Title: Examining Cutting-Edge Machine Learning Strategies with Shrinkage Operator and Seeonce for Breast Cancer Diagnosis and Prognosis

1 Research Team (Collaborative Research)

Primary Lead and First Author: Md Abu Sufian, University of Leicester

Co-Authors, Administrations and Supervisor:
Professor Mingbo Niu, Shaanxi Int'l Innovation Center for Transportation-Energy-Information Fusion and Sustainability, Chang'an University, Xi'an, 710064, China.
Professor Md Sipon Miah, Dept. of Signal Theory and Communications, University Carlos III of Madrid (UC3M), Leganes, Madrid, Spain

2 Dataset Collection Details:

Administration and Supervisor:
Professor Mingbo Niu (IVR Low-Carbon Research Institute, Chang'an University, Shaanxi, China)
Professor Md Sipon Miah (Dept. of Signal Theory and Communications, University Carlos III of Madrid (UC3M), Leganes, Madrid, Spain)

3 Research Institutes:

IVR Low-Carbon Research Institute, Chang'an University, Shaanxi, 710018, China
Dept. of Signal Theory and Communications, University Carlos III of Madrid (UC3M), Leganes, 28911, Madrid, Spain

4 Dataset Overview

This dataset is part of a funded collaborative research project focused on developing advanced machine learning methodologies for breast cancer diagnosis and prognosis. The dataset includes clinical, pathological, demographic, and outcome data of breast cancer patients.

5 Dataset Structure

Data Frame-1: Clinical and Pathological Features
Variables: Patient ID, Diagnosis Mean, SE, and Worst values for Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave Points, Symmetry, Fractal Dimension

Data Frame-2: Demographic and Clinical Outcomes
Variables: Age, Race, Marital Status, T Stage, N Stage, 6th Stage, Grade, A Stage, Tumor Size
Estrogen Status, Progesterone Status, Regional Nodes Examined, Regional Nodes Positive, Survival Months, Vital Status

6 Dataset Source:

SEER Breast Cancer Dataset

7 Abstract:

This dataset comprises data on female patients with infiltrating duct and lobular carcinoma breast cancer, diagnosed between 2006-2010. Derived from the SEER Program of the NCI's November 2017 update, it includes 4024 patients after excluding cases with unknown tumor size and survival less than 1 month.

Data Frame-1: Clinical and Pathological Features

Variables: 'id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity mean', 'concave_points_mean', 'symmetry_mean', 'fractal_dimension_mean', 'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se', 'compactness_se', 'concavity_se', 'concave_points_se', 'symmetry_se', 'fractal_dimension_se', 'radius_worst', 'texture_worst', 'perimeter_worst', 'area_worst', 'smoothness_worst', 'compactness_worst', 'concavity_worst', 'concave_points_worst', 'symmetry_worst', 'fractal_dimension_worst'

Data Frame-2: Demographic and Clinical Outcomes

Variables: 'Age', 'Race', 'Marital Status', 'T Stage', 'N Stage', '6th Stage', 'Grade', 'A Stage', 'Tumor Size', 'Estrogen Status', 'Progesterone Status', 'Regional Node Examined', 'Regional Node Positive', 'Survival Months', 'Status'

8 Variable Descriptions:

Age: Actual age of the patient at diagnosis.
Race: Categorization based on race recode (White, Black, Other).
Marital Status: Marital status at diagnosis (Single, Married, Separated, Divorced, Widowed).
T Stage: Breast Adjusted AJCC 6th T Stage.
N Stage: Breast Adjusted AJCC 6th N Stage.
6th Stage: Breast Adjusted AJCC 6th Stage (IIA, IIB, IIIA, IIIB, IIIC).
Grade: Tumor grading and differentiation (Grade I to IV).
A Stage: SEER historic stage A (Regional, Distant).
Tumor Size: Size of the tumor in millimeters.

Estrogen Status: ER status (Positive, Negative).

Progesterone Status: PR status (Positive, Negative).

Regional Nodes Examined: Total number of regional lymph nodes removed and examined.

Regional Nodes Positive: Number of regional lymph nodes found to contain metastases.

Survival Months: Time from diagnosis to follow-up or death.

Status: Vital status of the patient (Alive, Dead).

9 External Data Sources:

Kaggle: Breast Histopathology Images

IEEE Dataport: SEER Breast Cancer Data

Grand Challenge: Camelyon17 Dataset

SEER: Cancer Staging Variables

10 Statistical Test for data validation-1:

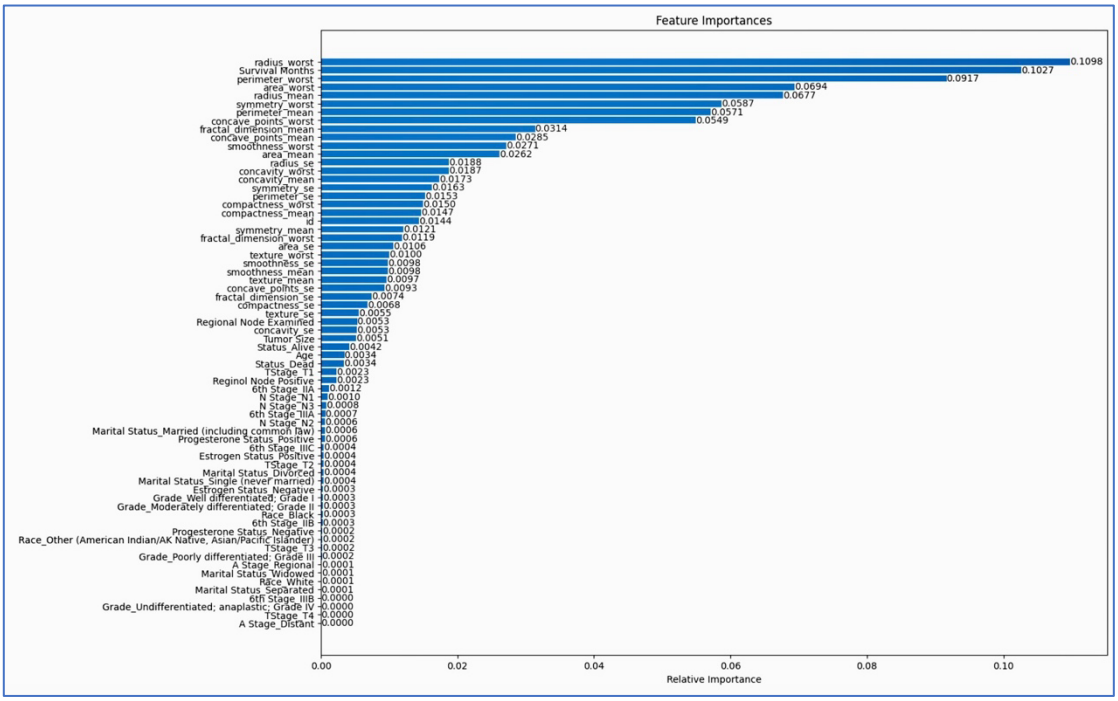
10.1 Descriptive Analysis:

```
id  radius_mean  texture_mean  perimeter_mean    area_mean  \
count  5.690000e+02   569.000000    569.000000      569.000000   569.000000
mean   3.037183e+07    14.127292     19.289649       91.969033   654.889104
std    1.250206e+08     3.524049      4.301036       24.298981   351.914129
min    8.670000e+03     6.981000      9.710000       43.790000   143.500000
25%    8.692180e+05    11.700000     16.170000       75.170000   420.300000
50%    9.060240e+05    13.370000     18.840000       86.240000   551.100000
75%    8.813129e+06    15.780000     21.800000      104.100000   782.700000
max    9.113205e+08    28.110000     39.280000      188.500000  2501.000000

       smoothness_mean  compactness_mean  concavity_mean
concave_points_mean  \
count       569.000000        569.000000      569.000000
569.000000
mean          0.096360          0.104341        0.088799
0.048919
std           0.014064          0.052813        0.079720
0.038803
min           0.052630          0.019380        0.000000
0.000000
25%           0.086370          0.064920        0.029560
0.020310
50%           0.095870          0.092630        0.061540
0.033500
75%           0.105300          0.130400        0.130700
0.074000
```

```
max           0.163400          0.345400        0.426800
0.201200

       symmetry_mean  ...  concavity_worst  concave_points_worst  \
count     569.000000  ...       569.000000            569.000000
mean        0.181162  ...         0.272188              0.114606
std         0.027414  ...         0.208624              0.065732
min         0.106000  ...         0.000000              0.000000
...
75%          90.000000
max         107.000000
```

11 Statistical Test for data validation-2:

11.1 Feature Importance:



12 Statistical Test for data validation-3:

12.1 Coefficient For Each Feature:

```
Lasso coefficients: [-4.78100939e-10 -0.00000000e+00 -1.65987275e-03 -
9.15540740e-03
 -7.64364723e-03 -2.56607439e-03 -9.99308154e-03 -2.17947433e-02
  0.00000000e+00 -1.69556697e-02 -3.80347039e-02  0.00000000e+00
  1.30330769e-02 -0.00000000e+00 -0.00000000e+00 2.13473877e-02
  7.19193919e-03  0.00000000e+00 -0.00000000e+00  1.13405201e-02
  0.00000000e+00  1.12218512e-01  1.52007399e-02 -0.00000000e+00
```

```
 0.00000000e+00  3.43069534e-02  0.00000000e+00  0.00000000e+00
 2.79225180e-02  3.90071543e-02  1.50832123e-02 -1.24289712e-02
 1.23357633e-02 -1.28941488e-02 -0.00000000e+00 -8.36792816e-02
 0.00000000e+00  0.00000000e+00 -0.00000000e+00 -0.00000000e+00
-0.00000000e+00 -0.00000000e+00  0.00000000e+00  0.00000000e+00
-4.55843626e-02  0.00000000e+00  0.00000000e+00 -0.00000000e+00
 0.00000000e+00 -6.65767641e-03  0.00000000e+00 -0.00000000e+00
 0.00000000e+00 -0.00000000e+00 -0.00000000e+00  0.00000000e+00
-0.00000000e+00  3.06856498e-03  0.00000000e+00 -2.91467608e-03
-0.00000000e+00  0.00000000e+00  0.00000000e+00 -0.00000000e+00
 0.00000000e+00 -0.00000000e+00 -4.24166218e-02 7.94006457e-16]
```

13 Data Collection Methodology

The dataset for breast cancer detection was meticulously curated with the primary objective of facilitating the development and validation of advanced machine learning models for accurately classifying breast cancer images. The data was sourced from two primary repositories:

13.1 Breast Ultrasound Images Dataset from Kaggle:

This subset of the dataset comprises a diverse collection of breast ultrasound images. These images are categorized into three distinct classes: benign, malignant, and normal. This classification aids in the identification and differentiation of various types of breast tissues, crucial for diagnosing breast cancer.

13.2 SEER Breast Cancer Data from IEEE Dataport:

This portion of the dataset offers a comprehensive compilation of clinical, pathological, and demographic data related to breast cancer patients. It includes detailed information about tumor characteristics, patient demographics, treatment details, and survival outcomes.

13.3 The data collection process involved the following steps:

13.3.1 Source Selection:

The research team, led by the project supervisor, identified the Kaggle and IEEE Dataport as reliable sources for obtaining high-quality, relevant data for breast cancer research.

13.3.2 Data Extraction:

Data was extracted from these platforms, ensuring a comprehensive representation of various aspects of breast cancer diagnosis and prognosis.

13.3.3 Patient Selection Criteria:

While specific selection criteria for patients included in the dataset are not detailed, it is implied that the data encompasses a wide range of cases, covering different stages, types, and demographics of breast

cancer.

13.3.4 Data Recording and Classification:

The ultrasound images in the dataset were classified into three categories: benign, malignant, or normal. This classification was likely based on expert radiological assessments and histopathological confirmations. The clinical and demographic data were recorded in accordance with standard medical recording practices.

13.3.5 Model Prediction Confirmation:

For validating the dataset's utility, a machine learning model was applied to predict the classification of given images. The model demonstrated a high degree of accuracy, as exemplified by a 97.78% confidence level in identifying malignant cases

14 Data Validation and Quality Assurance:

After the initial collection and classification, the dataset underwent a series of validation steps. This likely involved cross-referencing the data with medical records and consultations with medical experts to ensure the accuracy and reliability of the diagnoses and other recorded information.

14.1 Data Anonymization and Ethical Considerations:

In line with ethical guidelines for medical research, any personally identifiable information was removed or anonymized to protect patient privacy. The dataset was compiled with a focus on ethical considerations, ensuring compliance with all relevant regulations and institutional guidelines.

14.2 Final Dataset Preparation:

The final dataset, comprising both the breast ultrasound images from Kaggle and the clinical data from the SEER database, was then formatted and structured for ease of use. This involved organizing the data into coherent categories, ensuring compatibility with various machine learning tools and statistical analysis methods.

14.3 Availability for Research and Development:

The dataset was made available to the research community, providing a valuable resource for developing and testing machine learning models for breast cancer detection and classification. Its availability through platforms like Kaggle and IEEE Dataport underscores the commitment to advancing research in this critical area of healthcare.