GeneNet

Gene co-expression networks & ensemble learning for cancer tissue classification

Turning complex RNA-Seq data into interpretable computational structure.



SHAMOEEL ALI + MUSAB ABBASI

C++ • DATA STRUCTURES • BIOINFORMATICS

The dimensionality challenge

A single tissue sample contains thousands of expression measurements — far more features than samples.

20,531
GENE FEATURES

801
PATIENT SAMPLES

5
CANCER CLASSES

WHY IT MATTERS

High dimensionality increases computational cost and makes relevant expression patterns difficult to isolate.

OUR RESEARCH QUESTION

Can we reduce, connect and classify gene expression efficiently?

From transcriptomics to network learning

Existing methods motivate GeneNet's combination of feature selection, gene relationships and ensemble prediction.

01

RNA-Seq / TCGA

High-throughput profiling reveals tissue-specific expression signatures.

02

Co-expression analysis

Correlation-based networks describe genes with similar expression patterns.

03

Ensemble learning

Bagged decision trees stabilize classification by aggregating predictions.

GENENET CONTRIBUTION **A custom, end-to-end C++ implementation connecting all three ideas.**

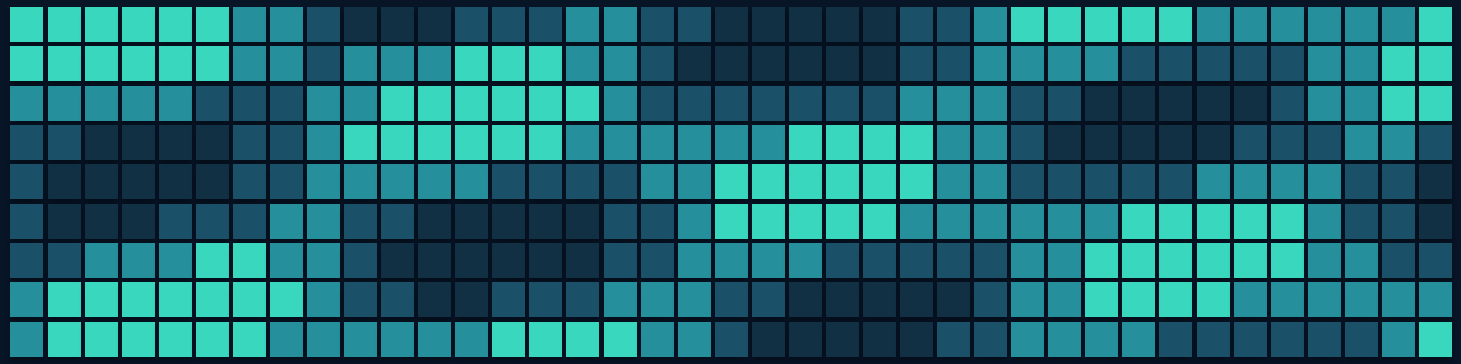
Foundations: TCGA Research Network (2013); Langfelder & Horvath (2008); Breiman (1996).

Five tissue classes, one expression matrix

PANCAN RNA-Seq: each row is a sample; each column is a measured gene-expression feature.

BRCA	LUAD	COAD	KIRC	PRAD
Breast	Lung	Colon	Kidney	Prostate

ILLUSTRATIVE EXPRESSION HEATMAP



samples →

Expression intensity varies across samples and genes.

Diagram is illustrative, not a heatmap computed from the dataset.

A five-stage computational workflow

GeneNet uses purpose-built data structures to move from raw measurements to tissue predictions.

01

READ

CSV matrix

Arrays / vectors

02

REDUCE

Variance → top 500

Merge sort

03

CONNECT

Pearson correlation

Hash table / graph

04

EXPLORE

Gene modules

BFS / DFS

05

PREDICT

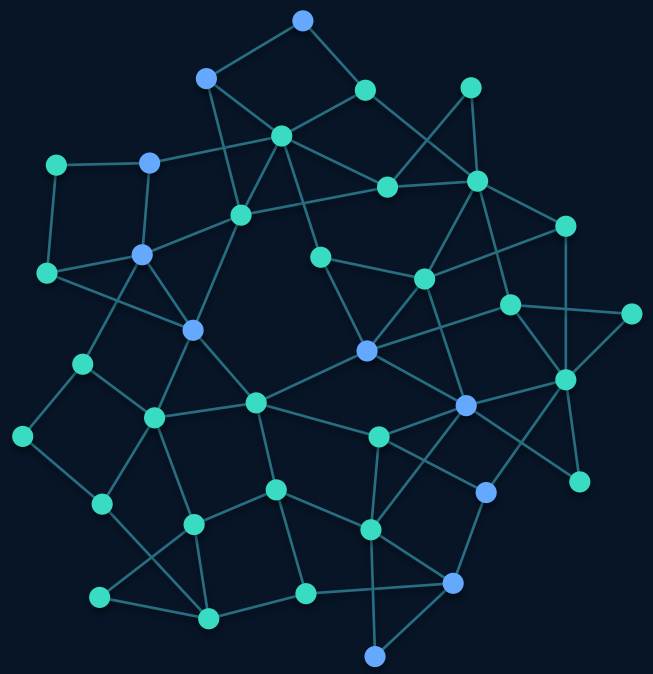
Five-class output

Trees + bagging

FEATURE REDUCTION **20,531 → 500 genes | 97.6% fewer features**

Genes become a co-expression network

An edge links two selected genes when the absolute Pearson correlation exceeds the project threshold (|r| ≥ 0.80).



500
GENE NODES

1,814
GRAPH EDGES

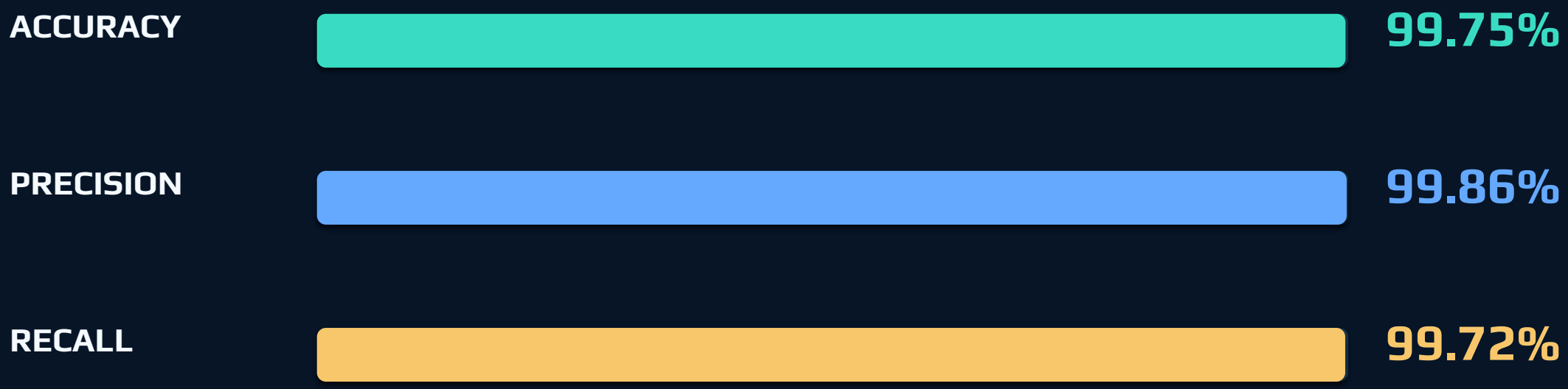
124,750
PAIRS EVALUATED

304
COMPONENTS

Interpretation: edges indicate co-variation, not causality.

Reported classification performance

Project-reported results on five tissue classes. Values have not been independently reproduced.



OUT-OF-BAG ERROR 3.87% Validation split and leakage controls require documentation.

What the results mean — and what they do not

Strong internal results demonstrate a functioning research prototype, not a validated clinical diagnostic.

DEMONSTRATED

A complete pipeline from RNA-Seq features to five-class predictions.

COMPUTATIONAL VALUE

A 97.6% reduction in feature count before correlation analysis.

LIMITATION

No independent external cohort or clinical validation reported.

NEXT SCIENTIFIC STEP

Cross-validation • external datasets • per-class confusion matrix • biological interpretation

From gene measurements to structured insight

GeneNet illustrates how core computer-science methods can support computational cancer research.

01

ORGANIZE

20,531 measurements

02

CONNECT

A sparse gene network

03

CLASSIFY

Five cancer tissue types

FUTURE DIRECTION Independent validation → robust evaluation → translational research

THANK YOU / QUESTIONS?

gene-net-gui.vercel.app