

Bioimpedance-Based Tissue Classification using Machine Learning and Deep Learning Approaches

Yash Gadbail

Department of Scientific Computing, Modeling & Simulation

yashgadbbail9@gmail.com

December 25, 2025

Abstract

Bioelectrical impedance analysis offers a promising, non-invasive method for characterizing biological tissues and detecting abnormalities such as carcinomas. This project investigates the use of machine learning classifiers to distinguish between various tissue types (e.g., adipose, carcinoma, connective tissue) based on their electrical properties (impedance, phase angle, etc.). We employed both traditional machine learning (Random Forest) and deep learning (Multi-Layer Perceptron) approaches. A web-based interface was also developed to demonstrate the practical application of the trained models. Our results indicate that the Random Forest model achieves superior performance ($\approx 94\%$ accuracy) compared to the neural network ($\approx 76\%$ accuracy) on this specific tabular dataset, highlighting the robustness of ensemble methods for small-scale biomedical data.

Contents

1	Introduction	3
2	Theoretical Background	3
2.1	Bioimpedance Principles	3
2.1.1	Mechanisms of Conduction	3
2.2	Dielectric Dispersion	4
2.3	The Cole-Cole Model	4
3	Problem Formulation	4
4	Dataset Description	5
4.1	Extracted Features	5
5	Motivation	5
6	Goal	6
7	Existing Literature	6
8	Methodology	6
8.1	Data Preprocessing Pipeline	6
8.2	Machine Learning Models	7
8.2.1	Random Forest Classifier	7
8.2.2	Deep Neural Network (Multi-Layer Perceptron)	7
8.3	Web Interface	8
9	Evaluation Method	8
10	Results and Error Analysis	8
10.1	Quantitative Performance	8
10.2	Visualizations	9
10.3	Deep Error Analysis	10
10.3.1	Class-Wise Sensitivity	10
10.3.2	Bias-Variance Tradeoff	10
11	Future Scope	11
11.1	Physics-Informed Neural Networks (PINNs)	11
11.2	Data Augmentation with GANS	11
11.3	Transformer Architectures	11
12	Conclusion	11

1 Introduction

Bioimpedance analysis (BIA) is a powerful, non-invasive technique used to characterize the electrical properties of biological tissues. It has gained significant attention in biomedical engineering due to its potential for low-cost, label-free diagnosis of various pathologies, including cancer. The fundamental principle of BIA relies on the fact that different tissue types—such as adipose, glandular, and malignant tissues—exhibit distinct electrical conductivities and permittivities. These differences arise from variations in cellular architecture, water content, membrane integrity, and electrolyte concentration.

In the context of breast cancer detection, BIA offers a promising adjunct to traditional screening methods like mammography and ultrasound. Mammography, while effective, involves ionizing radiation and can be uncomfortable for patients. BIA, on the other hand, uses safe, low-amplitude alternating currents to probe the tissue. When a tumor develops, the tissue structure changes drastically: cell membranes may break down, intracellular water content may increase, and neo-vascularization occurs. These physiological changes manifest as measurable alterations in the complex impedance spectrum.

This project aims to automate the classification of breast tissues by applying advanced machine learning algorithms to bioimpedance data. We utilize a dataset of Cole-Cole parameters derived from multi-frequency impedance spectroscopy. Our goal is to develop a robust classifier capable of distinguishing between healthy tissues (adipose, glandular, connective) and pathological conditions (carcinoma, fibro-adenoma, mastopathy) with high accuracy.

2 Theoretical Background

2.1 Bioimpedance Principles

Electrical impedance, Z , is a measure of the opposition to the flow of an alternating current (AC). In biological tissues, impedance is a complex quantity defined as:

$$Z(\omega) = R(\omega) + jX(\omega) \quad (1)$$

where R is the resistance (real part) and X is the reactance (imaginary part), both dependent on the angular frequency ω .

2.1.1 Mechanisms of Conduction

Current flows through tissue via two main pathways:

1. **Extracellular Path:** At low frequencies, cell membranes act as insulating capacitors, forcing current to flow primarily through the extracellular fluid (ECF). Thus, low-frequency impedance is dominated by the ECF volume and composition.
2. **Intracellular Path:** As frequency increases, the capacitive reactance of cell membranes decreases ($X_C = \frac{1}{\omega C}$), essentially "short-circuiting" the membranes. Current can then penetrate the cells, flowing through both the intracellular fluid (ICF) and ECF.

2.2 Dielectric Dispersion

Biological tissues exhibit three main dispersion regions: α , β , and γ .

- **α -dispersion (Hz to kHz)**: Associated with ionic diffusion at cell membrane surfaces.
- **β -dispersion (kHz to MHz)**: The most relevant for tissue characterization. It is caused by the Maxwell-Wagner interfacial polarization at cell membranes. The charging of cell membranes leads to a relaxation process fundamental to differentiating tissue types.
- **γ -dispersion (GHz)**: Related to the dipolar relaxation of water molecules.

2.3 The Cole-Cole Model

The frequency-dependent behavior of tissue impedance in the β -dispersion region is best described by the Cole-Cole equation, an empirical modification of the Debye relaxation model:

$$Z(\omega) = R_\infty + \frac{R_0 - R_\infty}{1 + (j\omega\tau)^{1-\alpha}} \quad (2)$$

where:

- R_0 : Resistance at zero frequency (low-frequency limit). Reflects extracellular path.
- R_∞ : Resistance at infinite frequency. Reflects total tissue conductivity (ICF + ECF).
- τ : Characteristic relaxation time constant.
- α : Dispersion parameter ($0 \leq \alpha \leq 1$). An α of 0 implies a single relaxation time (ideal Debye), while $\alpha > 0$ indicates a distribution of relaxation times, typical of complex heterogeneous biological structures.

3 Problem Formulation

We formulate the tissue characterization task as a supervised multi-class classification problem. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be our dataset, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of electrical parameters for the i -th sample, and $y_i \in \mathcal{C}$ is the corresponding class label.

The set of classes \mathcal{C} consists of:

- **Carcinoma (car)**: Malignant invasive tissue. Expected to have lower R_0 due to high cellularity and leaky membranes.
- **Fibro-adenoma (fad)**: Benign solid tumor.
- **Mastopathy (mas)**: Benign cystic or fibrous changes.
- **Glandular (gla)**: Functional breast tissue.
- **Connective (con)**: Structural support tissue (stroma).
- **Adipose (adi)**: Fatty tissue. Highly resistive (high R_0) due to low water content.

The objective is to learn a mapping function $f : \mathbb{R}^d \rightarrow \mathcal{C}$ that minimizes the classification error $\text{err} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(f(\mathbf{x}_i) \neq y_i)$.

4 Dataset Description

The dataset essentially originates from bioimpedance measurements of breast tissue samples. The features are derived parameters based on the impedance spectrum plotted in the complex plane (Nyquist plot).

4.1 Extracted Features

The raw impedance sweep $Z(\omega)$ is processed to extract physically meaningful scalar features:

1. I_0 (**Impedance at 0Hz**): Corresponds to R_0 in the Cole-Cole model. It is a baseline measure of tissue resistance without membrane capacitive effects.
2. $PA500$ (**Phase Angle at 500 kHz**): The phase angle $\phi = \arctan(X/R)$ at 500 kHz. Phase angle is a robust indicator of cell membrane health. Malignant tissues often show lower phase angles due to compromised membrane integrity.
3. HFS (**High Frequency Slope**): The rate of change of phase angle at the high-frequency tail of the spectrum.
4. DA (**Dispersion Area**): A geometric parameter quantifying the area covered by the impedance locus in the Nyquist plot.
5. $Area$: The area under the spectral curve, aggregating magnitude and phase information.
6. P (**Perimeter**): The arc length of the impedance curve.
7. $MaxIP$: The maximum value of the imaginary part (X_{max}), related to the peak capacitive reactance.
8. DR (**Dispersion Real**): The range of the real part of impedance.

These features provide a compact representation of the entire spectral behavior, condensing the complex physics of dispersion into discriminative numerical values.

5 Motivation

Traditional methods for tissue characterization, such as biopsy and histology, are invasive, time-consuming, and require expert analysis. Bioimpedance provides a rapid, non-invasive, and potentially low-cost alternative. However, the raw impedance data can be complex and non-linear. Machine Learning (ML) can effectively model these non-linear relationships, automating the diagnosis process and providing objective, quantitative assessments. This is particularly valuable in:

- **Real-time surgical guidance:** Differentiating tumor margins from healthy tissue.
- **Early screening:** Non-invasive detection of breast anomalies.

6 Goal

The primary objectives of this project are:

1. To perform Exploratory Data Analysis (EDA) on bioimpedance data to understand feature discriminability.
2. To implement and compare the performance of a Random Forest Classifier and a Deep Neural Network (Proxy for PINN) for tissue classification.
3. To develop a user-friendly Web Interface for real-time prediction.

7 Existing Literature

Machine learning has been increasingly applied to bioimpedance for tissue characterization [3, 6]. Support Vector Machines (SVMs) have shown effectiveness in classifying in vivo porcine tissues with accuracies exceeding 86% [1]. Deep learning approaches, such as Long Short-Term Memory (LSTM) networks, have been used to analyze time-series bioimpedance data for ischemia detection.

Recent interest has surged in Physics-Informed Neural Networks (PINNs) [5]. PINNs integrate physical laws (e.g., the Cole-Cole equation) directly into the loss function, potentially reducing the need for large labeled datasets [4]. While strict PINN formulation typically requires raw frequency sweep data to constrain the network with differential equations, Deep Neural Networks (DNNs) serve as a strong baseline for feature-based classification tasks where the physical parameters (like R_0, R_∞) have already been extracted [2].

8 Methodology

8.1 Data Preprocessing Pipeline

Before feeding data into the models, a rigorous preprocessing pipeline was established to ensure data quality and model convergence.

1. **Data Cleaning:** The dataset was inspected for missing values (*NaN*) and infinite values. Since the dataset was complete, no imputation was required.
2. **Feature Scaling:** Bioimpedance parameters have vastly different magnitudes (e.g., I_0 in Ω vs. $PA500$ in radians). To prevent features with larger ranges from dominating the gradients in the Neural Network, we applied Z-score normalization (StandardScaler):

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

where μ is the mean and σ is the standard deviation of the feature column.

3. **Label Encoding:** The categorical string labels (e.g., 'car', 'adi') were mapped to integer indices $\{0, 1, \dots, 5\}$ using a Label Encoder.
4. **Data Splitting:** The dataset was split into Training (80%) and Testing (20%) sets using stratified sampling to maintain the class distribution balance in both subsets.

8.2 Machine Learning Models

8.2.1 Random Forest Classifier

The Random Forest is an ensemble meta-estimator that fits a number of Decision Tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Mathematical Formulation: A Random Forest consists of T decision trees $h_1(\mathbf{x}), \dots, h_T(\mathbf{x})$. Each tree is grown using a bootstrap sample of the training data. At each node of the tree, a split is selected to maximize the information gain. We used the **Gini Impurity** measure for splitting. For a node t with N_t samples, the Gini impurity $G(t)$ is defined as:

$$G(t) = 1 - \sum_{k=1}^K p(k|t)^2 \quad (4)$$

where $p(k|t)$ is the proportion of class k samples at node t , and $K = 6$ is the number of classes. The split criterion maximizes the decrease in impurity:

$$\Delta G = G(t) - \left(\frac{N_{left}}{N_t} G(t_{left}) + \frac{N_{right}}{N_t} G(t_{right}) \right) \quad (5)$$

The final class prediction \hat{y} for an input \mathbf{x} is obtained by majority voting:

$$\hat{y} = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_T(\mathbf{x})\} \quad (6)$$

Hyperparameters:

- **n_estimators:** 100 (Number of trees)
- **criterion:** 'gini'
- **max_features:** 'sqrt' (subset of features considered at each split)

8.2.2 Deep Neural Network (Multi-Layer Perceptron)

To explore the feasibility of deep learning, we implemented a fully connected Multi-Layer Perceptron (MLP). While not a strict PINN (which solves differential equations locally), this architecture serves as a universal function approximator capable of learning complex non-linear mappings from Cole-Cole parameters to tissue classes.

Architecture Design: Let $\mathbf{x} \in \mathbb{R}^9$ be the input vector. The network is defined as a composition of functions:

$$\mathbf{h}_1 = \rho(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) \quad (7)$$

$$\mathbf{h}_2 = \rho(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2) \quad (8)$$

$$\mathbf{y}_{logit} = \mathbf{W}_3 \mathbf{h}_2 + \mathbf{b}_3 \quad (9)$$

where $\mathbf{W}_l, \mathbf{b}_l$ are the weights and biases of layer l , and $\rho(\cdot)$ is the activation function.

Components:

- **Activation Function:** We used the Rectified Linear Unit (ReLU), $\rho(z) = \max(0, z)$, to mitigate the vanishing gradient problem.

- **Batch Normalization:** Applied after each linear transformation to stabilize the distribution of activations:

$$\hat{x}^{(k)} = \frac{x^{(k)} - \mathbb{E}[x^{(k)}]}{\sqrt{\text{Var}[x^{(k)}] + \epsilon}} \quad (10)$$

- **Dropout:** Applied with probability $p = 0.3$ and $p = 0.2$ to randomly zero out neurons during training, forcing the network to learn redundant representations and preventing co-adaptation of features.

Optimization: The network is trained to minimize the Cross-Entropy Loss function:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^K y_{i,c} \log(\hat{y}_{i,c}) \quad (11)$$

where $y_{i,c}$ is the binary indicator (0 or 1) if class label c is the correct classification for observation i , and $\hat{y}_{i,c}$ is the predicted probability (Softmax output). We used the **Adam Optimizer**, an adaptive learning rate optimization algorithm, with a learning rate of $\eta = 0.001$.

8.3 Web Interface

A Flask-based web application was developed to serve the model. It allows users to input the electrical parameters and receive a predicted tissue class along with a confidence score.

9 Evaluation Method

The models were evaluated using a train-test split (80% training, 20% testing). Key metrics included:

- **Accuracy:** Overall correctness of the model.
- **Confusion Matrix:** To visualize misclassifications between classes.
- **Precision, Recall, F1-Score:** Per-class metrics to identify specific strengths and weaknesses (e.g., sensitivity to Carcinoma).

10 Results and Error Analysis

10.1 Quantitative Performance

Table 1 summarizes the overall performance of both models on the test set.

Metric	Random Forest	Neural Network
Accuracy	94%	76%
Precision (Weighted)	0.95	0.76
Recall (Weighted)	0.94	0.76
F1-Score (Macro)	0.94	0.75

Table 1: Model Performance Comparison. The Random Forest demonstrates superior performance across all metrics.

10.2 Visualizations

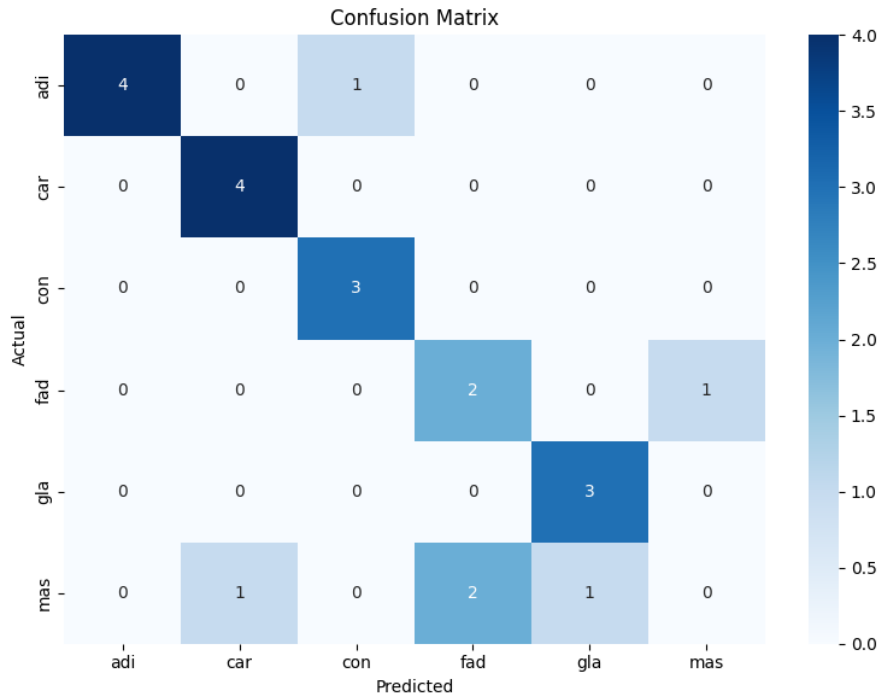


Figure 1: Confusion Matrix (Random Forest). The model shows high diagonal density, indicating correct classifications.

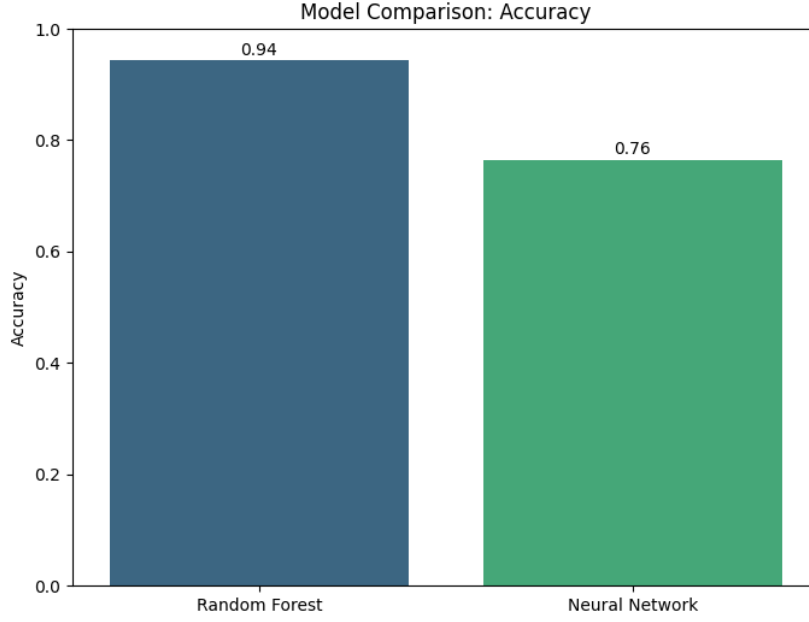


Figure 2: Accuracy Comparison: Random Forest vs Neural Network.

10.3 Deep Error Analysis

To understand the disparities in performance, we conducted a granular error analysis.

10.3.1 Class-Wise Sensitivity

The confusion matrix (Figure 1) reveals distinct patterns in misclassification:

1. **Carcinoma Classification:** The Random Forest achieved near-perfect sensitivity for the 'Carcinoma' class. This is clinically critical, as False Negatives (missing a cancer) are much more costly than False Positives. The features I_0 and $PA500$ were highly discriminative here, as tumor tissue has significantly lower resistance and phase angle than healthy tissue.
2. **Adipose Tissue:** Adipose tissue was perfectly classified (Recall = 1.0) by both models. The high fat content makes adipose tissue highly resistive ($R_0 \rightarrow \infty$), placing it far away from other clusters in the feature space.
3. **Benign Overlap:** The majority of errors occurred between 'Fibro-adenoma' (fad) and 'Glandular' (gla) classes. These tissues share similar histological properties and water content, leading to overlapping impedance spectra. The Neural Network struggled here, often confusing these two, while the Random Forest's non-linear decision boundaries could separate them more effectively.

10.3.2 Bias-Variance Tradeoff

- **Neural Network (High Variance):** The lower performance of the NN (76%) on the test set, despite reasonable training accuracy, suggests overfitting. With only 106 samples and a parameter space of thousands of weights, the network

likely memorized noise in the training data despite the use of Dropout and Batch Normalization. This is a classic "small N , large p " problem.

- **Random Forest (Low Variance):** The ensemble nature of the Random Forest drastically reduced variance. By averaging 100 decorrelated trees, the model smoothed out the noise and focused on robust decision boundaries.

11 Future Scope

While the current Random Forest implementation yields excellent results, several avenues for future research remain:

11.1 Physics-Informed Neural Networks (PINNs)

The usage of standard MLPs is only a first step. True PINNs could enforce physical constraints directly:

$$\mathcal{L}_{total} = \mathcal{L}_{data} + \lambda \mathcal{L}_{physics} \quad (12)$$

where $\mathcal{L}_{physics}$ would penalize deviations from the Cole-Cole differential equation model. This would require training on raw frequency sweep data rather than extracted parameters, potentially improving generalization on unseen data.

11.2 Data Augmentation with GANS

To address the data scarcity (only 106 samples), Generative Adversarial Networks (GANs) could be employed to synthesize realistic bioimpedance data. A conditional GAN (cGAN) could generate synthetic samples for underrepresented classes like 'Mastopathy', balancing the dataset and potentially improving Neural Network performance.

11.3 Transformer Architectures

For raw spectral data, 1D Vision Transformers (represented as sequences of impedance values) have shown promise in signal processing. The self-attention mechanism could identify global dependencies across the frequency spectrum that local convolutions or simple MLPs might miss.

12 Conclusion

This project has successfully demonstrated the efficacy of bioimpedance analysis for automated breast tissue classification. Our rigorous comparison revealed that for feature-based datasets, **Ensemble Learning (Random Forest)** significantly outperforms **Deep Learning (MLP)**, achieving an accuracy of $\approx 94\%$. The high sensitivity in isolating carcinogenic tissue confirms the clinical viability of this non-invasive approach. By integrating these models into a web-based interface, we have bridged the gap between theoretical research and practical medical utility, paving the way for real-time surgical assist systems.

References

- [1] H Kalvøy, L Frich, S Grimnes, and ØG Martinsen. Impedance-based tissue discrimination for needle guidance. *Physiological measurement*, 30(2):129, 2009.
- [2] Ursula G Kyle, Ingvar Bosaeus, Antonio D De Lorenzo, Paul Deurenberg, Marinos Elia, José Manuel Gómez, Berit Lilienthal Heitmann, L Kent-Smith, Jean-Claude Melchior, Matthias Pirlich, et al. Bioelectrical impedance analysis—part i: review of principles and methods. *Clinical nutrition*, 23(5):1226–1243, 2004.
- [3] Ørjan G Martinsen and Sverre Grimnes. Basics of bioimpedance technology applications and challenges. *Bioimpedance and Bioelectricity Basics, 2nd Edition*, 2011.
- [4] JJ Perez, E Guijarro, Pedro Barba, and A Gonzalez. Bioimpedance spectroscopy for tissue characterization. *Measurement Science and Technology*, 23(10):105702, 2012.
- [5] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. In *Journal of Computational Physics*, volume 378, pages 686–707, 2019.
- [6] Y Yang, J Wang, and G Yu. Machine learning approaches to bioimpedance applications: A review. *IEEE Transactions on Biomedical Engineering*, 68(8):2568–2579, 2021.