Project Report

Breast Cancer Prediction Using Artificial Neural Networks & Genetic Algorithm

NAME: S. Varshitha

HT.NO: 2203A51519

1. Objective / Problem Statement

Breast cancer remains one of the most critical and prevalent health concerns, particularly affecting women across the globe. With millions of new cases reported each year, early detection and accurate diagnosis play a crucial role in determining the prognosis and effectiveness of treatment options. Traditional diagnostic methods, while effective, can be time-consuming and sometimes yield inconclusive results.

This project aims to develop an intelligent computational model to predict whether a breast tumour is benign or malignant using clinical diagnostic data. The ultimate goal is to assist medical professionals in making quicker and more accurate decisions. The core methodology involves the use of Artificial Neural Networks (ANN), a type of deep learning model known for its ability to learn and identify complex, non-linear relationships in data. To further enhance the performance of the ANN, a Genetic Algorithm (GA) is employed for feature selection. This bio-inspired optimization technique mimics the process of natural selection to identify the most relevant features, thereby improving accuracy and reducing computational complexity.

Through this project, we intend to demonstrate that the synergy between neural networks and genetic algorithms can offer powerful tools for the medical community, particularly in the early diagnosis of breast cancer.

2. Dataset Used

2.1 Source and Format

The dataset used in this project is derived from real-world clinical records and is provided in a simplified CSV format named simplified_dataset.csv. This dataset is representative of the types of information collected during standard breast cancer diagnostic procedures.

2.2 Features

The dataset contains a variety of features extracted from patient diagnostic data. Key features include:

- **Radius Mean**: Average distance from the centre to points on the perimeter
- **Texture Mean**: Standard deviation of Gray-scale values

- **Perimeter Mean**: Length of the tumour border
- **Area Mean**: Size of the tumour region
- **Smoothness Mean**: Local variation in radius lengths

Other features include concavity, symmetry, fractal dimension, and more, each offering unique insights into the tumour's characteristics.

2.3 Target Variable

The target variable is a binary classification label:

- **0**: Benign tumour
- **1**: Malignant tumour

The balanced distribution of the target classes ensures that the model is not biased towards one class over another.

2.4 Preprocessing

Before training the model, the dataset underwent several preprocessing steps:

- **Handling Missing Values**: Checked and cleaned any missing or inconsistent entries
- **Encoding Categorical Data**: Applied Label Encoding and One-Hot Encoding
- **Feature Scaling**: Normalized numerical features using StandardScaler
- **Train-Test Split**: Divided the data into 80% training and 20% testing sets

---

3. Model Details

3.1 Artificial Neural Network Architecture

The ANN used in this project is a multi-layer perceptron consisting of the following components:

- **Input Layer**: Accepts the features selected from the dataset
- **Hidden Layers**: Two dense layers with ReLU activation functions
- **Dropout Layers**: Applied between hidden layers to prevent overfitting
- **Output Layer**: Single neuron with sigmoid activation for binary classification

3.2 Hyperparameters

- **Optimizer**: Adam
- **Loss Function**: Binary Cross entropy
- **Epochs**: 100

- **Batch Size**: 10

3.3 Genetic Algorithm for Feature Selection

The Genetic Algorithm optimizes the feature selection process by:

- Creating an initial population of random feature sets
- Evaluating each set based on model accuracy
- Applying crossover and mutation to generate new populations
- Iteratively improving the feature set

This approach significantly reduced the number of features while improving the model's performance.

---

4. Implementation

4.1 Environment

- **Programming Language**: Python 3.x
- **Libraries Used**:
  - NumPy, Pandas for data manipulation
  - Scikit-learn for preprocessing and GA
  - TensorFlow and Kera's for ANN model development
  - Matplotlib and Seaborn for visualization

4.2 Steps Followed

1. **Data Loading**: Loaded the CSV file and explored basic statistics and correlations.
2. **Preprocessing**: Encoded categorical data and scaled numerical features.
3. **Feature Selection**: Applied the Genetic Algorithm to choose optimal features.
4. **Model Building**: Constructed and compiled the ANN model.
5. **Training**: Trained the model on the training set.
6. **Evaluation**: Tested the model on the unseen test data.
7. **Visualization**: Plotted loss and accuracy trends over epochs.

---

5. Results and Metrics

5.1 Performance Without Feature Selection

- **Training Accuracy**: ~91%
- **Validation Accuracy**: ~89%
- **Loss**: ~15%

5.2 Performance With Genetic Algorithm Feature Selection

- **Training Accuracy**: ~94%
- **Validation Accuracy**: ~93%
- **Loss**: ~10%

5.3 Confusion Matrix

The confusion matrix showed:

- High true positive and true negative rates
- Very few false positives and false negatives

5.4 Other Evaluation Metrics (Proposed for Future Work)

- **Precision**
- **Recall**
- **F1-Score**
- **ROC-AUC Curve**

---

6. Conclusion

In this project, an Artificial Neural Network model was successfully developed and optimized using a Genetic Algorithm for the prediction of breast cancer based on clinical data. The integration of GA significantly improved the model's accuracy and reduced its complexity by selecting only the most relevant features.

The final model demonstrated a high level of accuracy (~94%) with low loss (~10%), confirming that deep learning models, when combined with evolutionary feature selection techniques, can serve as powerful diagnostic tools in healthcare.

This approach not only enhances diagnostic accuracy but also contributes to faster decision-making in medical practice.

---

7. Future Scope

While the project yielded excellent results, there are several directions for future improvement:

7.1 Model Enhancement

- Explore deeper neural network architectures
- Experiment with alternative activation functions such as Leaky ReLU or ELU
- Implement batch normalization to stabilize training

7.2 Evaluation Improvements

- Use more robust metrics like ROC-AUC, F1-Score, and Precision-Recall curves

7.3 Feature Expansion

- Add more clinical parameters (e.g., hormone receptor status, family history)
- Integrate imaging data (e.g., mammograms)

7.4 Explainability

- Apply SHAP or LIME to understand which features most influence predictions

7.5 Real-World Deployment

- Develop a user-friendly interface for doctors to input patient data
- Deploy the model as a web app or integrate with hospital information systems

---

8. References

- Dataset: https://www.kaggle.com/datasets?search=breast+cancer
- Google Dataset Search: https://datasetsearch.research.google.com
- TensorFlow Documentation: https://www.tensorflow.org
- Genetic Algorithm Library: https://pypi.org/project/geneticalgorithm/
