BreastNet: Early Breast Cancer Detection

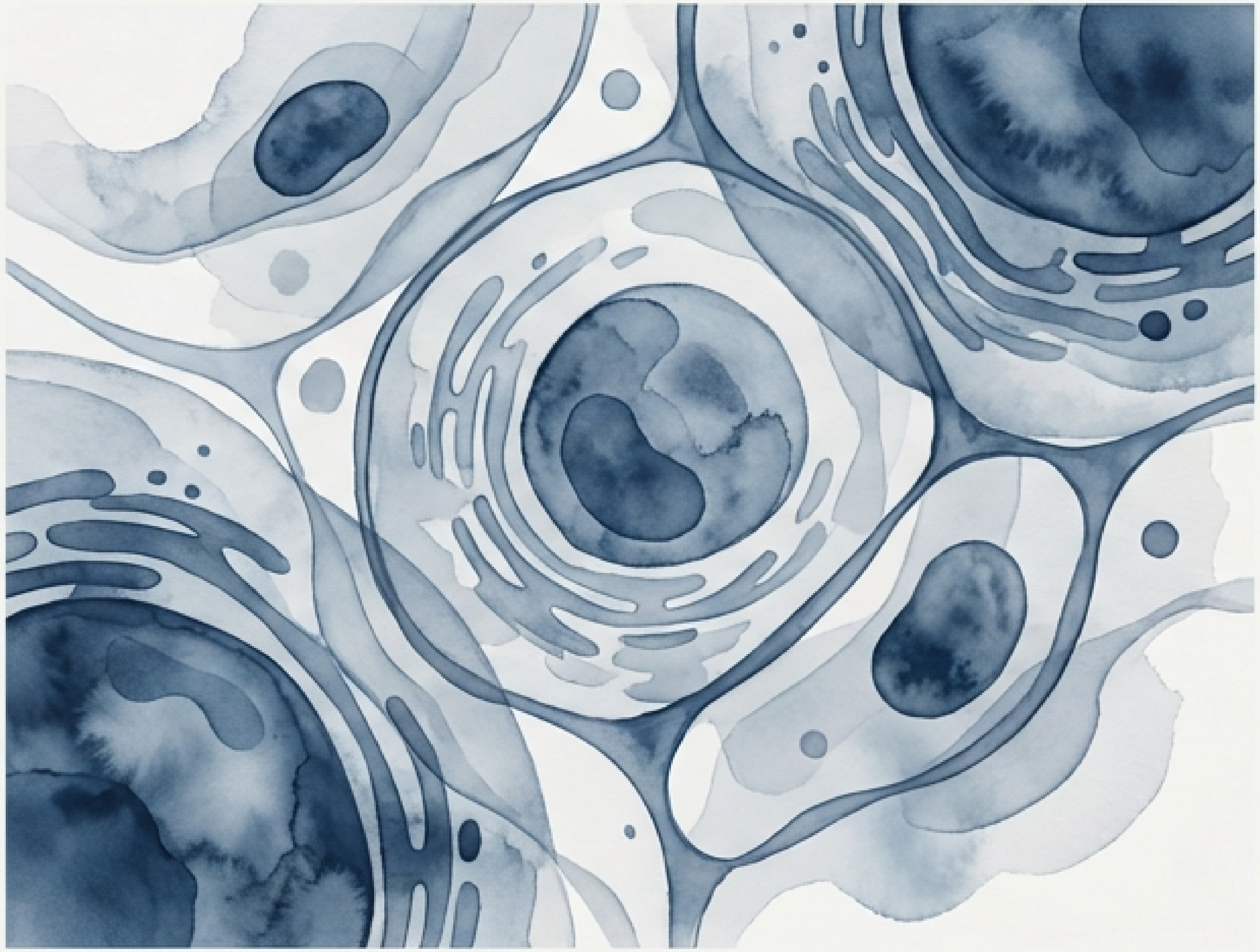
A neural network achieving **95.6%** accuracy and **100% sensitivity** for identifying malignant tumors.

The Challenge: Improving Early Detection in Breast Cancer

— Breast cancer remains a major global health issue where early detection is the single most important factor for improving survival rates.

— The traditional diagnostic process relies on manual interpretation of cell measurements, which can be time-consuming and subject to inconsistency.

— There is a critical need for tools that can support clinicians, increase efficiency, and provide consistent, data-driven analysis.



The Goal: A Fast and Reliable Clinical Decision Support Tool

Project Objective: To develop a neural network that automatically classifies tumors as benign or malignant based on structured numeric data from cell nuclei.

Why This Approach Matters:

Speed & Consistency

Algorithmic assistance can provide rapid, repeatable analysis to support pathologists.

Pattern Recognition

Machine learning can identify complex patterns across multiple features that are too subtle for manual inspection.

Improved Outcomes

Augmenting clinical workflows with AI can lead to earlier, more accurate detections.

The Foundation: The Breast Cancer (Wisconsin) Dataset

- **Source:** Kaggle (yasserh)
- **Total Samples:** 699
- **Features for Learning:** 9
- **Target:** 'Class' column (Benign or Malignant)

Annotation: All features are integer scores from 1-10. Higher values generally correlate with a higher risk of malignancy.

The Blueprint: A Structured Data Preprocessing Workflow

1. Load Data

Ingest CSV and inspect for quality.

→

2. Clean

Drop irrelevant "Id" column and remove rows with missing values.

→

3. Encode Target

Convert class labels from (2, 4) to (0, 1) where 0=Benign, 1️⃣ 1=Malignant.

→

4. Split

Partition data into 80% for training and 20% for testing.

→

5. Scale Features

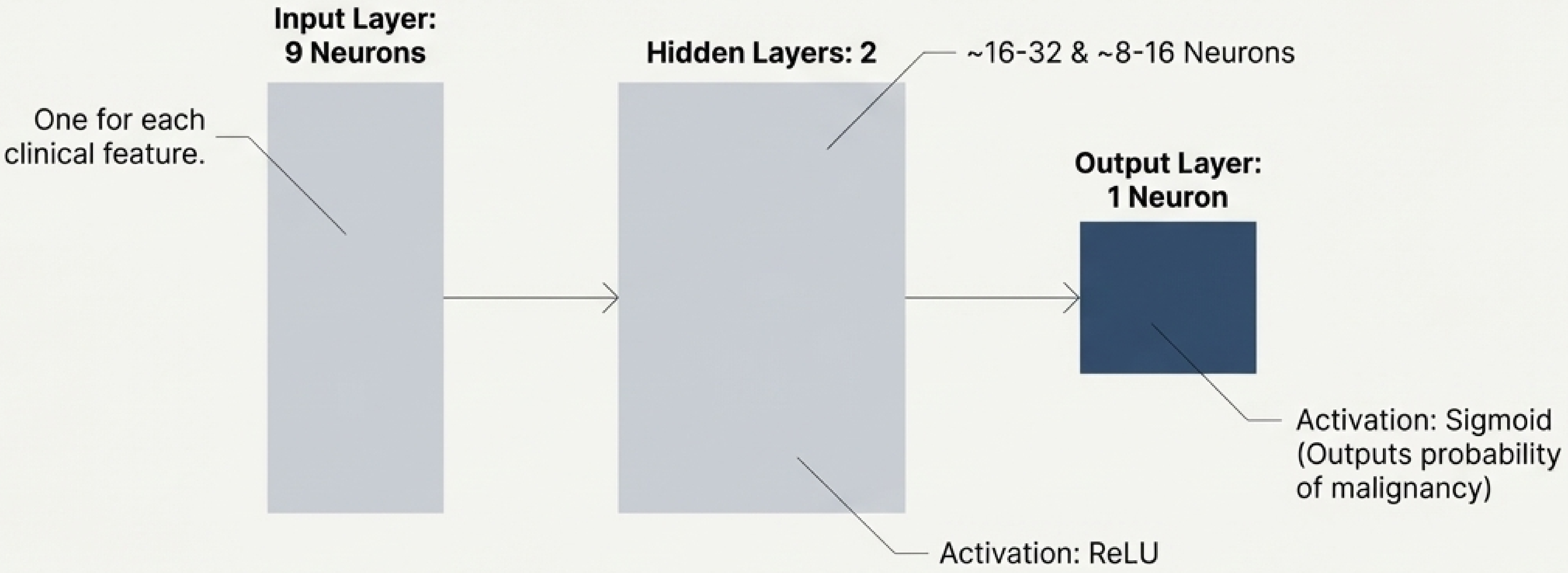
Apply `StandardScaler` to normalize feature ranges, improving model convergence.

→

6. Prepare Inputs

Convert scaled arrays into the final format for the neural network.

The Engine: BreastNet's Neural Network Architecture



Key Parameters

- **Loss Function:** Binary Cross-Entropy
- **Optimizer:** Adam / Gradient Descent
- **Learning Rate:** ~0.03

Rationale: A Design Focused on Simplicity and Performance

Why a simple architecture?

The dataset is small (699 samples). A shallow network avoids overfitting and promotes generalization.

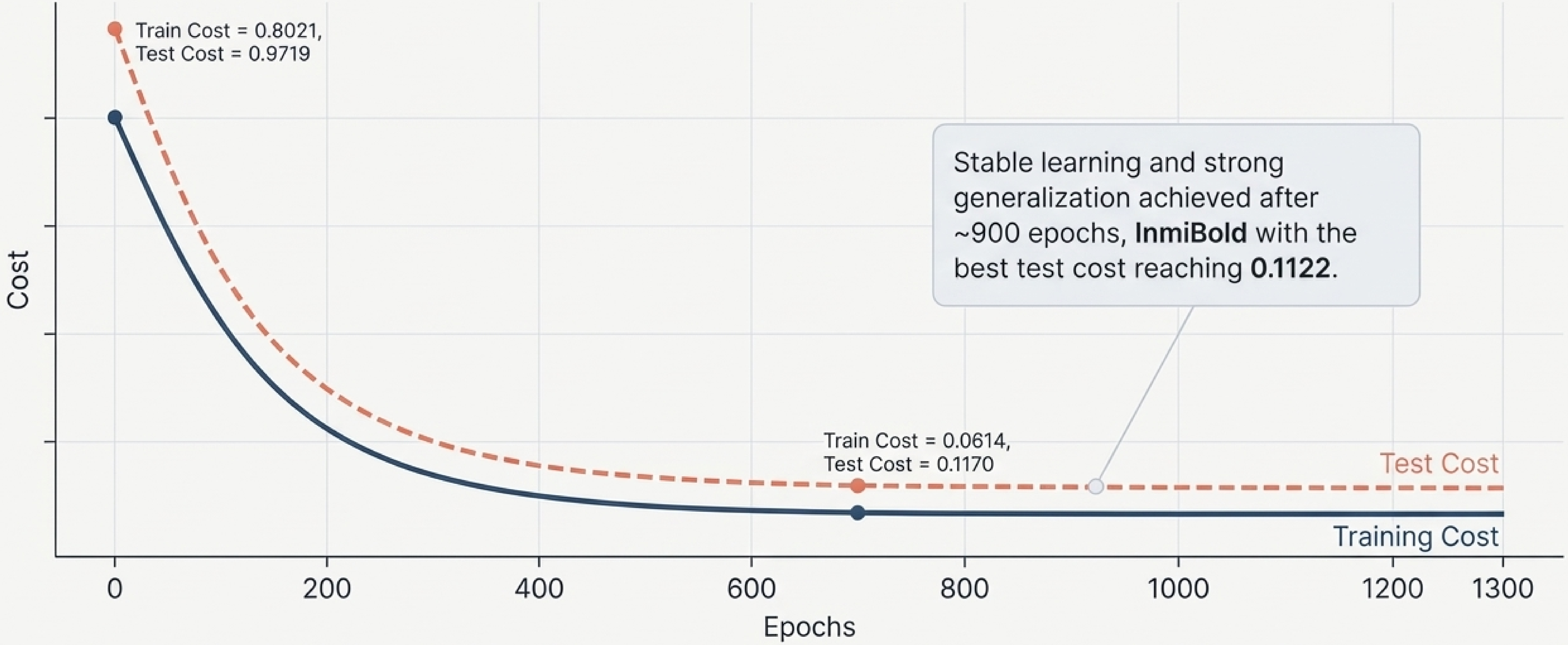
Why ReLU activation?

The Rectified Linear Unit is computationally efficient and helps speed up the training process compared to other functions.

Why Sigmoid output?

The Sigmoid function constrains the output to a value between 0 and 1, providing a true probability of malignancy that is easy to interpret.

Learning in Action: Model Convergence Over 1300 Epochs



The Results: A High-Performance Classification Model

95.6% Accuracy	**91.23%** F1 Score	**100%** Recall (Sensitivity)
83.87% Precision	**94.3%** Specificity	**0.1122** Final Test Cost

The model demonstrates a strong balance of accuracy and recall, showcasing its reliability on unseen test data.

The Critical Finding: The Model Never Misses a Malignant Case

		Actual	
		Benign	Malignant
Predicted	Benign	83 True Negatives	5 False Positives
	Malignant	0 False Negatives	26 True Positives

Zero false negatives on the test set is a critical outcome. For a cancer screening tool, avoiding missed cases (high sensitivity) is the highest priority.

Key Observations and Insights

Excellent recall is achievable: The model's architecture and training were highly effective at identifying all positive cases, which is the most important clinical requirement.

False positives are the main error type: The model is 'cautious,' preferring to misclassify a benign case as malignant (FP) rather than the other way around. This is a safer failure mode for a diagnostic aid.

Feature scaling is essential: Applying `StandardScaler` was critical for achieving stable training and convergence.

Training stabilizes predictably: The model learned effectively and stabilized after approximately 900 epochs, indicating a well-behaved training process.

An Honest Assessment: Project Limitations

Dataset Size: With only 699 samples, the model's ability to generalize to diverse real-world populations may be limited.

Data Modality: The model is based solely on tabular data. A comprehensive diagnosis in a real clinical setting incorporates medical imaging (e.g., mammograms).

Validation: This is a proof-of-concept model and has not been medically validated for clinical use.

Overfitting Risk: While the simple architecture mitigates this, the small dataset and long training time present a potential risk of overfitting.

The Road Ahead: Future Improvements and Directions

Model Enhancement

Explore deeper networks, regularization techniques (like Dropout), and conduct ROC/AUC analysis for threshold tuning.

Data Strategy

Use oversampling or class weights to address any data imbalance and validate the model on additional breast cancer datasets.

Multimodal Integration

Combine the current tabular data model with an image-based model trained on mammograms for a more holistic analysis.

Deployment

Develop the model into a web-based tool for demonstration and research purposes.

Conclusion: Augmenting Clinical Decisions with AI

- BreastNet successfully demonstrates the power of neural networks for early breast cancer detection using structured clinical data.
- The model achieves **95.6% accuracy** (in clini blue) and a clinically vital **100% sensitivity** (in warn coral) on the test set.
- It provides a framework for developing fast, consistent, and data-driven tools to support healthcare professionals.

This project shows how AI can serve as a powerful analytical tool, augmenting the expertise of clinicians and contributing to the goal of earlier, more effective cancer detection.

Tools and Technologies

Python

pandas

NumPy

Scikit-learn

Matplotlib

Custom Neural Network

Jupyter Notebook

Kaggle Breast Cancer Dataset (by yasserh)