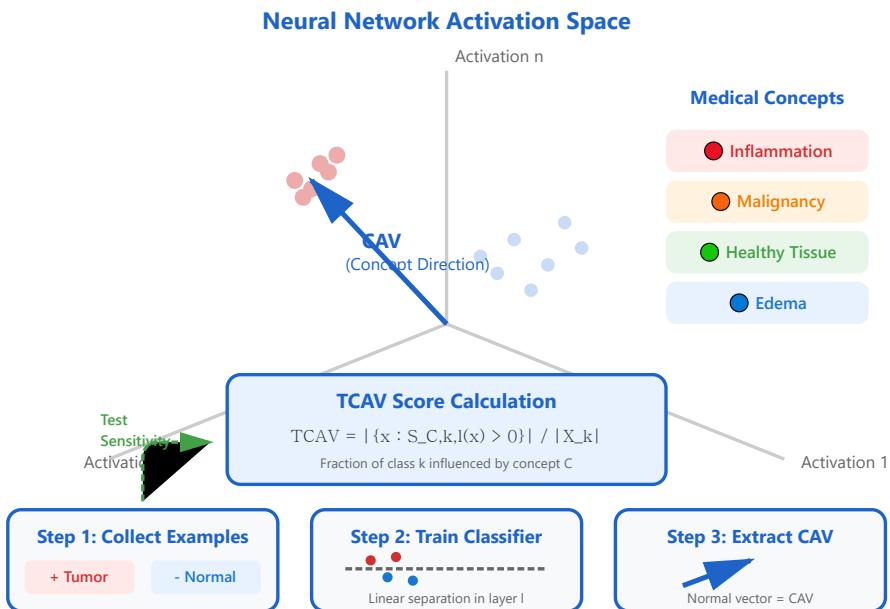


Concept Activation Vectors (CAV)



High-level Concepts

CAVs test whether neural networks learn human-interpretable concepts

→ Direction Vectors

Concepts represented as directions in activation space

→ Sensitivity Testing

Measure how much a concept influences model predictions

→ Medical Validation

Verify if models use clinically relevant concepts