Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777848

WOMEN'S BREAST CANCER PREDICTED USING THE RANDOM FOREST APPROACH AND COMPARISON WITH OTHER METHODS

Amit Bhanushali[1]*, Kulbir Singh[2], Krishnakumar Sivagnanam[3], Kaushik Kumar Patel[4]

[*1]Amit Bhanushali, Independent researcher, Quality Assurance Manager, West Virginia University, WV, USA.

[2]Kulbir Singh, Independent researcher, Health information Manager, IL, USA.

[3]Krishnakumar Sivagnanam, Independent researcher, Solutions Architect, Tech Mahindra (Americas) Inc. VA, USA.

[4]Kaushikkumar Patel, Independent researcher, Director Data Development, White Plains, NY, USA.

[*1]E-mail: amitbhanushali9855@gmail.com

***CORRESPONDING AUTHOR**

*Amit Bhanushali, Independent researcher, Quality Assurance Manager, West Virginia University, WV, USA

Abstract:

The early and precise detection of breast cancer is one of the most crucial measures in the fight against it. Unfortunately, breast cancer is asymptomatic in the early stages, but certain symptoms may appear later on. However, when breast cancer is symptomatic, therapy may be difficult or even impossible, which can result in death. Future technique, info gain method, and random forest method are the three approaches employed. Thus, accurate risk assessment is crucial for lowering mortality. Due to the different risk profiles of women, such as delayed menarche, low drug misuse, and low smoking rates, certain computational algorithms for assessing breast cancer risk have been established in the developed world. However, these strategies do not function well in developing countries. We attempted to demonstrate the superiority of the random forest approach. In this study, we use the Random Forest Classifier (RFC) machine learning approach drinking, dangers at work, and menopausal age. Four strategies — utilizing Chi-Square, common data gain, Spearman relationship, and all elements — were exactly utilized in the component choice. When all risk factors were taken into account. The findings of the selected characteristics for mutual information gain and Chi-Square were identical. The Random Forest Classifier has a fair chance of accurately predicting a woman's risk of developing breast cancer. The study assisted in identifying the female breast cancer risk factors. This is important information that can assist women in focusing on those risk factors in an effort to lower the incidence of breast cancer.

Keywords: random forest, breast, cancer, classifier

1.0. Introduction:

About a million women worldwide are affected by breast cancer each year, and approximately 50% of cases result in death [1–5]. Breast cancer is most frequent in women. According to a

recent epidemiological research, three million new instances of breast cancer would be diagnosed annually worldwide by 2050 [6], indicating that the disease poses a serious threat to human life. If deterioration and mortality are to be avoided, early detection and intervention of breast cancer are best done when it is asymptomatic [7]. Numerous risk factors, some of which may be changed (behavioural risk factors) and others of which cannot (biological risk factors), make women more susceptible to breast cancer [8[While the behavioural risk factors can be reduced or eliminated by changing one's lifestyle. whereas examples of behavioural risk factors include smoking frequency, alcohol use, and workplace hazards. [9,10]. To compute the gamble of measurable model has been demonstrated to be a useful and reliable device. For ladies with an extensive family background of bosom malignant growth and comparative illnesses, the Gail model has been found to essentially downplay the gamble of creating bosom disease [1, 11-13]. This has confined the model's application to specific sorts of ladies. Concentrating on bosom disease risk factors in ladies, particularly Nigerian ladies, was our primary objective. Given the eccentricities of ladies, this is significant. Women have certain unique characteristics, including differences menarche ages. [14,15,16,17].

Machine learning models are a few of the computational techniques that have previously been presented [19-26]. None of these models were created for women, and they also have a low level of predictability. Additionally, several of the computer models that are now being used [27,28 ,29–33]. A set of machine learning classification algorithms known as the (RFC) is comprised of a few individual choice trees that cooperate as a gathering [34].

Each tree in the RF makes a forecast and casts a vote; the class with the largest number of votes decides the expectation for the entire model. One advantage [35–42]. Due to RFC's good performance [43–45] and the fact that it outperformed other machine learning algorithms, it has been suggested for the categorization of breast cancer in American women.

The pre-processing technique will be incorporated into the modelling process and used to Nigerian women in order to improve the model's performance. Because RFC mixes many AI techniques thus utilizes their benefits, we speculated that it would perform better compared to other AI calculations that utilize a solitary model using three methods future method, info gain method and random forest method.

2.0. Literature Review:

Currently, there is a lot of interest in the research of picture categorization utilising deep and machine learning. Mammography images were classified using a variety of methodologies, including binary, multi, and dual classification, to demonstrate the effectiveness of the recommended methods. Recent research (Shen, Wu, & Suk, 2017) show that profound learning improves profound organization preparing by haphazardly erasing layers from convolutional brain organization (CNN) models.

The Mobile Nets employ a realistic architecture based on depth-wise convolutions to build their deep neural networks (Howard et al., 2017). It was recommended to utilize ResNet to classify images. (Xie, Girshick, Dollár, Tu, & He, 2017) recommended VGG, Xception, or ResNet for breast cancer classification due to their excellent training accuracy (around 98 percent). For the game plan of chest threatening development, it has been proposed (Li, Shen, Zhou, Wang, and Li, 2020) that histological pictures be used connected with DenseNet and

SENet. Results from five-overlay cross-endorsement uncovered an AUC of 0.9468, responsiveness of 0.886, and identity of 0.876 for the overhauled InceptionV3 configuration portrayed in (Wang et al., 2020). The Multiscale Assessment of the MIAS dataset The All Convolutional Cerebrum Association (Mom CNN) made by Shin et al. (2016) got 0.99 locale under the twist (AUC) and 96.2% mindfulness. (Zhu et al., 2019) suggest using a Smash Excitation-Pruning (SEP) block in a hybrid CNN designing to bunch chest sickness from histopathology pictures. We looked through the writing and found that mammography examination utilizing the DenseNet121+ELM model has not been finished.

A study team in 2016 (Fabio A Spanhol, Oliveira, Petitjean, & Heutte, 2015) exhibited a precision of around 85.1% at the patient level utilizing SVM and PFTAS highlights. Scientists recognized cores in a recent report utilizing a dataset of 500 examples from 50 patients utilizing a wide scope of techniques, including fluffy C-implies, K-implies, serious learning brain organizations, and Gaussian blend models. Only benign and malignant diagnoses were taken into account in order to ensure accurate reporting. F-100% were attained (Kowal, Filipczuk, Obuchowicz, Korbicz, & Monczak, 2013). Scientists in 2013 showed 94% ID precision on a dataset of 92 examples for bosom malignant growth discovery utilizing a brain organization (NN) and support vector machine (SVM) based procedure with dismissal choice. They then, at that point, assessed 361 examples from the Israel Establishment of Innovation dataset and tracked down around 97% characterization precision (Zhang, Zhang, et al. The complexity and limitations of several publicly available annotated datasets were recently thoroughly discussed in a paper on histological image processing for breast cancer detection and classification (Veta, Pluim, Van Diest, & Viergever, 2014). using the gave engineering tone and surface elements and various classifiers using a democratic component, the typical acknowledgment rate for BC grouping at the patient not entirely set in stone to be 87.53%. This outcome was arrived at utilizing support vector machines, choice trees, nearest neighbor classifiers, discriminant investigation, and gathering classifiers. This approach fared better compared to all others in light of AI up till 2017 (Gupta & Bhavsar, 2017).

Several papers that examine DL methods for breast cancer diagnosis rely on classification utilising CNN variants. The BreakHis dataset is essential for many of these experiments. A convolutional brain organization (CNN) was proposed in 2016 as an amplification free method for distinguishing bosom malignant growth utilizing convolution parts of differed sizes (7, 7, 5 5, and 3 3).Convolutional neural network (CNN) and multi-task CNN (MTCNN) models were used to classify breast cancer patients with an 83.25% identification rate (Bayramoglu, Kannala, & Heikkilä, 2016). In a further research from the same year, breast cancer images and patients were categorised using an AlexNet-like model and several fusion techniques (sum, product, and max). According to Fabio Alexandre Spanhol, Oliveira, Petitjean, and Heutte (2016), the max fusion method allowed researchers to further develop acknowledgment exactness to a normal of 90% for picture grouping and 85.6% for patient characterization. This year witnessed the release of yet another method based on deep learning. In this study, feature vectors were gathered using a convolutional neural network (CNN) and input into a classifier. Fabio A. Spanhol's DeCAF Another challenging dataset, H&E-stained images from breast biopsies, was categorised in 2017 using the CNN model. The four labels for the images were normal tissue, benign lesion, in situ malignancy, and aggressive cancer. Images included both

"invasive" and "in situ" cancer. For more details, go to (Arajo et al., 2017), which includes the findings of the patch-based and image-based evaluations. The binary class trial on BC using images from histopathology, categories breast cancer. The BreakHis dataset has been subjected to a number of machine learning and deep learning techniques, but our new DL architecture has shown to be the most successful. The best findings were obtained by Hayani, & Algamal (2017), although our method yields cutting-edge results based on methods utilised to identify breast cancer in 2017). By utilising a cutting-edge deep learning model on the Break dataset, our research tackles the issue of BC categorization. the datasets from the 2015 Breast Cancer Classification Challenge and his. The model in question is the Inception Recurrent Residual Convolutional Neural Network (IRRCNN). A novel hybrid Extreme Learning Machine Model (ELM) based on DenseNet121 is presented in which breast cancer is detected from mammography pictures. The mammograms underwent preprocessing and data augmentation to improve their quality. In the subsequent step of categorization, attributes were independently collected after initially being pooled and flattened. The completely con-ELM model for attributes.

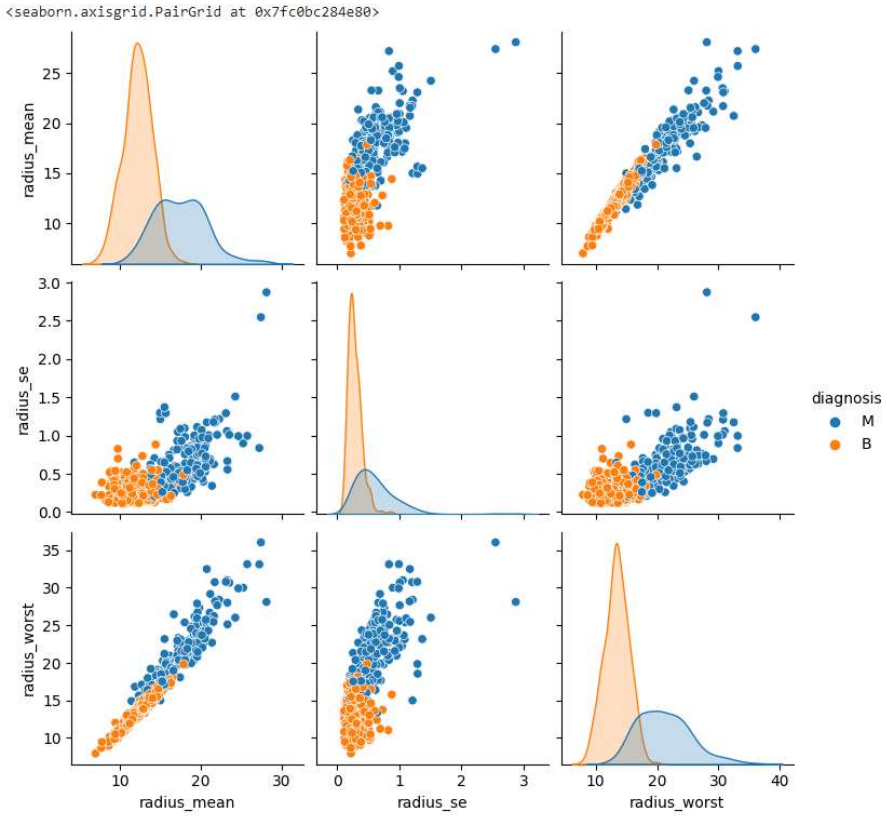
We applied a machine learning method to replace the fully connected layer. The weights of the model employed in the extreme learning machine were optimised using AdaGrad to make it more reliable and effective. AdaGrad was chosen as the preferred optimisation methodology because, in comparison to other techniques, it converges rapidly. The Convolutional Brain Organization (IRRCNN) model is utilized to represent these ideas. The IRRCNN is a strong profound convolutional brain network that joins the upsides of the Initiation Organization (Commencement v4), the Remaining Organization (ResNet), and the Intermittent Convolutional Brain Organization (RCNN).

3.0. Methodology:

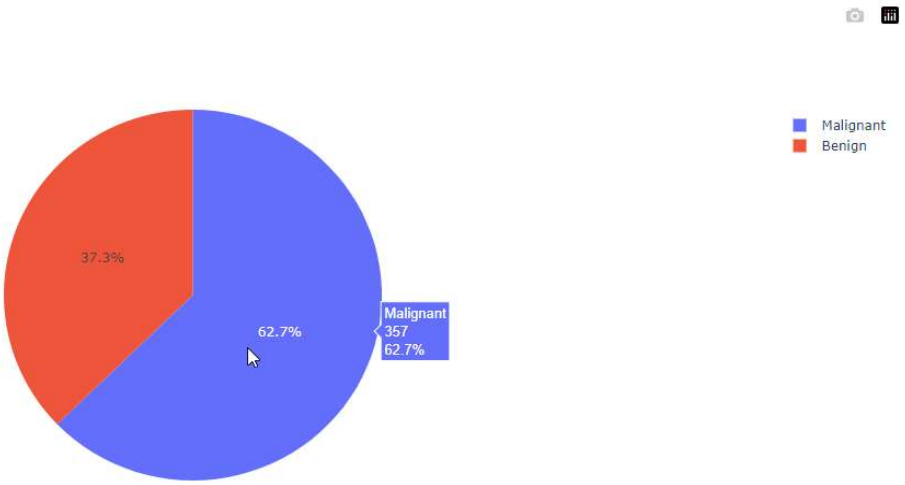
Using the training dataset, data was preprocessed by computationally and objectively choosing a subset of features from all risk categories. These features were then utilised to build classification models. Three methods were used to choose the features. The initial method involved calculating the correlation coefficient between all the risk variables and the subject's actual diagnosis, which was determined to be either verified malignant or benign. The risk variables that had a strong link with the actual data were chosen and classified. Since all of the risk factors are categorical categories, Spearman correlation was applied. Three feature selection algorithms were applied in the second strategy. It is generally known that these feature selection techniques are good at spotting characteristics that could perform well in classification experiments [46–49]. The 25 risk variables included in this study were all employed as the third strategy using three methods future method, info gain method and random forest method. By employing these strategies, we may find the most effective feature selection technique and produce a robust model with a high predictive value for breast cancer. For the training and testing sets, the dataset was divided 70 to 30 accordingly. This suggests that we divided the 90 malignant into 63 for training and 27 for testing at random. It was done again for benign. Therefore, 126 data in total were used for training and 54 for testing. The instruction using three methods future method, info gain method and random forest method.

4.0. Result and Discussion:

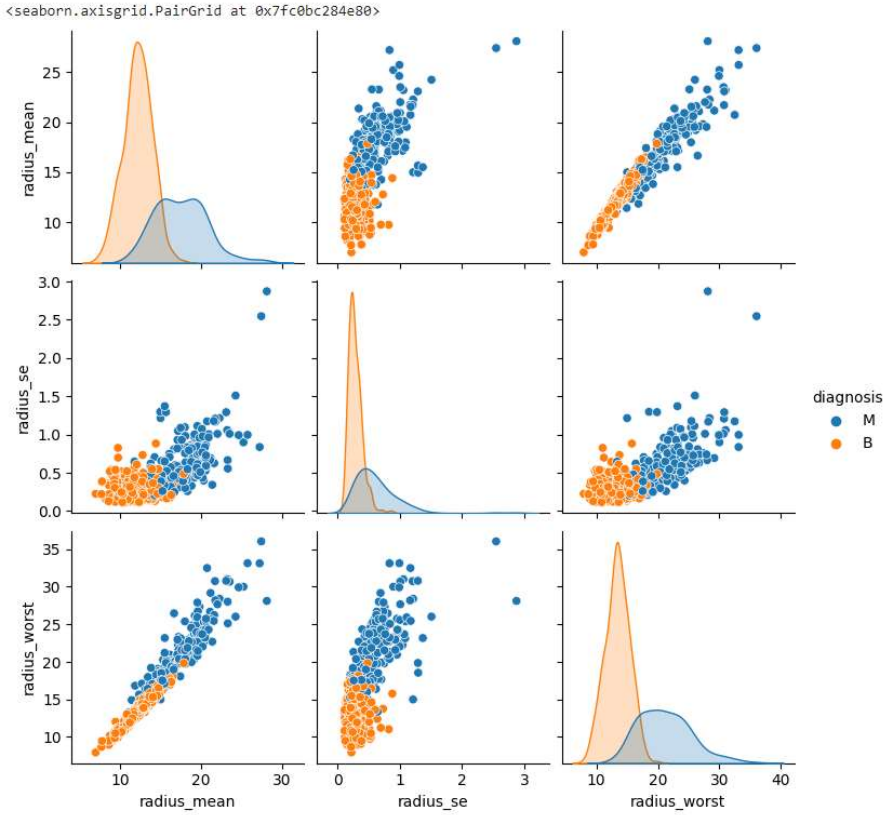
It demonstrates the output of the classification method for the RFC and SVM models utilising the three sets of features. The results demonstrate that when increased to 97%. Chi-Square features with 98.33% accuracy, 100% sensitivity, 96.55% specificity, and 98% AUC provided the greatest performance.



[21], Additionally, we contrasted our model with the well-liked and extensively applied Gail model [50]. 95.00% accuracy, 90.32% sensitivity, and 100% specificity were provided by the Gail model. This demonstrates and even outperforms it since it has higher sensitivity, which means that it is more accurate at predicting the risk of developing cancer in women using three methods future method, info gain method and random forest method.

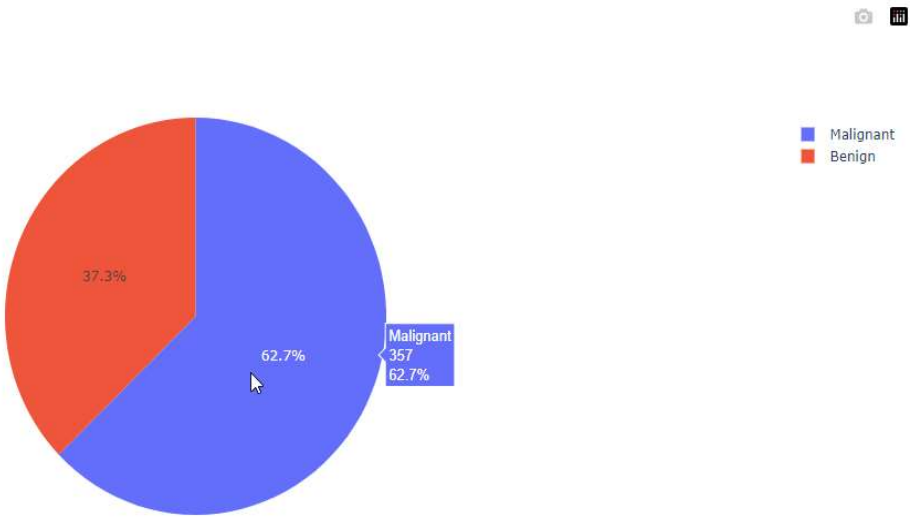


Data visualization (features)

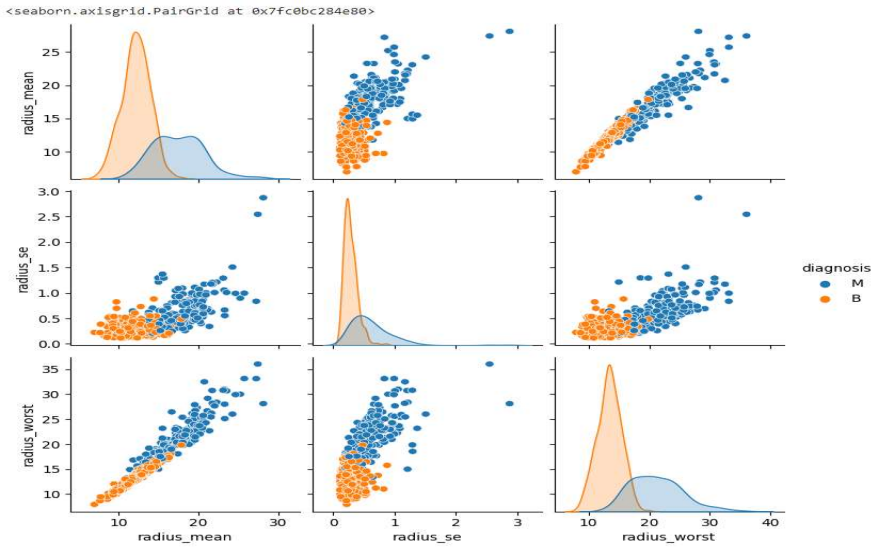


Well-known Gail model. Defects in Gail's model [51]. The Gail model has already been used to Caucasian women without a major breast cancer history. However, our suggested approach can handle predicting breast cancer in black women, whether or not they have a strong history of breast cancer. A rough estimate of 36% of our participants had a history of benign breast illness, and 9% of them had a major. These are the instances where the Gail model was found to overestimate the risk of breast cancer [52] using three methods future method, info gain method and random forest method.

We discovered that five risk factors—exercise, pesticide usage, fruit intake. The two most significant risk variables are also exercise and pesticide usage. This result matches those of previous research conducted in the industrialised world [53–56]. The research that originally identified these risk variables was this one using three methods future method, info gain method and random forest method.



Data visualization (features)



This study has a lot of advantages. The first is the propensity for risk variables to occur. Our accuracy rate for predicting breast cancer using risk factors alone was 98%. This is noteworthy because it suggests that breast cancer may be accurately predicted using the proposed model without the need of imaging or any laboratory tests. Because women have less access to high-quality healthcare facilities, this is good news for them. The lack of good diagnostic hospitals is a result of funding restrictions. Being able to predict breast cancer using just risk factors

suggests that the disease can be detected even before symptoms appear, leading to early treatment and subsequently lower mortality.

The usage of the ensemble machine learning method RFC is the second major strength of this study. Because they blend many machine learning algorithms and use their individual capabilities. Additionally, RFC provided a specificity of about 97%, demonstrating that 97 out of 100 patients without breast cancer were properly identified as such. This holds great promise using three methods future method, info gain method and random forest method.

```
df_model_eval = pd.DataFrame({'model': model_names, 'feature_count': feature_counts,
                              'acc': model_acc_scores,
                              'prc': model_prc_scores, 'rec': model_rec_scores,
                              'f1': model_f1_scores})
df_model_eval
```

	model	feature_count	acc	prc	rec	f1
0	rfc_model_1	30	0.935673	0.935673	0.936040	0.935952
1	dtc_model_1	30	0.906433	0.906433	0.912675	0.907368
2	nbc_model_1	30	0.912281	0.912281	0.907929	0.912415
3	svc_model_1	30	0.923977	0.923977	0.901577	0.922261
4	knc_model_1	30	0.935673	0.935673	0.929761	0.935569

After Applying Feature Selection (look row no 5 & 6) classifier results.

```
df_model_eval = pd.DataFrame({'model': model_names, 'feature_count': feature_counts,
                              'acc': model_acc_scores,
                              'prc': model_prc_scores, 'rec': model_rec_scores,
                              'f1': model_f1_scores})
df_model_eval
```

	model	feature_count	acc	prc	rec	f1
0	rfc_model_1	30	0.935673	0.935673	0.936040	0.935952
1	dtc_model_1	30	0.906433	0.906433	0.912675	0.907368
2	nbc_model_1	30	0.912281	0.912281	0.907929	0.912415
3	svc_model_1	30	0.923977	0.923977	0.901577	0.922261
4	knc_model_1	30	0.935673	0.935673	0.929761	0.935569
5	rfc_model_1a	15	0.959064	0.959064	0.961011	0.959242
6	dtc_model_1a	15	0.953216	0.953216	0.950058	0.953216

Seq model with Adam optimizer

Seq model

+ Code + Text

```
[91] classifier_2 = Sequential()
     classifier_2.add((LSTM(units = 50, return_sequences = True, input_shape = (X_train2.shape[1], 1))))
     classifier_2.add(Dropout(0.3))

     classifier_2.add(LSTM(units = 50, return_sequences = True))
     classifier_2.add(Dropout(0.1))

     classifier_2.add(LSTM(units = 50, return_sequences = True ))
     classifier_2.add(Dropout(0.1))

     classifier_2.add(LSTM(units = 50 ))
     classifier_2.add(Dropout(0.2))

     classifier_2.add(Dense(units = 1  ))
     #classifier_2.add(Dropout(0.1))
     classifier_2.summary()
```

Results of Seq model with Adam optimizer
Precision: 0.703125
Recall: 1.000000
F1 score: 0.825688
==========
Seq model 3: with Adam Optimizer Accuracy: 0.889

****
Hyper Tuning – GridSearch CV

```
cv_scoring = ['balanced_accuracy','f1', 'roc_auc', 'recall','precision']
rfc_model_A = RFC(random_state=0,max_features='sqrt',n_estimators=50,class_weight='balanced')
cv_results_rfc_model_A = CV(rfc_model_A ,XXX,yyy, cv=5, n_jobs=5 , verbose=10 , scoring=cv_scoring)
#display(cv_results_rfc_model_A)

rfc_model_A_auc = cv_results_rfc_model_A['test_roc_auc'].mean()
rfc_model_A_f1  = cv_results_rfc_model_A['test_f1'].mean()
rfc_model_A_rec = cv_results_rfc_model_A['test_recall'].mean()
rfc_model_A_prec = cv_results_rfc_model_A['test_precision'].mean()

print('\t\tModel Training Evaluation (Cross-Validation) Results for rfc_model_X_1: ')
print(' \n\t\t\tauc: {}  \n\t\t\tf1: {}    \n\t\t\trec: {}     \n\t\t\tprec: {}'
          .format( rfc_model_A_auc,rfc_model_A_f1,rfc_model_A_rec,rfc_model_A_prec))

model_names_A.append('rfc_model_X_A')
feature_counts_A.append(XXX.shape[1])
model_auc_scores_A.append(rfc_model_A_auc)
model_f1_scores_A.append(rfc_model_A_f1)
model_rec_scores_A.append(rfc_model_A_rec)
model_prec_scores_A.append(rfc_model_A_prec)
```

```
[Parallel(n_jobs=5)]: Using backend LokyBackend with 5 concurrent workers.
                Model Training Evaluation (Cross-Validation) Results for rfc_model_X_1:

                        auc: 0.9912804778715358
                        f1: 0.9527591209655277
                        rec: 0.9485049833887043
                        prec: 0.9580021141649049
```

```
sdc_model_1 = SDC(random_state=0,class_weight='balanced',max_iter=1000,loss='log')
cv_results_sdc_model_1 = CV(sdc_model_1 ,XXX,yyy, cv=5, n_jobs=5 , verbose=10 , scoring=cv_scoring)
#display(cv_results_sdc_model_1)
sdc_model_1_auc = cv_results_sdc_model_1['test_roc_auc'].mean()
sdc_model_1_f1  = cv_results_sdc_model_1['test_f1'].mean()
sdc_model_1_rec = cv_results_sdc_model_1['test_recall'].mean()
sdc_model_1_prec = cv_results_sdc_model_1['test_precision'].mean()

print('\t\tModel Training Evaluation (Cross-Validation) Results for sdc_model_1: ')
print('\n\t\t\tauc: {}  \n\t\t\tf1: {}   \n\t\t\trec: {}     \n\t\t\tprec: {}'
      .format( sdc_model_1_auc,sdc_model_1_f1,sdc_model_1_rec,sdc_model_1_prec))
model_names_A.append('sdc_model_1')
feature_counts_A.append(XXX.shape[1])
model_auc_scores_A.append(sdc_model_1_auc)
model_f1_scores_A.append(sdc_model_1_f1)
model_rec_scores_A.append(sdc_model_1_rec)
model_prec_scores_A.append(sdc_model_1_prec)
```

```
                Model Training Evaluation (Cross-Validation) Results for sdc_model_1:

                        auc: 0.9637634159825448
                        f1: 0.8214042731462087
                        rec: 0.8264673311184939
                        prec: 0.8607165989518931
```

```
'LogisticRegression'
lrc_model_X = LRC(random_state=0, max_iter=1000, multi_class='ovr', class_weight='balanced',
cv_results_lrc_model_X = CV(lrc_model_X, XXX,yyy, cv=5, n_jobs=5 , verbose=10 , scoring=cv_sc
lrc_model_X_auc = cv_results_lrc_model_X['test_roc_auc'].mean()
lrc_model_X_f1 = cv_results_lrc_model_X['test_f1'].mean()
lrc_model_X_rec = cv_results_lrc_model_X['test_recall'].mean()
lrc_model_X_prec = cv_results_lrc_model_X['test_precision'].mean()
print('\t\tModel Training Evaluation (Cross-Validation) Results for lrc_model_X: ')
print('\n\t\t\tauc: {}  \n\t\t\tf1: {}   \n\t\t\trec: {}  \n\t\t\tprec: {}'.format( lrc_model

model_names_A.append('lrc_model_X')
feature_counts_A.append(XXX.shape[1])
model_auc_scores_A.append(lrc_model_X_auc)
model_f1_scores_A.append(lrc_model_X_f1)
model_rec_scores_A.append(lrc_model_X_rec)
model_prec_scores_A.append(lrc_model_X_prec)
```

```
[Parallel(n_jobs=5)]: Using backend LokyBackend with 5 concurrent workers.
                Model Training Evaluation (Cross-Validation) Results for lrc_model_X:

                        auc: 0.9673939938133538
                        f1: 0.8885423946487336
                        rec: 0.863344407530454
                        prec: 0.9173616435244343
```

```
dtc_model_X = DTC(random_state=0, class_weight='balanced', max_features='sqrt')
cv_results_dtc_model_X = CV(dtc_model_X, XXX,yyy, cv=5, n_jobs=5 , verbose=10 , scoring=cv_scorin
dtc_model_X_auc = cv_results_dtc_model_X['test_roc_auc'].mean()
dtc_model_X_f1 = cv_results_dtc_model_X['test_f1'].mean()
dtc_model_X_rec = cv_results_dtc_model_X['test_recall'].mean()
dtc_model_X_prec = cv_results_dtc_model_X['test_precision'].mean()

print('\t\tModel Training Evaluation (Cross-Validation) Results for dtc_model_X: ')
print('\n\t\t\tauc: {}  \n\t\t\tf1: {}   \n\t\t\trec: {}   \n\t\t\tprec: {}'.format( dtc_model_X_

model_names_A.append('dtc_model_X')
feature_counts_A.append(XXX.shape[1])
model_auc_scores_A.append(dtc_model_X_auc)
model_f1_scores_A.append(dtc_model_X_f1)
model_rec_scores_A.append(dtc_model_X_rec)
model_prec_scores_A.append(dtc_model_X_prec)
```

```
                Model Training Evaluation (Cross-Validation) Results for dtc_model_X:

                        auc: 0.930031610853753
                        f1: 0.9104122136447017
                        rec: 0.9246954595791805
                        prec: 0.8983415186094661
```

Precision: 0.906250
Recall: 0.865672
F1 score: 0.885496
==========
'Seq model 4 : with Adam Optimizer Accuracy: 0.912

This study's inclusion of a data pre-processing phase in the modelling is its third main strength. Pre-processing is known to enhance computational models [29–33]. Feature extraction represented the main preprocessing phase. It is for the most part realized that include determination models can improve the characterization execution of AI calculations [46–49]. Three methods were utilised for the feature. This demonstrates how adding a pre-processing phase to our modelling approach enhanced the effectiveness of the suggested model. The study's participants' singularity is its fourth key strength. We looked at females. This demonstrates how distinctive our study is. This study is especially exceptional since it identified 11 risk variables that are reliable indicators of female breast cancer. There were 25 risk variables examined, however only 11 were shown to be highly predictive. The sample size is the study's main drawback using three methods future method, info gain method and random forest method. The fact that we only looked at 180 people suggests that the results might not be easily generalizable. Data availability, a frequent characteristic of, placed constraints on us. Future research should think about enlarging the sample size to improve the generalizability of the findings using three methods future method, info gain method and random forest method.

5.0. Conclusion

Our goal in studying a group of only based on risk indicators using three methods future method, info gain method and random forest method. For the effective prediction of female breast cancer, we created a Random Forest Classifier. Our findings demonstrate that risk variables are only highly predictive of breast cancer even before any symptoms appear. Additionally, our findings support the hypothesis that using three methods future method, info gain method and random forest method. The main causes of breast cancer in women were also found. This is important information that can assist women in focusing on those risk factors in an effort to lower the incidence of breast cancer. In comparison to the info gain approach and the future method, we have discovered that the random forest method produces better outcomes. The random forest approach is the most effective treatment for breast cancer.

6.0. Acknowledgement:

I am grateful to my guide and to God for providing me with such a significant chance. I also want to thank everyone who has supported me in my work, whether directly or indirectly.

7.0. References:

[1] S. Gupta, D. Kumar, A. Sharma, Data mining classification techniques applied for breast cancer diagnosis and prognosis, Indian J Comput. Sci. Eng. 2 (2) (2011) 188–195.
[2] R.A. Weinberg, R.A. Weinberg, The Biology of Cancer, Garland Science, 2013.
[3] E. Barillot, L. Calzone, P. Hup´e, J. Vert, A. Zinovyev, Computational systems biology of cancer, Biomed Eng Online 12 (76) (2012) 1–3.
[4] L. Pecorino, Molecular Biology of Cancer: Mechanisms, Targets, and Therapeutics, Oxford University Press, 2012.
[5] F. Bray, et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin 68 (6) (2018) 394–424.

[6] L.C. Seeff, et al., Patterns and predictors of colorectal cancer test use in the adult US population, Cancer: Interdiscip Int.l J. Am. Cancer Soc. 100 (10) (2004) 2093–2103.
[7] Z. Tao, et al., Breast cancer: epidemiology and etiology, Cell Biochem. Biophys. 72 (2) (2015) 333–338.
[8] K. Williams, et al., Breast cancer risk prediction using data mining classification techniques, Trans. Netw. Commun. 3 (2) (2015), 01-01.
[9] J. Dheeba, N.A. Singh, S.T. Selvi, Computer-aided detection of breast cancer on mammograms: a swarm intelligence optimized wavelet neural network approach, J Biomed Inform 49 (2014) 45–52.
[10] V. Chaurasia, S. Pal, A novel approach for breast cancer detection using data mining techniques, Int. J. Innovative Res. Comput. Commun. Eng. 2017 (2) (2014) 2456–2465.
[11] A. Hüsing, et al., Validation of two US breast cancer risk prediction models in German women, Cancer Causes & Control 31 (6) (2020) 525–536.
[12] J.J. Pitt, et al., Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features, Nat Commun 9 (1) (2018) 1–12.
[13] Kumar, N., V. Singh, and G. Mehta, Assessment of Common Risk Factors and Validation of the Gail Model for Breast Cancer: A hospital-Based Study from Western India.2020.
[14] S. Wang, et al., Development of a breast cancer risk prediction model for women in Nigeria, Cancer Epidemiol. Prev. Biomarkers 27 (6) (2018) 636–643.
[15] L.A. Brinton, et al., Breast cancer in Sub-Saharan Africa: opportunities for prevention, Breast Cancer Res. Treat. 144 (3) (2014) 467–478.
[16] D.A. Boggs, et al., Validation of a breast cancer risk prediction model developed for Black women, J. Natl. Cancer Inst. 105 (5) (2013) 361–367.
[17] C.A. Adebamowo, O. Ajayi, Breast cancer in Nigeria, West Afr J Med 19 (3) (2000) 179–191.
[18] E. Alpaydin, Introduction to Machine Learning, MIT press, 2020.
[19] G. Lee, H. Fujita, Deep Learning in Medical Image analysis: Challenges and Applications, Springer, 2020.
[20] D.B. Carsten Henneges, Richard Fux, Natascha Friese, Harald Seeger,
[21] Hans Neubauer, Stefan Laufer, Christoph H Gleiter, Matthias Schwab, Andreas Zell, Bernd Kammerer, Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection, Biomed. Centre 9 (104) (2009) 1–11.
[22] S.A. Mojarad, S.S. Dlay, W.L. Woo, G.V. Sherbet, Breast Cancer Prediction and Cross Validation Using Multilayer Perceptron Neural Networks, IEEE, 2010.
[23] I. Saritas, Prediction of breast cancer using artificial neural networks, J. Med Syst. 36 (2012) 2901–2907.
[24] A.H. Al-Timemy, F.M. Al-Naima, N.H. Qaeeb, Probabilistic Neural Network for Breast Biopsy Classification, IEEE Computer Society, 2009, pp. 101–106.
[25] T.T. Anothaisintawee, Y. Wiratkapun, V. C., Kasamesup, A. Thakkinstian, Risk prediction models of breast cancer: a systematic review of model performances, Breast Cancer Res. Treat. 133 (2012) 1–10.
[26] N. Cruz-Ramíreza, H.G. Acosta-Mesaa, H. Carrillo-Calvetb, L.A. Nava-Fern´andezc,

R.E. Barrientos-Martíneza, Diagnosis of breast cancer using Bayesian networks:acase study, Comput. Biol. Med. 37 (2007) 1553–1564.
[27] Y.U. Ryu, R. Chandrasekaran, V.S. Jacob, Breast cancer prediction using the isotonic separation technique, Eur J Oper Res 181 (2007) 842–854.
[28] T.F. Adam Krzy˙zak, M.ehdi Habibzadeh, Ł.ukasz Jele´n, Application of pattern recognition techniques for the analysis of histopathological images, Comput Recognit Syst. 4 (2011) 623–644.
[29] H. Nasser, A.A. Sweilam a, N.K. Tharwat b, Abdel Moniem c, Support vector machine for diagnosis cancer disease: a comparative study, Egypt Inform. J. 11 (2) (2010) 81–92.
[30] D. Lavanya, D.K.U. Rani, Analysis of feature selection with classification: breast cancer datasets, Indian J. Comput. Sci. Eng. 2 (5) (2011) 756–763.
[31] G.I. Salama, M. Abdelhalim, M.A.-e. Zeid, Breast cancer diagnosis on three different datasets using multi-classifiers, Breast Cancer (WDBC) 32 (569) (2012) 2.
[32] D. Lavanya, K.U. Rani, Ensemble decision tree classifier for breast cancer data, Int.
J. Inform. Technol. Convergence Serv. 2 (1) (2012) 17.
[33] Kharya, S., Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease.arXiv preprint arXiv:1205.1923, 2012.
[34] P. Chandrasekar, et al., Improving the prediction accuracy of decision tree mining with data preprocessing, in: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC, 2017. IEEE.
[35] L.I. Kuncheva, et al., Multi-modal biometric emotion recognition using classifier ensembles, in: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2011. Springer.
[36] R. Díaz-Uriarte, S.A. De Andres, Gene selection and classification of microarray data using random forest, BMC Bioinformatics 7 (1) (2006) 3.
[37] M. Khalilia, S. Chakraborty, M. Popescu, Predicting disease risks from highly imbalanced data using random forest, . BMC Med. Inform. Decis. Making 11 (1) (2011) 51.
[38]T.M. Khoshgoftaar, M. Golawala, J. Van Hulse, An empirical study of learning from imbalanced data using random forest, in: 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), 2007. IEEE.
[39] J.S. Evans, et al., Modeling Species Distribution and Change Using Random Forest, inPredictive species and Habitat Modeling in Landscape Ecology, Springer, 2011, pp. 139–159.
[40] Z Masetic, A. Subasi, Congestive heart failure detection using random forest classifier, Comput Methods Programs Biomed 130 (2016) 54–64.
[41] M. Mohammady, H.R. Pourghasemi, M. Amiri, Land subsidence susceptibility assessment using random forest machine learning algorithm, Environ Earth Sci 78 (16) (2019) 503.
[42] A. Subasi, E. Alickovic, J. Kevric, Diagnosis of chronic kidney disease by using random forest. CMBEBIH 2017, Springer, 2017, pp. 589–594.
[43] Mariana Belgiu a, Lucian Dra˘gut b, Random forest in remote sensing: a review of applications and future directions, ISPRS J. Photogrammetry Remote Sensing (114) (2016) 24–31.

[44] J. Thongkam, G. Xu, Y. Zhang, AdaBoost algorithm with random forests for predicting breast cancer survivability, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence, 2008. IEEE.
[45] D.J. Wu, et al., Privately evaluating decision trees and random forests, in: Proceedings on Privacy Enhancing Technologies 2016, 2016, pp. 335–355.
[46] J. Ali, et al., Random forests and decision trees, Int. J. Comput. Sci Issues 9 (5) (2012) 272.
[47] S. Bahassine, et al., Feature selection using an improved Chi-square for Arabic text classification, J. King Saud University-Comput Inform Sci. 32 (2) (2020) 225–231.
[48] S. Adi, Y. Pristyanto, A. Sunyoto, The best features selection method and relevance variable for web phishing classification, in: 2019 International Conference on Information and Communications Technology (ICOIACT), 2019. IEEE.
[48] N. Barraza, et al., Mutual information and sensitivity analysis for feature selection in customer targeting: a comparative study, J. Inform. Sci. 45 (1) (2019) 53–67.
[49] N. Bi, et al., A multi-feature selection approach for gender identification of handwriting based on kernel mutual information, Pattern Recognit Lett 121 (2019) 123–132.
[50] H. Abdel-Razeq, L. Zaru, A. Badeeb, S. Hijjawi, The application of gail model to predict the risk of developing breast cancer among jordanian women, J Oncol 2020 (2020) 1–6.
[51] X. Wang, Y. Huang, L. Li, H. Dai, F. Song, K. Chen, Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta- analysis with trial sequential analysis, Breast Cancer Research 20 (18) (2018) 1–19.
[52] S.H Ewaid, L.H. AliAl-Azzawi, Breast cancer risk assessment by Gail Model in women of Baghdad, Alexandria J. Med. (2016) 1–4.
[53] R. Chowdhury, B. Sinha, M.J. Sankar, S. Taneja, N. Bhandari, N. Rollins, R. Bahl, J. Martines, Breastfeeding and maternal health outcomes: a systematic review and meta-analysis, Acta Pædiatrica 104 (2015) 96–113.
[54] D. Akdeniz, M.M. Klaver, Z.A. Smith Chloe´, B. Koppert Linetta, J. Hooning Maartje, The impact of lifestyle and reproductive factors on the risk of a second new primary cancer in the contralateral breast: a systematic review and meta-analysis, Springer J. 35 (5) (2020).
[55] C. He, P. Kraft, I. Chasman Daniel, E. Buring Julie, C. Chen, E. Hankinson Susan, G. Pare, S. Chanock, M. Ridker Paul, J. Hunter David, A large-scale candidate gene association study of age at menarche and age at natural menopause, Springer J 128 (2010) 515–527.
[56]V.S. Blinder, M.M. Murphy, L.T. Vahdat, H.T. Gold, I. de Melo-Martin, M.K. Hayes,R.J. Scheff, E. Chuang, A. Moore, M. Mazumdar, Employment after a breast cancer diagnosis: a qualitative study of ethnically diverse urban women, Springer J. 37 (2012) 763–772.