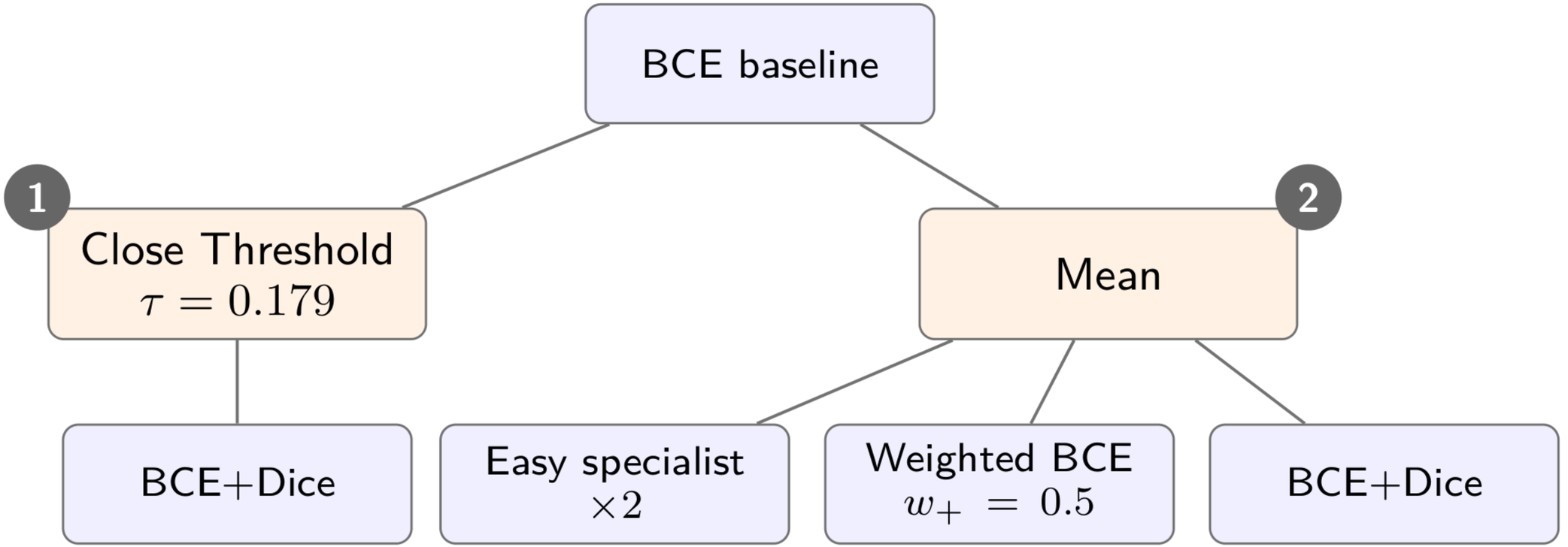
Parsimony reduces tree size

Benchmark	Mean tree nodes	Mean distinct models
MedMNISTv2	10.7 → 7.7	4.8 → 4.0
STARCOP	17.7 → 13.2	5.4 → 4.5

Without → with parsimony, at the same depth limit of 5. Test performance was similar.



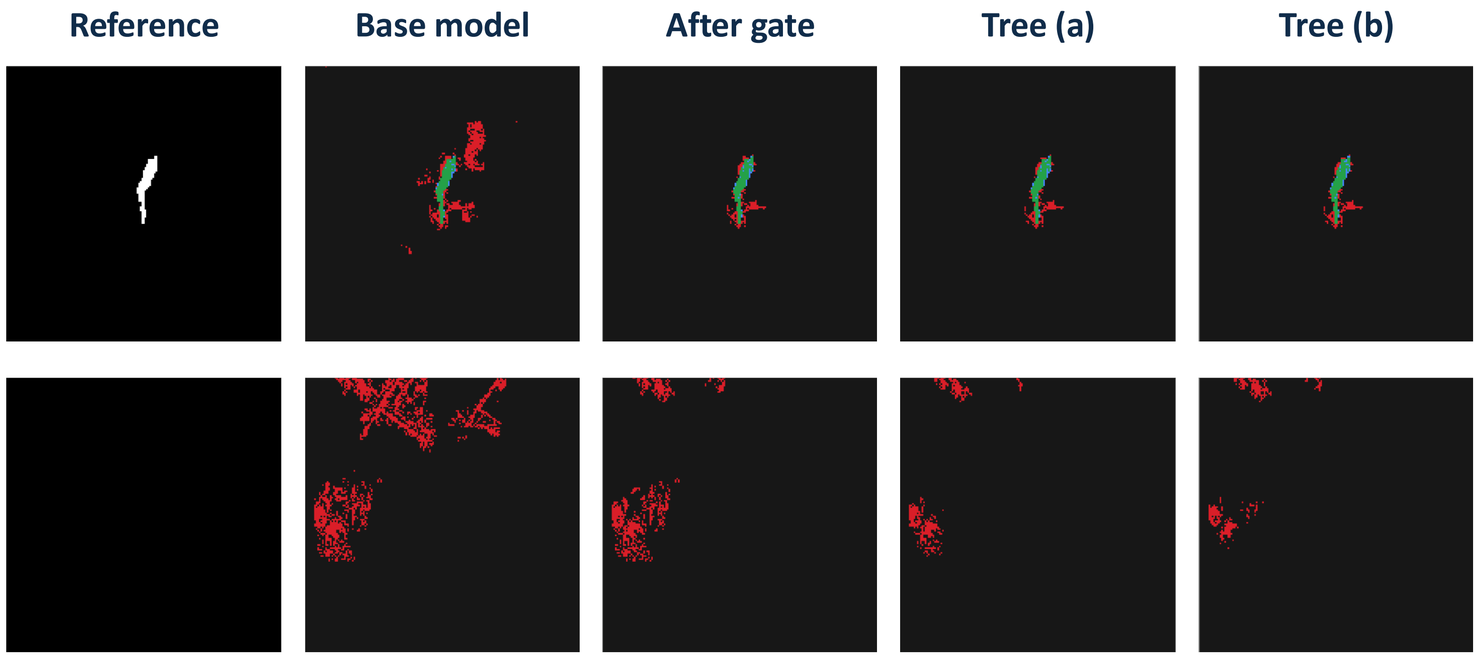
An evolved STARCOP procedure



BCE root · threshold gate · mean refinement

Source · Thesis, Fig. 6.1a

STARCOP: how fusion changes predictions



Model recruitment

Specialist: 95.2%

Individual F1: 0.486

Focal model: 8.8%

Individual F1: 0.669

■ Background ■ Correct plume ■ False alarm ■ Missed plume

Source · Thesis, Fig. 6.2 · Two representative test tiles

Evidence for the research hypotheses

H1 · Representation	Shared execution for classification and segmentation
H2 · Parsimony	Smaller trees at similar test performance
H3 · Validation size	Less validation data weakened transfer
H4 · Anchoring	Improved over matched unanchored OKAPI

Limitations and future work

Selection reliability	Estimate validation uncertainty before search
Search vs. representation	Separate the effects with controlled benchmarks
Deployment cost	Measure full ensemble inference, including the prior

Conclusions and discussion

> **OKAPI evolves model selection and fusion rules.**

Complexity	Parsimony keeps the fusion procedure smaller
Transfer	Validation and prior quality shape the gains

Thank you

Questions & discussion

Contact: contact@damiankucharski.com



Agata M. Wijata



Jakub Nalepa



Arkadiusz Czerwiński



Damian Kucharski