Research Article
Journal of Research and Review in Science, 33-38
Volume 7, December 2020
DOI: 10.36108/jrrslasu/0202.70.0160

ORIGINAL RESEARCH

Random Forest Classifier for Diagnosis of Breast Cancer in African Women

Babafemi Oluropo MACAULAY[1], Soji Alabi AKANDE[2], Olusola Aanu OLABANJO[1], Boluwaji Adeola AKINNUWESI[1] and Benjamin Segun ARIBISALA

[1]Department of Computer Science, Faculty of Science, Lagos State University, Nigeria

[2]Department of Surgery, Lagos State University Teaching Hospital, Lagos Nigeria

Correspondence
Benjamin Segun Aribisala, Department of Computer Science, Faculty of Science, Lagos State University, Nigeria.
Email:Benjamin.aribisala@lasu.edu.ng

Funding information

Self-Sponsored

Abstract:
Introduction: Early detection of breast cancer among women in Sub-Saharan Africa (SSA) is very challenging because of many factors e.g. low knowledge of breast cancer, lack of awareness of early detection treatment, treatment cost and poor perception of breast cancer. There is dearth of research efforts on computational approach to breast cancer diagnosis in SSA as against what obtains in developed countries.
Aim: Here, we propose a novel diagnosis model for African women using a computational technique called Random Forest (RF), a machine learning technique.
Materials and Methods: Study data comprised of technical indicators (diagnostic parameters) of breast cancer diagnosis, collected from breast cancer patients attending oncology clinic in Lagos State University Teaching Hospital. A total of 180 subjects were studied out of which 90 were confirmed cases of breast cancer and 90 were benign. Nine diagnostic parameters were included. These are clump thickness, marginal adhesion, uniformity of cell size, uniformity of cell shape, single epithelial cell, bare nuclei, bland chromatin, normal nucleoli and mitosis. Principal Component Analysis was used for feature selection and RF model was used for classification.
Results: The RF model gave an accuracy of 98.23%, sensitivity of 95.24%, specificity of 100.00% and Area under curve of 98%.
Conclusion: The proposed Random Forest model has a good potential at classifying breast cancer in African women. Adoption of computational diagnosis approach in SSA can lead to early diagnosis and reduction of mortality rate.

Keywords: Breast Cancer, Random Forest, Machine Learning, African women, Diagnosis, feature extraction.

All co-authors agreed to have their names listed as authors.

JRRS
https:sciencejournal.lasucomputerscience.com

1. INTRODUCTION

Breast cancer is the most diagnosed and is also known to be the highest cause of cancer-related mortality among women globally [1-4]. It is documented that 15% of all female cancer is breast cancer and in 2012, it was reported globally that 521,000 deaths among women was due to breast cancer [5]. Cancer belongs to a class of diseases that result from abnormal cell growth [6]. Diagnosis and treatment of breast cancer in its earliest stage remains the only way to improve its outcome and reduce mortality, thus early and accurate diagnosis of breast cancer is important [7, 8]. To survive breast cancer in the long term, metastasis of the cancerous cell must be halted through appropriate medical intervention at the early stage [7]. Early detection of breast cancer among women in Sub-Saharan Africa (SSA) is very challenging because of lack of awareness, treatment cost [9, 10].

Previous studies have demonstrated that there are some differences in breast cancer etiology in African women compared with those of European women [11-14]. For example, African women tend to have more aggressive breast cancer than the European counterparts [20]. Also, the tumour biology of African women is worse than those of European women [9], cancer management is more difficult to manage in Black women than in white women [10] and African women have poor access to Breast Cancer screening [10].

Additionally, some specific uniqueness have been identified in Nigerian women in relation to breast cancer [14, 15]. First, Nigerian women seems to have an earlier age of breast cancer onset compared with women from other countries [14], [16-18]. Second, the breast cancer tumour size looks bigger in Nigerian women than those of women in other countries [15]. The mortality rate of breast cancer is higher in Nigerian women than women from other countries [16,17]. This uniqueness suggests that increased research efforts toward early diagnosis and reduced mortality rate in Nigerian women is highly desirable.

Manual methods of breast cancer diagnosis are susceptible to human errors leading to between 10 and 30% misdiagnosis by Radiologists [19]. Computational methods have been identified to have better accuracy than manual methods and thus the potential to improve treatment outcomes and diminish mortality rate [4], [20, 21]. Several models have been proposed for breast cancer classification, with the aim of separating breast cancer into malignant and benign. However, most of the models are in use in the developed countries, where the level of education, breast cancer awareness, access to good healthcare facilities and standard of living are considerably better than what obtains in the SSA [19, 22, 23]. Computational research in the SSA are very limited. Additionally, most of the existing models have limited accuracy.

The importance of breast cancer diagnosis has made it to attract many research efforts in the developed nations and a number of computational techniques have been proposed. Some of the previously used techniques include: Artificial Neural Network (ANN) [24]; Genetic Algorithm (GA) [5]; Genetically Optimized Neural Networks Model (GONNM) [3]; Genetic Programming Generated Feature (GPGF) [25]; Support Vector Machine (SVM) [26]; K-Nearest Neighbour (KNN) [27]; Fuzzy-rough Nearest Neighbour (FrNN) [28]; and Ensemble Learning (EL) [2].

Our focus in this study is to develop a computational model for diagnosis of breast cancer among African women using Random Forest Classifier (RFC). RFC belongs to the family of machine learning classification algorithms that comprises of several number of individual decision trees operating as an ensemble [29]. Each tree in the random forest makes a classification and decide, the class with the most selection becomes the overall model’s classification. One of the merits of RFC over a single model is that each of the tree classifier is synonymous to team member and all members work synergistically to make the final classification. This will certainly perform better than when using a single decision tree. Other advantages of RFC are; it is applicable to binary classification, has tendency to cope with a dataset where the number of variables is greater number of observations, not susceptible to the problem of over-fitting, has capability to handle dataset that combines continuous and categorical classifiers, high efficiency and excellent accuracy, sensitivity and specificity [30-32]. RFC has been proposed for classification of breast cancer in American women because of its excellent performance [33] and it was found to outperform other machine learning techniques. To the best of our knowledge, there is no study that has investigated the classification power of RFC on African breast cancer dataset.

The aim of this study was to develop an ensemble machine learning model based on RFC for diagnosis of breast cancer. The model will be implemented on Nigerian women and a feature selection technique based on the Principal Component Analysis will be included in the modelling because of its potential to increase model’s performance [25]. We hypothesised that RFC will perform better than other machine learning algorithms that use a single model because it combines different machine learning algorithms thereby taking advantage of their strengths.

2. MATERIAL AND METHODS / EXPERIMENTAL DETAILS / METHODOLOGY

2.1 Subjects

The data used for this study were from the data of patients receiving treatment at the oncology department of Lagos State University Teaching

Hospital (LASUTH) Ikeja Lagos, Nigeria over a 10-year period covering 2009 to 2019. Included participants were 180 adult women, who had been clinically evaluated for breast cancer. 90 were malignant cases and 90 benign cases. Approval for use of data was granted, upon request, by the Health Research and Ethics Committee of the hospital (REF NO: LREC/06/10/1253). There was no direct contact between the researchers and any of the participants as data extraction was entirely done from their clinical history record. We did not have access to personal information of the participants, thus guaranteeing confidentiality of patients' information. All cancer cases that have no bearing with the breast were excluded from our selection.

2.2 Data Description

The data used in this study are the technical indicators of breast cancer which can be described as features or parameters that are used in clinical setting for diagnosis of breast cancer. Nine technical indicators of breast cancer diagnosis were used, these are: clump thickness, marginal adhesion, uniformity of cell size, uniformity of cell shape, single epithelial cell, bare nuclei, bland chromatin, normal nucleoli and mitosis.

The technical indicators were measured by attending doctors using images from various diagnostic procedures with values ranging from 1 – 10. These values were later standardized by converting to 0.00 – 1.00 to ensure uniformity. These technical indicators are conventionally used by Radiologists for manual identification of breast cancer. During manual identification, clump thickness is rated as benign if the value is between 0.00 – 0.49 and malignant otherwise. For marginal adhesion, 0.00 – 0.49 is classified as benign while 0.50 – 1.00 is classified as malignant. For uniformity of cell size, 0.00 – 0.49 is classified as benign and 0.50 – 1.00 is rated as malignant. For single epithelial cell, 0.00 – 0.49 is classified as benign while 0.50 – 1.00 is classified as malignant. For uniformity of cell shape, 0.00 – 0.49 is rated as benign and malignant otherwise. In the case of bare nuclei, 0.00 – 0.49 is rated and malignant otherwise. For bland chromatin, 0.00 – 0.49 is classified as benign while 0.50 – 1.00 is classified as malignant. For normal nucleoli, 0.00 – 0.49 is rated as benign and malignant otherwise. Lastly, for mitosis, 0.00 – 0.49 is rated as benign and malignant otherwise.

2.3 Data Pre-processing

Data preprocessing was done by computationally and objectively selecting some features from the nine technical indicators. Feature selection was done in order to identify features that have good discriminatory power at separating the benign breast cancer from the malignant ones. The computational approach of feature selection used in this study is called Principal Component Analysis (PCA). This feature selection model is well known to perform well at identifying features that could give high performance during classification experiment [33].

2.4 Classification Using Random Forest Classifier

2.4.1 Model Development and Evaluation

The dataset was divided into two groups, 70% for training while 30% was used for testing. The training set was used to build the RFC model while the testing case was used in the model evaluation. In the modelling, we experimentally tuned the hyperparameters in order to optimize performance. These parameters include the number of decision trees in the forest and the number of features considered by each tree while splitting a node. The hyperparameters of the maximum depth was set to 5 while estimator was set to 10 and the maximum feature value was set to 1. In order to test our hypothesis that RFC performs better than other machine learning algorithm that uses a single model, we also classified the data using SVM. SVM is a machine learning technique that is well known to perform very well on biological data [20]. Standard performance metrics such as accuracy, sensitivity, specificity and AUC were used to assess model performance. The contributions of the technical indicators to the breast cancer diagnosis, that is the features, were computed using feature importance. RFC and SVM were implemented in python programming language.

3. RESULTS AND DISCUSSION

3.1 Technical Indicators

Table 1 shows the distribution of the dataset using the technical indicators and a threshold of 0.5. The threshold was used because that is the threshold commonly used by Radiologists during manual assessment for discriminating between the benign and the malignant group. The result shows that about 70% of the dataset falls into the group of 0 – 0.49 when characterized using Uniformity of cell size, clump thickness, marginal adhesion and bland chromatin, while about 60% of the dataset falls into the group of 0 – 0.49 when characterized using uniformity of cell shape, single epithelia cell, bare nuclei, normal nucleoli and mitosis.

Table 1: Segregation of Participants Based on Values of Technical Indicators

S/N	Technical Indicators	0.00 – 0.49	0.50 – 1.00
1	Uniformity of Cell Size	72.78%	27.22%
2	Uniformity of Cell Shape	59.44%	40.56%
3	Clump Thickness	78.89%	21.11%
4	Marginal Adhesion	77.78%	22.22%
5	Single Epithelial Cell	65.56%	34.44%
6	Bare Nuclei	64.44%	35.56%
7	Bland Chromatin	71.11%	28.89%
8	Normal Nucleoli	66.67%	33.33%
9	Mitosis	66.11%	33.89%

The result of PCA shows that bare nuclei is the most viable technical indicator for breast cancer diagnosis (Table 2)

Table 2: **Communalities Resulting from Principal Components Analysis**

Technical Indicators	Initial	Extraction
Uniformity of Cell Size	1.000	0.597
Uniformity of Cell Shape	1.000	0.667
Clump Thickness	1.000	0.640
Marginal Adhesion	1.000	0.536
Single Epithelial Cell	1.000	0.604
Bare Nuclei	1.000	0.726
Bland Chromatin	1.000	0.693
Normal Nucleoli	1.000	0.678
Mitosis	1.000	0.586

3.2 Feature Extraction

From Table 2, the order of importance of the features with respect to breast cancer diagnosis was ranked as: bare nuclei, bland chromatin, normal nucleoli, uniformity of cell shape, clump thickness, single epithelial cell, uniformity of cell size, mitosis and marginal adhesion. The three most important features were bare nuclei, bland chromatin and normal nucleoli, and they were used in classification to enable a comparison with the use of the entire nine features.

3.3 Results of Classification using Random Forest Classifier

In the first classification experiment where all the nine features were used, the RFC model gave an accuracy of 97.56%, sensitivity of 95.25%, and specificity of 100% and Area under curve (AUC) of 92%. In comparison with SVM, the RFC model out-performed SVM, which had 85.11% accuracy, 84.29% sensitivity, 92.56% specificity and 90% AUC.

In the second classification experiment where only three features (bare nuclei, bland chromatin and normal nucleoli) were used, the RFC model gave an accuracy of 98.23%, sensitivity of 95.24%, and specificity of 100.00% and Area under curve (AUC) of 98%. In comparison with SVM, the RFC model out-performed SVM, which has 96.33% accuracy, 89.67% sensitivity, 92.67% specificity and 94% AUC.

Figure 1 shows the ROC of PCA used for feature extraction when only three features (bare nuclei, bland chromatin and normal nucleoli).

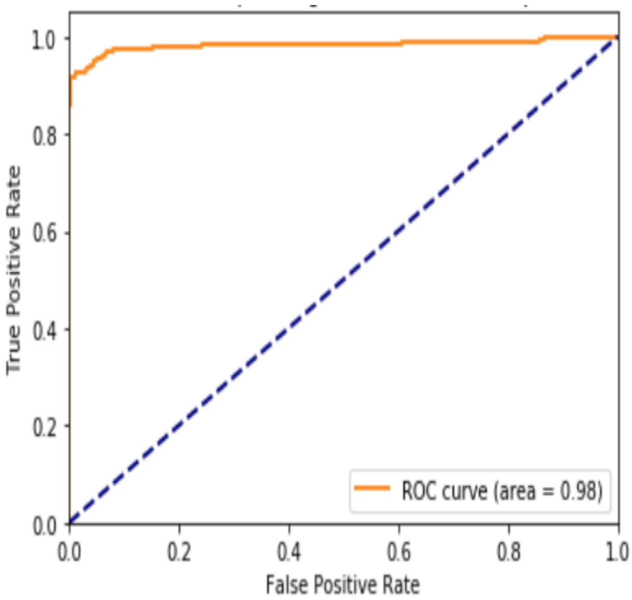

Figure 1: ROC Analysis showing the performance of the proposed Principal Components Analysis based feature extraction Random Forest Classification model

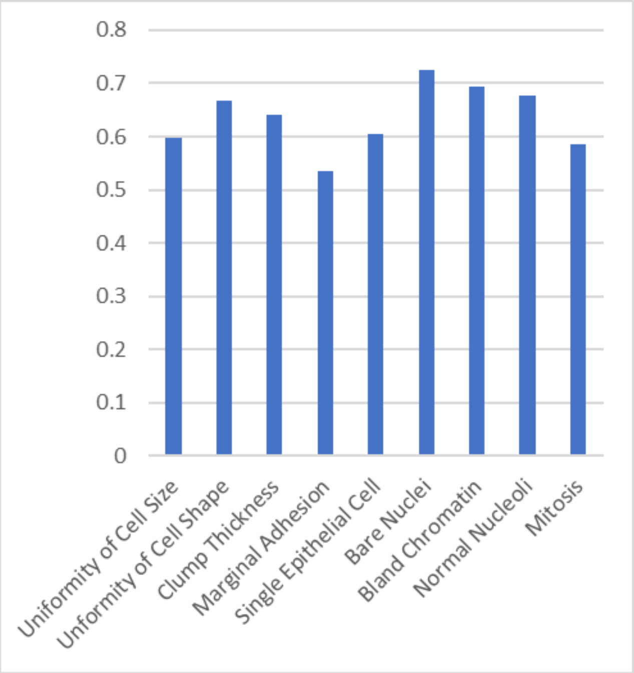

Figure 2: Contributions of the Technical Indicators to Diagnosis using Feature importance

The aim of this study was to develop a Random Forest Classifier for breast cancer diagnosis in African women using standard technical indicators only. We used nine technical factors obtained from 180 participants and our feature extraction was based on Principal Components Analysis. Our results showed that PCA extracted bare nuclei as the strongest technical indicator for breast cancer diagnosis upon which our RFC model obtained an accuracy of 98.23%, sensitivity of 95.24%, and specificity of 100.00% and Area under curve (AUC) of 98%.

This study has many strengths. Firstly, it uses an ensemble machine learning algorithm called RFC.

Ensemble machine learning algorithms performs better than other machine learning algorithms that use a single model because they combine different machine learning algorithms thereby taking advantage of their strengths [25, 27, 28]. This assertion is supported by the fact that when we used SVM we got sensitivity and specificity of 89% and 88% respectively. On the contrary, RFC gave a sensitivity of 95.24%, this is obviously a very good result because it implies that 95 out of 100 breast cancer patients were correctly diagnosed. Also, RFC gave a and specificity of almost 100% showing that all non-breast cancer patients were correctly classified as non-breast cancer patients.

The second major strength of this study is the inclusion of data pre-processing step in the modeling. It is well known that pre-processing improves computational models [30, 31]. The major preprocessing step was feature extraction. Feature extraction models are well known to improve the classification performance of machine learning algorithms [28, 33].

The third strength is the use of local data. There are very limited research efforts on breast cancer diagnosis in Nigeria and the Sub Sahara Africa as a whole because of poor facility and scarce data. We collected real data from one of the best teaching hospitals in Nigeria and used the data in this study.

The main limitation of this research is the sample size. We used only 180 subjects and that could limit the generalization of the results interpretation. Our future plan is to increase the sample size by approaching other teaching hospitals in Nigeria for data. The other limitation is the scope of the data. Data was collected only in Lagos; a nationwide data collection will obviously increase generalizability. Future efforts will increase the scope of data.

COMPETING INTERESTS

The authors do not have any financial and personal relationships with other people or organizations that could inappropriately influence this work. We, therefore, declare that no competing interests exist.

AUTHORS' CONTRIBUTIONS

BOM Conducted literature review, collected data, performed experiments and wrote the first draft of the paper.
SAA Provided information on the relevance of technical indicators to breast cancer diagnosis
OAO Was involved in data collection, contributed to the first draft and was involved in software development.
BAA Conceived and designed the study, performed experiments and co-supervised the work.
BSA Conceived and designed the study, performed experiments, supervised the entire work.
All authors read and approved the final draft of the manuscript.

REFERENCES

[1] Hortobagyi, G.N., Salazar, J.G., Pritchard, K., Amadori, D., Haidinger, R., Hudis, C.A., Khaled, H., Liu, M., Martin, M., Namer, M., O'Shaughnessy, J.A., Shen, Z.Z. and Albain, K.S, *The Global Breast Cancer Burden: Variations in Epidemiology and Survival.* Clinical Breast Cancer, 2005. **6**(5): p. 391-401.

[2] Gupta, S., Kumar, D. and Sharma, A., *Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis.* Indian Journal of Computer Science and Engineering, 2011. **2**(2): p. 188 - 195.

[3] Bhardwaj, A.a.T., A., *Breast Cancer Diagnosis Using Genetically Optimized Neural Network Model.* Expert Systems with Application, 2015.

[4] Yassin, N.I.R., Omran, S. El Houby, E.M.F. and Allam, H., *Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review.* Computer Methods and Programs in Biomedicine, 2018. **156**: p. 25-45.

[5] Alicˇkovic´, E.a.S., A., *Breast cancer diagnosis using GA feature selection and Rotation Forest.* Neural Computer and Application, 2015.

[6] Abdelaal, M.M.A., Farouq, M.W., Sena, H.A. and Salem, A.M, *Using Data Mining for Assessing Duagnosis of Breast Cancer.* Processdings of International Multiconference on Computer Science and Information Technology, 2010. **5**: p. 11-17.

[7] Onan, A., *On the Performance of Ensemble Learning for Automated Diagnosis of Breast Cancer.* Advances in Intelligent Systems and Computing, 2015.

[8] Akay, M.F., *Support vector machines combined with feature selection for breast cancer diagnosis.* Expert Systems with Applications, 2009. **36**: p. 3240–3247.

[9] Akuoko, C.P., Armah, E., Sarpong, T., Quansah, D.Y., Amankwaa, I. and Boateng, D., *Barriers to early presentation and diagnosis of breast cancer among African women living in sub-Saharan Africa.* PLOS ONE, 2017: p. 1-18.

[10] Alharbi, A.a.T., F., *A Fuzzy-Genetic Algorithm Method for the Breast Cancer Diagnosis Problem.* IARIA, 2015.

[11] Dunn, B.K., Agurs-Collins, T., Browne, D., Lubet, R. and Johnson, K.A., *Health disparities in breast cancer: biology meets socioeconomic status.* Breast Cancer Research and Treatment, 2010. **121**: p. 281-292.

[12] Martin, D.N., Boersma, B.J., Yi, M., Reimers, M., Howe, T.M., Yfantis, H.G., Tsai, Y.C., Williams, E.H., Lee, D.H., Stephens, R.M., Weissman, A.M. and Ambs, S., *Differences in the Tumor Microenvironment between African-American and European-American Breast*

Cancer Patients. PLOS ONE, 2009. **4**(2): p. 1 - 14.

[13] Dignam, J.J., *Differences in Breast Cancer Prognosis Among African-American and Caucasian Women.* A cancer Journal for Clinicians, 2000. **50**(1): p. 50-64.

[14] Pitt, J.J.e.a., *Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features.* Nature Communications, 2018.

[15] Elmore, J.G., Moceri, V.M., Carter, D. and Larson, E.B., *Breast Carcinoma Tumor Characteristics in Black and White Women.* American Cancer Society, 1998. **83**(12): p. 2509-2515.

[16] Fregene, A.a.N., L.A, *Breast Cancer in Sub-Saharan Africa: How Does It Relate to Breast Cancer in African-American Women?* American Cancer Society, 2004.

[17] Ghartey, F.N., Anyanful, A., Eliason, S., Adamu, S.M. and Debrah, S., *Pattern of Breast Cancer Distribution in Ghana: A Survey to Enhance Early Detection, Diagnosis, and Treatment.* International Journal of Breast Cancer, 2016. **2016**: p. 1-10.

[18] Zheng, Y., TomWalsh, Gulsuner, S., Casadei, S., Lee, M.K., Ogundiran, T.O., Ademola, A., Falusi, A.G., Adebamowo, C.A., Oluwasola, A.O., Adeoye, A., Odetunde, A., Babalola, C.P., Ojengbede, O.A., Odedina, S., Anetor, I., Wang, S., Huo, D., Yoshimatsu, T.F., Zhang, J., Felix, G.E.S., King, M. and Olopade, O.I., *Inherited Breast Cancer in Nigerian Women.* American Society of Clinical Oncology, 2018. **36**(28): p. 2820-2827.

[19] Shah, S.P., *Issues In Medical Diagnosis Using Computational Techniques.* Fourth International Conference on Computational Intelligence and Communication Networks, 2012: p. 348-354.

[20] Aruleba, K., Obaido, G., Ogbuokiri, B., Fadaka, A.O., Klein, A., Adekiya, T.A. and Aruleba, R.T., *Applications of Computational Methods in Biomedical Breast Cancer Imaging Diagnostics: A Review.* Journal of Imaging, 2020.

[21] Chowdhary, C.L., and Acharjya, D.P., *A Hybrid Scheme for Breast Cancer Detection using Intuitionistic Fuzzy Rough Set Technique.* International Journal of Healthcare Information Systems and Informatics, 2016. **11**(2): p. 38-61.

[22] Gullatte, M.M., Brawley, O., Kinney, A., Powe, B. and Mooney, K., *Religiosity, Spirituality, and Cancer Fatalism Beliefs on Delay in Breast Cancer Diagnosis in African American Women.* Journal of Religious Health, 2010. **49**: p. 62-72.

[23] Hunter, C.P., Redmond, C.K., Chen, V.W., Austin, D.F., Greenberg, R.S., Correa, P., Muss, H.B., Forman, M.R., Wesley, M.N., Blacklow, R.S., Kurman, R.J., Dignam, J.J., Edwards, B.K. and Shapiro, S., *Breast Cancer: Factors Associated With Stage at Diagnosis in Black and White Women.* Journal of the National Cancer Institute, 1993. **85**(14): p. 1129-1137.

[24] He, S., Wu, Q.U. and Saunders, J.R., *Breast cancer diagnosis using an artificial neural network trained by group search optimizer.* Transactions of the Institute of Measurement and Control, 2009. 00(0): p. 1-15.

[25] Guo, H.A.N., A.K., *Breast Cancer Diagnosis Using Genetic Programming Generated Feature.*