COMPUTER SCIENCE AND DATA ANALYTICS

REPORT 5
COURSE: GUIDED RESEARCH I
PROJECT TITLE : FINDING INFORMATIVE REGIONS IN GRAYSCALE MAMMOGRAM IMAGES.

Arzu Fatullayeva
arzu.fatullayeva@gwmail.gwu.edu

Baku 2023

```
In [9]: print(train_descriptor)

[[237 237  58  99 149 205  70 167 227  32 234   8 179 209  17  48 128 238
   72  90  91 184 218  63 239 253   5 112  74  27 202  10]
 [160 128 106 169 116 195 186  99  94  37  88  19 149 136  50  64 218  43
   19 129  54 186  68  31 204  15 186  97  21   8 117 138]
 [ 41 173  56 123 149 141 119  55 122   9 236   8 178 177   0  49 129  42
   74  24 122 232  80  63 239 223  20 121  80   8 202 170]
 [142  92  22 191 167 226 189 101 156  22  87 179 252 119  43  79  99 115
  187 163  48 213  11 238 228  34 218 131 188 127 165 211]
 [ 40 167  60 113 188 205  81 191 118  61 172  24 179  49   2  16 137   2
   72   8  58 248 115  63 239 223  52 113  64  72 214 106]
 [ 30 199 206 159 174  98  63 236 148  87 223 179  77 103 171  79 127  83
  191 227  48 197  47 224 248  34 255 135 183 255 165 115]
 [ 32 165  61  99 144 237  81  63 118 117  47  24 215  49  74  16 145  10
  104  24  58 120 121  23 205 159  16 116 192 202 198 234]
 [ 44 180 181  97  17 111  99 154 118 251  43  24 215 117  72  60  17  28
   72  56  27 244 249  18 205 222  49 116 192 203  70 240]
 [ 48 181  56  57 187 233  17  39  62 101  42 185 166  61   2  16 135  59
  205  17  54 184  81  23 239 245  16 115 208   0 135  91]
 [195 223 102  31 243 148 190 183 185  81  70 245 104 238 150 247 166 111
  241 205 247 216 158 253  46  58 158 194  15 132 187 221]]
```

Fig. 1 Descriptors for 10 features.

Descriptors obtained from ORB (Fig 1.) used as features for ML models. The descriptor is derived from the rBRIEF (Rotation-aware BRIEF) algorithm and is used to describe the local image content around a detected keypoint.

Threshold for feature was taken 8, and due to this reason 680 pictures(340/340 negative and positive ones) was given to ML models, such as KNN, SVM, Random Forest for training. 250 negative and 250 positive cancer images were given for testing and below in table1 results obtained while testing.

	Accuracy	Recall	Precision	F1 Score
Random Forest	**0.638**	**0.6683**	**0.548**	**0.6022**
SVM	**0.656**	**0.6875**	**0.572**	**0.6245**
KNN	**0.654**	**0.6667**	**0.616**	**0.6403**
Logistic Regression	**0.588**	**0.548**	**0.5957**	**0.5708**

Table 1. Results of ML models for classification of cancer

As this images had interfering text information present in the corners of the mammograms, the keypoints extracted after Brute Force matching includes some of the keypoints from text in this images, which prevents to classify images with high metrics.

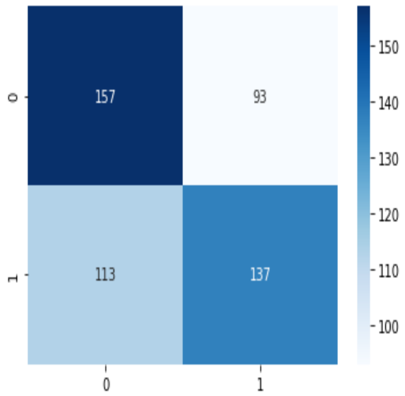

Fig 2. Confusion matrix for Logistic regression

Future Work

The research proposal outlined for further steps focuses on a novel approach to breast cancer detection using augmented ORB descriptors extracted from Digital Imaging and Communications in Medicine (DICOM) mammograms. The key points of approach include:

Augmentation of Mammograms: Extensively augment mammograms to create diverse variations by applying rotations and resizing. The goal is to remove any interfering text information present in the corners of the mammograms.

ORB Descriptor Extraction: The core of this method involves extracting ORB descriptors from the augmented mammograms. These descriptors are keypoints identified within the images, each containing a 32-byte binary string.

Descriptor Selection: Then we plan to select a subset of these descriptors (e.g., 5) that are closely related to the original descriptor. Different closeness criteria such as Hamming distance and Levenshtein distance are explored to identify the most suitable comparison rule. This process aims to identify the most effective and common features.

Classification Model: The selected descriptors are then used to construct a final classification model using advanced machine learning techniques. This model will be evaluated on an independent test dataset to assess its performance.

Promise of Superior Performance: This approach aims to outperform conventional CNN-based architectures by applying innovative augmentation, preprocessing, and feature extraction techniques. The potential success of study could lead to improved breast cancer diagnosis, enabling earlier detection and better treatment outcomes.

Overall, future research proposal demonstrates a thoughtful and comprehensive approach to breast cancer detection, leveraging both image augmentation and advanced feature extraction methods to enhance the accuracy of diagnosis. The use of diverse comparison rules and a final classification model adds depth to the methodology, and the emphasis on rigorous evaluation on an independent dataset ensures the reliability of the results. If successful, then research could indeed have significant implications for breast cancer diagnosis and treatment.