Interpretable Solutions for Breast Cancer Diagnosis with Grammatical Evolution and Data Augmentation

Yumnah Hasan(✉), Allan de Lima, Fatemeh Amerehi,
Darian Reyes Fernández de Bulnes, Patrick Healy, and Conor Ryan

University of Limerick, Limerick, Ireland
{Yumnah.Hasan,Allan.Delima,Fatemeh.Amerehi,Darian.Reyesfernandezdebulnes,
Patrick.Healy,Conor.Ryan}@ul.ie

Abstract. Medical imaging diagnosis increasingly relies on Machine Learning (ML) models. This is a task that is often hampered by severely imbalanced datasets, where positive cases can be quite rare. Their use is further compromised by their limited interpretability, which is becoming increasingly important. While *post-hoc* interpretability techniques such as SHAP and LIME have been used with some success on so-called black box models, the use of inherently understandable models makes such endeavours more fruitful. This paper addresses these issues by demonstrating how a relatively new synthetic data generation technique, STEM, can be used to produce data to train models produced by Grammatical Evolution (GE) that are inherently understandable. STEM is a recently introduced combination of the Synthetic Minority Oversampling Technique (SMOTE), Edited Nearest Neighbour (ENN), and Mixup; it has previously been successfully used to tackle both between-class and within-class imbalance issues. We test our technique on the Digital Database for Screening Mammography (DDSM) and the Wisconsin Breast Cancer (WBC) datasets and compare Area Under the Curve (AUC) results with an ensemble of the top three performing classifiers from a set of eight standard ML classifiers with varying degrees of interpretability. We demonstrate that the GE-derived models present the best AUC while still maintaining interpretable solutions.

Keywords: Augmentation · Breast Cancer · Ensemble · Grammatical Evolution · STEM

1 Introduction

In medical imaging diagnoses, where decisions can have significant implications for individual's health, it is essential to gain a thorough understanding of the factors influencing these decisions. While Machine Learning (ML) models have proven effective in diagnosing a variety of medical conditions in medical imaging [29], their limited interpretability poses a challenge to their broader adoption. Moreover, the recently introduced European Union (EU) Communication

S. Smith et al. (Eds.): EvoApplications 2024, LNCS 14634, pp. 224–239, 2024.
https://doi.org/10.1007/978-3-031-56852-7_15

on Fostering a European approach to AI [1] specifically targets explainability as a key concern for the deployment of ML and Artificial Intelligence (AI) models.

Another prevalent challenge in the medical imaging domain is the issue of class imbalance within the dataset. Methods such as Synthetic Minority Oversampling Technique (SMOTE), Edited Nearest Neighbour (ENN), and Mixup combined together as STEM [16], which leverages the full distribution of minority classes, can effectively address both inter-class and intra-class imbalances. In [16], STEM was applied in-conjunction with an ensemble of ML classifiers, producing promising outcomes. However, understanding the reasoning behind ML model predictions remains a complex task. Furthermore, as the volume of instances and the specificity of problems grow, the complexity of the derived solutions also increases.

Building trust in ML classifiers and understanding the behaviour of the solutions is pivotal to their broader acceptance. Employing inherently explainable models is a useful strategy when generating Explainable AI models. Grammatical Evolution (GE) [26], an Evolutionary Computation (EC) technique, has been used to leverage grammars to define and constrain the syntax of potential solutions, producing inherently explainable models [22].

To address these challenges, we developed a classification system based on GE. Our study includes a comprehensive comparison with an ensemble of other ML classifiers. Notably, GE models show enhanced interpretability compared to other traditional ML models. GE provide solutions in the form of symbolic expressions, offering a more intuitive understanding of the decision-making process. This emphasis on interpretability is crucial, especially in healthcare, where understanding the rationale behind decisions is of paramount importance.

Our research hypothesises that the use of the STEM augmentation technique combined with an approach rooted in GE produces more interpretable solutions as compared to the other ensemble ML classifiers.

The contributions of this paper are as follows. Firstly, we develop a method that combines a GE classifier with STEM, outperforming an ensemble of ML classifiers, as indicated by the superior AUC. Secondly, our approach distinguishes itself by offering more interpretable solutions compared to the ensemble method. Finally, the paper presents rigorous statistical analyses to comprehensively evaluate the performance of implemented data augmentation techniques on each data setup.

The rest of the paper is structured as follows: Sect. 2 reviews the existing literature. Section 3 outlines the proposed methodology, and Sect. 4 addresses experimental details performed in this work. Results and discussion are described in Sect. 5, and Sect. 6 presents the conclusion and future guidelines.

2 Literature Review

In the realm of medical applications, particularly in the context of breast cancer diagnosis, the issue of imbalanced datasets is a critical concern. Imbalances, where one class significantly outweighs the other, can introduce biases and compromise the reliability of ML models. Implementing effective strategies for class balancing, such as oversampling, undersampling, and their combination, results in a more balanced and representative training dataset [9]. Previous studies [14,17] have recognized the impact of class imbalance in medical datasets for ML tasks.

Moreover, ML algorithms have demonstrated notable efficiency in the classification of medical data. A compelling study showcases the effectiveness of ensembles, where Bayesian networks and Radial Basis Function (RBF) classifiers with majority voting resulted in an accuracy of 97% [20] when applied to the Wisconsin Breast Cancer (WBC) dataset. Furthermore, an approach that combined linear and non-linear classifiers using Micro Ribonucleic Acid (miRNA) profiling achieved an impressive accuracy of 98.5% [28].

While these findings are promising, ML algorithms may struggle to contextualize information and are susceptible to unexpected or undetected biases originating from input data. Additionally, they often lack transparent justifications for their predictions or decisions [25]. In response to this, employing GE can yield interpretable solutions. As a variant of Genetic Programming, GE evolves human-readable solutions, offering explanations for the rationale behind its classification decisions, which is a significant advantage over current paradigms in unsupervised and semi-supervised learning [10].

Previous studies have already demonstrated the effectiveness of GE across a range of ML tasks. It has proven valuable for feature generation and feature selection [11], as well as for hyperparameter optimization [24]. The GenClass system [3], built upon GE, demonstrates promising outcomes and outperforms RBF networks in certain classification problems. They utilized thirty benchmark datasets from the UCI and KEEL repositories, including Haberman, which consists of breast cancer instances. While it has excelled in these areas, there are still avenues for further exploration.

In this paper, we aim to investigate the efficiency of utilizing GE as a medical imaging classifier combined with STEM to handle imbalance distributions of data samples, particularly in breast cancer diagnosis. Leveraging the interpretive and adaptable features of GE, our objective is to achieve accurate and reliable outcomes that can be easily explained.

3 Methodology

For analysis, we utilize two primary breast cancer datasets. One consists of images, the Digital Database for Screening Mammography (DDSM) [18], while the other consists of tabular data, the WBC [31] dataset. $DDSM$ is a comprehensive collection of mammograms, encompassing both normal and abnormal images. For this study, we focused on $DDSM's$ Cancer 02 volume and three volumes of normal samples (Volume 01-03). By selecting one volume of cancer images compared to three volumes of normal images, we maintain a realistic class imbalance ratio. These images come from the Craniocaudal (CC) and Mediolateral Oblique (MLO) views of both the left and right breasts. We work with

152 cancerous images and 876 healthy ones from volumes 1-3. Each image was divided into four segments: the entire breast (**I**), the top segment (**It**), the middle segment (**Im**), and the bottom segment (**Ib**).

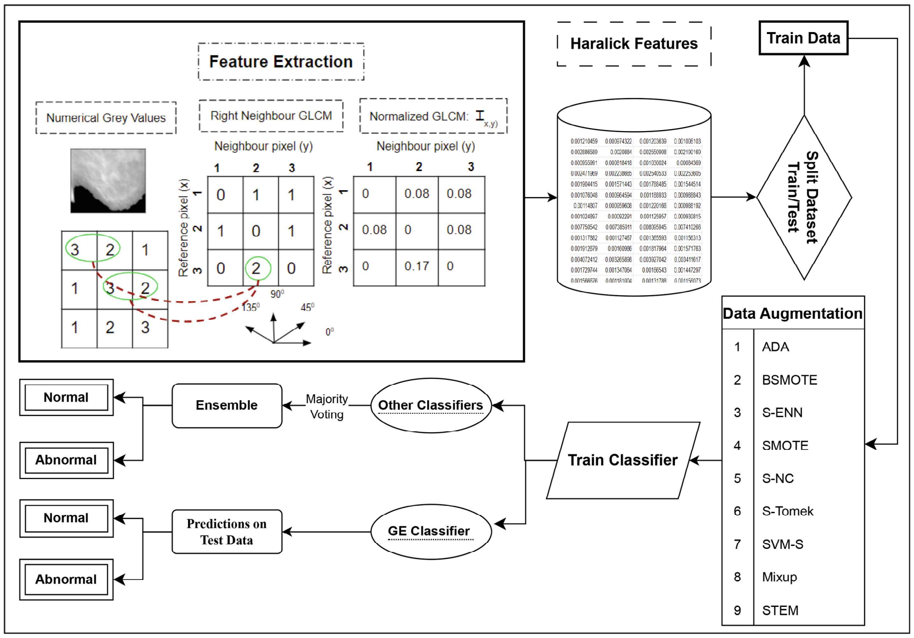


Fig. 1. Outline of the proposed approach for breast cancer classification using GE and other classifiers.

The *WBC* dataset consists of 30 features derived from Fine Needle Aspiration (FNA) samples of breast masses, categorising patients into benign (non-cancerous) and malignant (cancerous) cases. It comprises 212 malignant samples and 357 benign samples.

To create a dataset containing breast cancer images from the *DDSM* image for evaluating the proposed methodology, we first need to extract features that will be used for training. This involves isolating the breast region, eliminating irrelevant background data, segmenting the breast region, and extracting pertinent features to generate a comprehensive training dataset of breast segments. Initially, a median filter is applied to reduce noise within the images. Subsequently, non-essential background data, often containing machine-generated labels such as 'CC' or 'MLO', is removed. For this step, we employed a precise Otsu thresholding technique. Following this, the segmenting process proposed in [27] effectively partitioned images into three overlapping segments.

Feature extraction is the next critical phase. In our study, we extracted a set of Haralick's Texture Features [15] from both whole and segmented images. The selection of these features is based on the hypothesis that there are discernible textural differences between normal and abnormal images. Specifically, we compute thirteen distinct Haralick features from the Gray-Level Co-Occurrence (GLCM) matrix, employing four orientations corresponding to two diagonal (grey-level numeric values of the images) and two adjacent neighbours. This process results in generating a total of 52 features per segment or image.

High class imbalance present in the utilized datasets poses a significant challenge in developing robust and accurate predictive models. Therefore, explicit data augmentation has been implemented in the training set to effectively address this class imbalance challenge. Using nine distinct augmentation approaches outlined in Sect. 4.3, synthetic samples are generated to enrich the dataset with more discriminative information, ultimately improving the learning capabilities of the model.

In the last step, the GE classifier and an ensemble of other ML classifiers are trained separately to make predictions on the test set. Augmented training data is used, while the original imbalanced test set is used for testing. For ensembling, eight ML classifiers are used as mentioned in Sect. 4.5. The top three classifiers, based on AUC, are selected and combined through majority voting to create the final predictions. The complete pipeline of the proposed approach is shown in Fig. 1.

4 Experimental Details

The DDSM and WBC datasets are used to evaluate the proposed technique. The study employs five different data setups to train the classifiers. For the WBC dataset, a single setup is utilized, consisting of 30 breast mass features per sample acquired through FNA.

In contrast, the DDSM dataset includes images from two views, CC and MLO. To conduct experiments, the dataset is categorized into four distinct configurations based on these views. In the initial setup, denoted as "S_{CC}", data is exclusively extracted from segments of the CC view. Conversely, the second category, "S_{MLO}", comprises segmented images exclusively from the MLO view. The third configuration, "S_{CC+MLO}", combines segments from both views. Lastly, the fourth setup, "F_{CC+MLO}", considers the full image (non-segmented) features from both the CC and MLO views for comprehensive analysis. The number of features for each segment or image is 52, used in all these setups

We divided the datasets into training and testing sets at an 80:20 ratio, respectively. Notably, all $DDSM$ configurations exhibit significant class imbalances, with class ratios ranging from 6:94 S_{CC}, S_{MLO} and S_{CC+MLO} setups. For F_{CC+MLO} the ratio between the positive versus negative class is 15:85. Likewise, the WBC dataset has a class distribution of 37% positive and 63% negative classes as illustrated in Fig. 2.

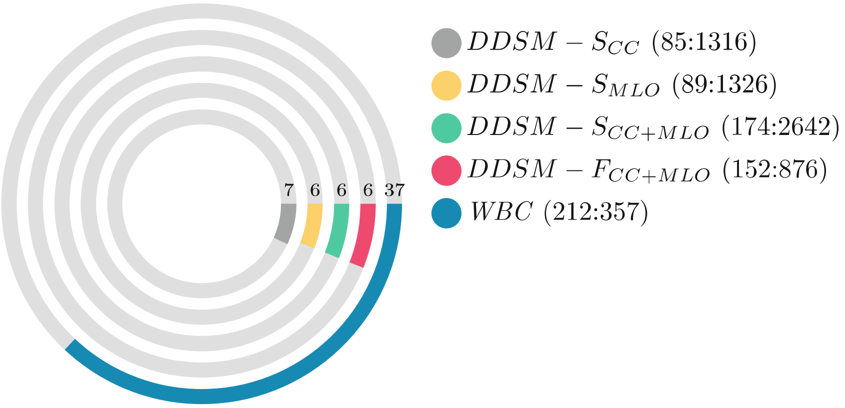


Fig. 2. Concentric ring chart for setup description. Rings are setups, and the coloured areas indicate training positive percent. Legend includes the training positive and negative total samples.

4.1 System Settings

All the ML experiments were conducted using the PyCaret library [2]. The GRAPE [8] framework was used to perform GE experiments. For statistical analysis, we employed the AutoRank Python library [19] to evaluate the performance of the implemented augmentation approaches. Our code, along with our dataset configurations, is available in our GitHub repository[1].

4.2 Performance Metric

To evaluate the performance of the designed approach, AUC has been selected as the assessment metric which uses Trapezoidal rule for its computation. AUC has become a widely accepted performance measure in classification problems due to its reliability, particularly in the context of imbalanced datasets [13,21].AUC serves as a comprehensive metric, encompassing both sensitivity (Eq. 1) and specificity (Eq. 2), considering various threshold values. $\mathrm{T_{Pos}}$ denotes true positives, $\mathrm{T_{Neg}}$ true negatives, $\mathrm{F_{Pos}}$ false positives, and $\mathrm{F_{Neg}}$ denotes false negatives.

$$Sensitivity = \frac{T_{Pos}}{T_{Pos} + F_{Neg}} \tag{1}$$

$$Specificity = \frac{T_{Neg}}{T_{Neg} + F_{Pos}} \tag{2}$$

4.3 Class Balancing

The methods utilized for generating synthetic data with the aim of equalizing the class distribution ratio include the Synthetic Minority Oversampling Technique

[1] https://github.com/yumnah3/Interpretable-Breast-Cancer-Diagnosis.git.

(SMOTE) [7], Borderline SMOTE (BSMOTE) [14], SMOTENC (S-NC) [7], Support Vector Machine SMOTE (SVM-S) [23], Mixup [32], and ADASYN (ADA) [17]. Additionally, three hybrid methods, SMOTE Edited Nearest Neighbour (S-ENN) [30] SMOTE-Tomek (S-Tomek) [5] and combination of SMOTE, ENN, and Mixup (STEM) are also implemented to compare against each other.

Notably, STEM generates a balanced number of samples for each class. Compared to other methods, it demonstrates the ability to increase the number of data samples more extensively, resulting in improved model performance.

4.4 Grammatical Evolution

GE's grammars are typically defined in Backus-Naur Form (BNF), a notation represented by the tuple N, T, P, S, where N is the set of $non-terminals$, transitional structures usually with semantic meaning, T is the set of $terminals$, items in the phenotype, P is a set of production rules, and S is a start $non-terminal$. The following simple grammar was created to evolve solutions for the first four data setups with 52 numerical features, whereas, for the last setup, 30 numerical features were used:

$$\begin{aligned}
\langle\text{expression}\rangle &::= \langle\text{operator}\rangle\texttt{(}\langle\text{expression}\rangle\texttt{,}\langle\text{expression}\rangle \mid \langle\text{operand}\rangle\\
\langle\text{operator}\rangle &::= \texttt{add} \mid \texttt{sub} \mid \texttt{mul} \mid \texttt{pdiv}\\
\langle\text{operand}\rangle &::= \langle\text{x}\rangle \mid \langle\text{digit}\rangle\langle\text{digit}\rangle\texttt{.}\langle\text{digit}\rangle\langle\text{digit}\rangle\\
\langle\text{x}\rangle &::= \texttt{x[0]}\ldots\texttt{x[51]}\\
\langle\text{digit}\rangle &::= \texttt{0} \mid \texttt{1} \mid \texttt{2} \mid \texttt{3} \mid \texttt{4} \mid \texttt{5} \mid \texttt{6} \mid \texttt{7} \mid \texttt{8} \mid \texttt{9}
\end{aligned}$$

This grammar permits the use of basic arithmetic operations (addition, subtraction, multiplication, and division –protected in case the divisor is equal to 0) and the inclusion of real numbers constants. These constants are helpful because GE can explore beyond the parameter space given to minimize the error between expected and predicted outputs, something that does not happen with other ML classifiers. The $non-terminal$ X encompasses the fifty-two numerical features for the first four setups of the DDSM dataset and the thirty numerical features for the WBC dataset.

The output domain of the evaluations is $o \in [-\infty, \infty]$. Subsequently, a sigmoid function is applied to constrain the values to $\sigma(o) \in [0, 1]$. For binary classification, the typical interpretation of the sigmoid function is the probability of belonging to class 1, and therefore we use $\sigma(o)$ to calculate AUC. Table 1 presents the experimental parameters used in this work:

Table 1. List of parameters used to run GE

Parameter type	Parameter value
Number of runs	30
Number of generations	100
Population size	200
Mutation probability	0.01
Crossover probability	0.8
Elitism size	1
Codon size	255
Initialisation	Sensible
Maximum initial depth	10
Maximum depth	35
Wrapping	0

4.5 Other Classifiers

We also used the augmented training data to train a diverse ensemble of eight ML classifiers. This ensemble includes Random Forest (RF), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis, LightGBM, XGBoost, AdaBoost, KNN, and Extra Trees models. Initially, a comprehensive model is trained using all eight classifiers. Subsequently, based on the AUC metric, the three best-performing models are selected. These selected models are then combined through a majority voting approach. The final predictions are made on the test dataset, which consists of imbalanced and unseen samples.

5 Results and Discussion

To evaluate the performance of the proposed method, five distinct data setups are employed. Four configurations are derived from the DDSM dataset, considering variations in views, segments, and full images. The fifth setup is from the WBC dataset. To enhance the robustness of the training setups, nine augmentation approaches are applied and compared. The assessment is conducted using an ensemble of other ML classifiers, alongside GE.

The performance of the classifiers is compared based on AUC for each dataset. The ensemble classifiers are denoted by their respective initials: L_d for Linear Discriminant Analysis, Q for Quadratic Discriminant Analysis, E for ExtraTree, R for Random Forest, L_i for Lightgbm, K for KNN, A for Adaboost, and X for Xgboost. It is important to note that the AUC values of the other ensemble classifiers are presented for a single run, and they are then compared against the median AUC derived from 30 runs conducted with GE.

Table 2 provides an overview of the results. In the first setup, S_{CC}, an AUC of 0.91 was achieved, outperforming the ensemble of L_dQE, which obtained an

AUC of 0.90. Similarly, in the second setup, S_{MLO}, an AUC of 0.90 was attained, while the ensemble of L_dQE achieved a slightly lower AUC of 0.84.

For the third setup S_{CC+MLO}, an AUC of 0.92 was observed using the GE classifier, outperforming other classifiers that yielded the highest AUC of 0.87 using L_dQE. When the classifiers were trained on full image features in setup F_{CC+MLO}, the highest AUC values were 0.94 and 0.85, obtained by the GE classifier and the ensemble of L_dQE, respectively.

When comparing the AUC using the WBC dataset, both GE and the ensemble of AKL_r achieved an AUC of 0.99.

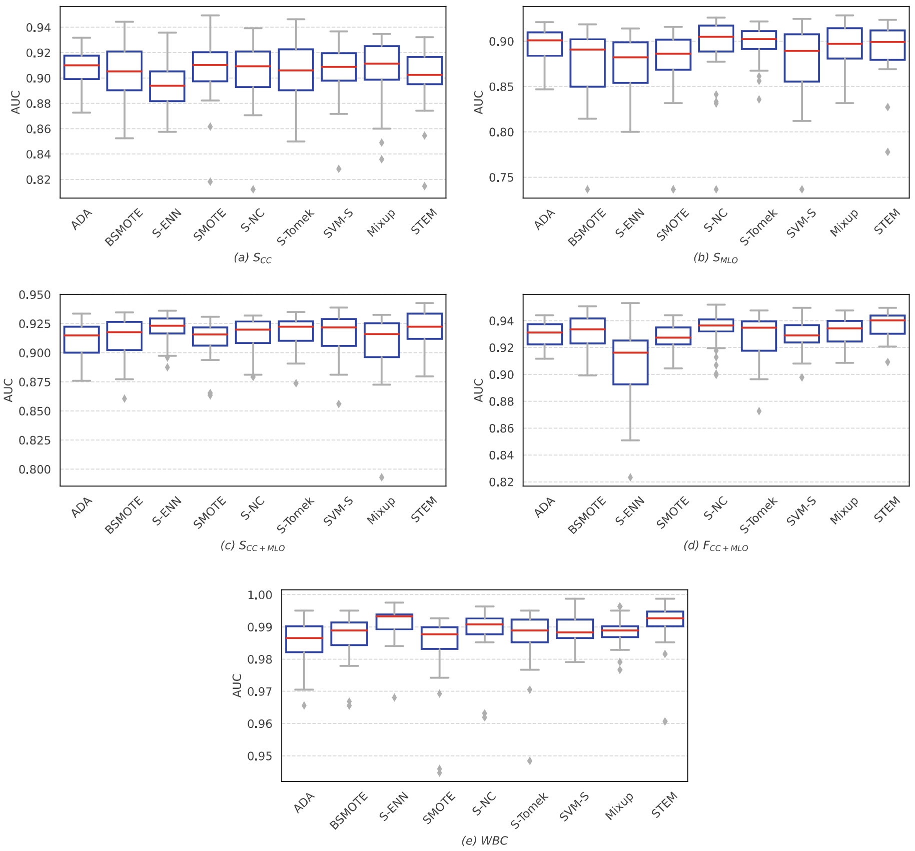
Table 2. A comparison of the AUC for GE and the ensemble approaches using the nine different augmentation techniques for each data setup.

Setups	Classifiers	ADA	BSMOTE	S-ENN	SMOTE	S-NC	S-Tomek	SVM-S	Mixup	STEM
S_{CC}	GE	0.91	0.90	0.89	0.91	0.90	0.90	0.90	0.91	0.90
	Others	0.76	0.73	0.93	0.77	0.82	0.77	0.73	0.90	0.90
		L_dQE	L_dQE	L_dQE	L_dQE	L_dQE	L_dQE	L_dQE	L_dQE	L_dQE
S_{MLO}	GE	0.90	0.90	0.90	0.90	0.87	0.90	0.89	0.90	0.89
	Others	0.80	0.80	0.80	0.82	0.78	0.82	0.81	0.81	0.84
		EL_iR	EL_iR	L_dQE	EL_iR	EL_iX	EL_iX	EL_iR	L_dQE	L_dQE
S_{CC+MLO}	GE	0.91	0.91	0.92	0.91	0.92	0.91	0.91	0.90	0.91
	Others	0.75	0.68	0.77	0.75	0.70	0.76	0.62	0.76	0.87
		EL_iR	EL_iR	EL_iR	EL_iR	EL_iX	EL_iR	EL_iR	EL_iR	L_dQE
F_{CC+MLO}	GE	0.93	0.91	0.90	0.92	0.93	0.94	0.93	0.93	0.93
	Others	0.78	0.84	0.72	0.81	0.82	0.82	0.82	0.81	0.85
		EQR	EL_iR	ERX	EQR	EL_iQ	EL_iR	EQR	L_iQL_d	L_dQE
WBC	GE	0.98	0.98	0.99	0.98	0.99	0.98	0.99	0.98	0.99
	Others	0.94	0.94	0.94	0.94	0.95	0.94	0.94	0.94	0.99
		L_dQE	L_dQE	EKL_i	L_dQE	L_dQE	L_dQE	L_dQE	L_dEL_i	AKL_r

The augmentation approaches are compared using the boxplot presented in Fig. 3. The plot indicates the AUC obtained from all nine augmentation approaches for each setup across all 30 runs. The horizontal line in red indicates the median value of the respective group.

GE provides valuable insights into the most informative features used in the solutions, as demonstrated in Table 3, which present the most frequently used features for each setup. The features extracted and presented in these tables are sorted by their impact on the solutions. Common features consistently found in Table 3 for the $DDSM$ dataset include "Inverse Difference Moment (IDM)"(feature 17), "Contrast" (feature 5), and "Difference Entropy" (feature 41). Both contrast and IDM represent the difference in grey levels between pixels, while entropy indicates the level of randomness in the grey levels.

For the WBC dataset, as shown in Table 3, the top three features that consistently appear in the solutions are 21, 20, and 27, corresponding to "Concave Point Worst", "Fractal Dimension", and "Radius Worst" respectively. The concave point worst feature indicates the severity of the concave portion of the shape,



(a) S_{CC}

(b) S_{MLO}

(c) S_{CC+MLO}

(d) F_{CC+MLO}

(e) WBC

Fig. 3. Boxplot analysis comparing opponent approaches and their AUC distributions across multiple runs

with "worst" denoting the highest mean value. The "fractal dimension" is a crucial characteristic that provides information related to the geometric shape of the fractals. The third feature, radius worst, represents the largest mean value for the distances from the centre to points on the perimeter.

While other ML models may share the feature of interpretability, they often present challenges that GE does not encounter. Decision trees and RF, though interpretable, lose clarity with complex structures and aggregation [4]. LDA relies on the Gaussian distribution of the data and assumes that the covariance of two classes is the same [12], limiting its applicability. In contrast, GE does not depend on these factors and maintains transparency throughout its evolution, even when addressing complex and non-linear problems.

Table 3. This analysis unveils prevalent features used by GE in all five setups. For S_{CC} and S_{MLO}, percentages are computed from 8684 and 7945 occurrences. Likewise, contributions to S_{CC+MLO} and F_{CC+MLO} are based on 8138 and 8522 occurrences, respectively. The features of WBC are also examined, with percentages drawn from 9076 appearances.

S_{CC}		S_{MLO}		S_{CC+MLO}		F_{CC+MLO}		WBC	
Feature	Usage	Feature	Usage	Feature	Usage	Feature	Usage	Feature	Usage
17	6.22%	5	4.93%	17	5.35%	41	3.78%	21	7.83%
41	4.87%	4	4.46%	5	4.36%	37	3.71%	20	7.64%
38	4.58%	41	3.95%	7	4.33%	4	3.63%	27	6.47%
5	4.19%	7	3.75%	41	4.02%	38	3.46%	24	5.56%
18	3.88%	17	3.65%	18	3.93%	11	3.38%	1	5.1%
7	3.50%	34	3.34%	38	3.55%	17	3.18%	13	4.87%
36	2.73%	45	3.15%	11	3.08%	5	3.11%	7	4.23%

5.1 Statistical Analysis

The statistical comparison of implemented data augmentation techniques involved a non-parametric Bayesian signed-rank test [6] applied to each dataset. In our analysis, conducted on nine augmentation techniques with 30 paired AUC samples each, the test distinguished between methods being pair-wise larger, smaller or inconclusive. The approaches listed in the rows are compared with the methods presented in the corresponding column. The subsequent Bayesian signed-rank test revealed significant distinctions among the techniques. In the cases where STEM has outperformed the other approaches are underlined in the Table 4.

In the S_{CC} setup, as illustrated in Table 4(a), STEM, Mixup, SMOTE, ADA, S-NC, SVM-S, S-Tomek and BSMOTE all exhibit larger medians than S-ENN.

The statistical comparison of medians depicted in Table 4(b) among various augmentation populations reveals notable differences for S_{MLO} setup. STEM, S-NC, S-Tomek, ADA, and Mixup exhibit larger medians compared to BSMOTE, SVM-S, SMOTE, and S-ENN.

Similarly, for setup S_{CC+MLO} in the Table 4(c) STEM again showcases its effectiveness by outperforming S-NC, BSMOTE, Mixup, SMOTE, and ADA in medians. Additionally, S-ENN demonstrates superiority by exhibiting larger medians than Mixup, SMOTE, and ADA. Additionally, S-Tomek outperforms SMOTE in median values. SVM-S, in particular, stands out with a larger median than ADA.

Moreover, STEM stands out by consistently surpassing S-Tomek, Mixup, BSMOTE, ADA, SVM-S, SMOTE, and S-ENN in median values presented in Table 4(d) for F_{CC+MLO} . Additionally, S-NC demonstrates superiority over SMOTE and S-ENN, while S-Tomek outperforms S-ENN in median values. Mixup, BSMOTE, ADA, SVM-S, and SMOTE all exhibit larger medians than S-ENN.

Finally, in the *WBC* setup, as depicted in Table 4(e), STEM emerged as the top-performing method, surpassing S-NC, BSMOTE, S-Tomek, Mixup, SVM-S, SMOTE, and ADA. S-NC exhibited a higher median than SMOTE and ADA, while Mixup outperformed SMOTE in median value. SVM-S demonstrated a larger median than SMOTE and ADA.

The Bayesian analysis results are summarized in Fig. 4. It reveals that STEM, a combination of S-ENN and Mixup, emerges as the top-ranking approach. This result underscores the effectiveness of this combined strategy in enhancing performance. Notably, S-ENN and Mixup individually secure the second and third positions, further affirming the significance of this ensemble approach.

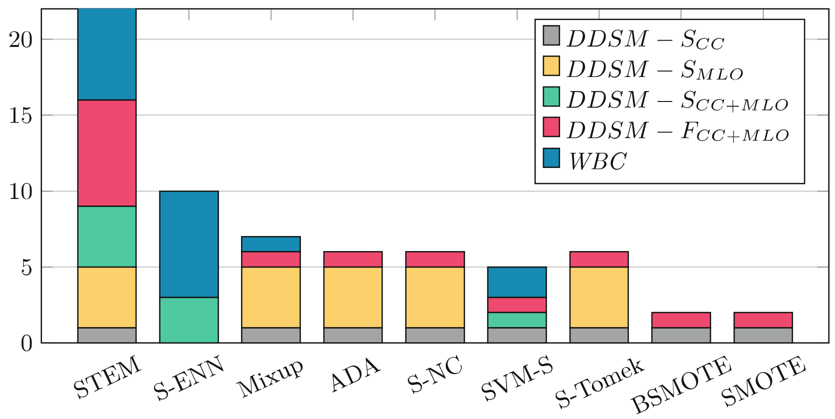


Fig. 4. The illustration of the overall results acquired from the Bayesian signed-rank test is shown here. The cumulative score is the total number of times one approach outperforms the other. STEM obtained a cumulative score of 23 where the maximum possible is 40 (comparing one versus another 8 approaches in 5 setups), outperforming the other approaches. Each color represents distinct test setups used for the evaluation.

Table 4. The results of the Bayesian signed-ranked test are summarized here for the nine augmentation approaches for each data setup. Arrows indicate the direction of differences: ⇑ for larger, ⇓ for smaller, - for inconclusive, and N/A for not applicable results. A family-wise significance level of $\alpha \equiv 0.05$ is employed.

(a) S_{CC}

	STEM	Mixup	SMOTE	ADA	S-NC	SVM-S	S-Tomek	BSMOTE	S-ENN
STEM	N/A	–	–	–	–	–	–	–	$\underline{\Uparrow}$
Mixup	–	N/A	–	–	–	–	–	–	⇑
SMOTE	–	–	N/A	–	-	–	–	–	⇑
ADA	–	–	–	N/A	–	–	–	–	⇑
S-NC	–	-	-	-	N/A	-	-	-	⇑
SVM-S	–	–	–	–	–	N/A	-	–	⇑
S-Tomek	–	–	-	-	–	–	N/A	–	⇑
BSMOTE	–	–		–	–	–	–	N/A	⇑
S-ENN	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	N/A

(b) S_{MLO}

	STEM	Mixup	S-NC	S-Tomek	ADA	BSMOTE	SVM-S	SMOTE	S-ENN
STEM	N/A	–	–	–	–	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$
Mixup	–	N/A	–	–	–	⇑	⇑	⇑	⇑
S-NC	–	–	N/A	–	–	⇑	⇑	⇑	⇑
S-Tomek	–	–	-	N/A	–	⇑	⇑	⇑	⇑
ADA	–	–	–	–	N/A	⇑	⇑	⇑	⇑
BSMOTE	⇓	⇓	⇓	⇓	⇓	N/A	–	–	–
SVM-S	⇓	⇓	⇓	⇓	⇓	–	N/A	–	–
SMOTE	⇓	⇓	⇓	⇓	⇓	–	–	N/A	–
S-ENN	⇓	⇓	⇓	⇓	⇓	–	–	–	N/A

(c) S_{CC+MLO}

	STEM	S-ENN	S-Tomek	SVM-S	S-NC	BSMOTE	Mixup	SMOTE	ADA
STEM	N/A	–	–	–	–	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$
S-ENN	–	N/A	–	–	–	–	⇑	⇑	⇑
S-Tomek	–	–	N/A	-	-	–	–	–	–
SVM-S	–	–	–	N/A	-	-	-	-	⇑
S-NC	–	–	–	–	N/A	–	–	–	–
BSMOTE	⇓	-	–	–	–	N/A	–	–	–
Mixup	⇓	⇓	–	–	–	–	N/A	–	–
SMOTE	⇓	⇓	—	–	–	–	–	N/A	–
ADA	⇓	⇓	–	⇓	–	–	–	–	N/A

(d) F_{CC+MLO}

	STEM	S-NC	Mixup	S-Tomek	BSMOTE	ADA	SVM-S	SMOTE	S-ENN
STEM	N/A	–	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$	$\underline{\Uparrow}$
S-NC	–	N/A	–	–	–	–	–	–	⇑
Mixup	⇓	–	N/A	–	–	–	–	–	⇑
S-Tomek	⇓	-	-	N/A	-	-	-	-	⇑
BSMOTE	⇓	–	–	–	N/A	–	–	–	⇑
ADA	⇓	–	–	–	–	N/A	–	–	⇑
SVM-S	⇓	–	–	–	–	–	N/A	–	⇑
SMOTE	⇓	–	–	–	–	–	–	N/A	⇑
S-ENN	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	N/A

(*continued*)

Table 4. (*continued*)

(e) *WBC*

	STEM	S-ENN	S-NC	SVM-S	Mixup	ADA	BSMOTE	S-Tomek	SMOTE
STEM	N/A	–	$\underline{\boldsymbol{\Uparrow}}$	$\underline{\boldsymbol{\Uparrow}}$	$\underline{\boldsymbol{\Uparrow}}$	$\underline{\boldsymbol{\Uparrow}}$	$\underline{\boldsymbol{\Uparrow}}$	$\underline{\boldsymbol{\Uparrow}}$	$\underline{\boldsymbol{\Uparrow}}$
S-ENN	–	N/A	⇑	⇑	⇑	⇑	–	⇑	⇑
S-NC	⇓	⇓	N/A	-	–	–	–	–	–
SVM-S	⇓	⇓	–	N/A	–	⇓	–	–	⇓
Mixup	⇓	⇓	–	–	N/A	–	–	–	⇑
ADA	⇓	⇓	–	⇓	–	N/A	–	–	–
BSMOTE	⇓	–	–	–	–	–	N/A	–	–
S-Tomek	⇓	⇓	–	–	–	–	–	N/A	–
SMOTE	⇓	⇓	–	⇓	⇓	–	–	–	N/A

6 Conclusion and Future Work

In this study, we addressed class imbalance and interpretability challenges in medical imaging diagnosis by using GE to produce classifier trained on data augmented by the recently-introduced STEM technique. Our approach not only delivers interpretable solutions but also outperforms an ensemble of other ML classifiers in terms of performance. The analysis conducted on the *DDSM* and *WBC* datasets emphasizes the effectiveness of GE, as evidenced by improvements in AUC and its ability to identify critical data features. Notably, our inclusion of Bayesian signed-rank test results confirms that STEM emerges as the best-performing approach for augmentation. The improved AUC and enhanced interpretability of our approach can help build trust and facilitate informed decisions. Thus, our study validates the proposed hypothesis, demonstrating the efficacy of the combined GE and STEM approach.

For future research, we suggest improving performance by incorporating additional image attributes, such as wavelet transformations and local binary patterns, to enhance the feature set and dataset diversity. Furthermore, exploring the mixture of different datasets to assess the robustness of our approach across various image data sources would be interesting.

Acknowledgements. The Science Foundation Ireland (SFI) Centre for Research Training in Artificial Intelligence (CRT-AI), Grant No. 18/CRT/6223 and the Irish Software Engineering Research Centre (Lero), Grant No. 16/IA/4605, both provided funding for this study.

References

1. Communication on Fostering a European approach to Artificial Intelligence — Shaping Europe's digital future (Apr 2021)
2. Ali, M.: Pycaret: an open source, low-code machine learning library in python version 2.3 (2020)

3. Anastasopoulos, N., Tsoulos, I.G., Tzallas, A.: Genclass: a parallel tool for data classification based on grammatical evolution. SoftwareX **16**, 100830 (2021)
4. Arrieta, A.B., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Inform. Fusion **58**, 82–115 (2020)
5. Batista, G.E., Bazzan, A.L., Monard, M.C., et al.: Balancing training data for automated annotation of keywords: a case study. Wob **3**, 10–8 (2003)
6. Benavoli, A., Corani, G., Mangili, F., Zaffalon, M., Ruggeri, F.: A bayesian wilcoxon signed-rank test based on the dirichlet process. In: International Conference on Machine Learning, pp. 1026–1034. PMLR (2014)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artifi. Intell. Res. **16**, 321–357 (2002)
8. de Lima, A., Carvalho, S., Dias, D.M., Naredo, E., Sullivan, J.P., Ryan, C.: GRAPE: grammatical Algorithms in Python for Evolution. Signals **3**(3), 642–663 (2022). https://doi.org/10.3390/signals3030039
9. Fernández, A., López, V., Galar, M., Del Jesus, M.J., Herrera, F.: Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches. Knowl.-Based Syst. **42**, 97–110 (2013)
10. Fitzgerald, J.M., Azad, R.M.A., Ryan, C.: GEML: Evolutionary unsupervised and semi-supervised learning of multi-class classification with Grammatical Evolution. In: 2015 7th International Joint Conference on Computational Intelligence (IJCCI), vol. 1, pp. 83–94 (Nov 2015)
11. Gavrilis, D., Tsoulos, I.G., Dermatas, E.: Selecting and constructing features using grammatical evolution. Pattern Recogn. Lett. **29**(9), 1358–1365 (2008). https://doi.org/10.1016/j.patrec.2008.02.007
12. Ghojogh, B., Crowley, M.: Linear and quadratic discriminant analysis: Tutorial. arXiv preprint arXiv:1906.02590 (2019)
13. Halimu, C., Kasem, A., Newaz, S.S.: Empirical comparison of area under roc curve (auc) and mathew correlation coefficient (mcc) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In: Proceedings of the 3rd International Conference on Machine Learning and Soft Computing, pp. 1–6 (2019)
14. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005). https://doi.org/10.1007/11538059_91
15. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Trans. Syst. Man Cybernet. 610–621 (1973)
16. Hasan, Y., Amerehi, F., Healy, P., Ryan, C.: Stem rebalance a novel approach for tackling imbalanced datasets using smote, edited nearest neighbour, and mixup (2023). https://arxiv.org/abs/2311.07504
17. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. IEEE (2008)
18. Heath, M., et al.: Current status of the digital database for screening mammography. In: Digital Mammography: Nijmegen, pp. 457–460. Springer (1998). https://doi.org/10.1007/978-94-011-5318-8_75
19. Herbold, S.: Autorank: a Python package for automated ranking of classifiers. J. Open Source Softw. **5**(48), 2173 (2020). https://doi.org/10.21105/joss.02173

20. Jabbar, M.A.: Breast cancer data classification using ensemble machine learning. Eng. Appli. Sci. Res. **48**(1), 65–72 (2021)
21. Liang, X., Jiang, A., Li, T., Xue, Y., Wang, G.: Lr-smote-an improved unbalanced data set oversampling based on k-means and svm. Knowl.-Based Syst. **196**, 105845 (2020)
22. Murphy, A., Murphy, G., Amaral, J., MotaDias, D., Naredo, E., Ryan, C.: Towards incorporating human knowledge in fuzzy pattern tree evolution. In: Hu, T., Lourenço, N., Medvet, E. (eds.) EuroGP 2021. LNCS, vol. 12691, pp. 66–81. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72812-0_5
23. Nguyen, H.M., Cooper, E.W., Kamei, K.: Borderline oversampling for imbalanced data classification. Inter. J. Knowl. Eng. Soft Data Paradigms **3**(1), 4–21 (2011). https://doi.org/10.1504/IJKESDP.2011.039875
24. Noorian, F., de Silva, A.M., Leong, P.H.W.: gramEvol: grammatical evolution in R. J. Stat. Softw. **71**, 1–26 (2016). https://doi.org/10.18637/jss.v071.i01
25. Rashed, B.M., Popescu, N.: Machine learning techniques for medical image processing. In: 2021 International Conference on E-Health and Bioengineering (EHB), pp. 1–4 (Nov 2021). https://doi.org/10.1109/EHB52898.2021.9657673
26. Ryan, C., Collins, J.J., Neill, M.O.: Grammatical evolution: Evolving programs for an arbitrary language. In: Banzhaf, W., Poli, R., Schoenauer, M., Fogarty, T.C. (eds.) EuroGP 1998. LNCS, vol. 1391, pp. 83–96. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0055930
27. Ryan, C., Krawiec, K., O'Reilly, U.-M., Fitzgerald, J., Medernach, D.: Building a stage 1 computer aided detector for breast cancer using genetic programming. In: Nicolau, M., et al. (eds.) EuroGP 2014. LNCS, vol. 8599, pp. 162–173. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44303-3_14
28. Sharma, S.K., Vijayakumar, K., Kadam, V.J., Williamson, S.: Breast cancer prediction from microRNA profiling using random subspace ensemble of LDA classifiers via Bayesian optimization. Multimedia Tools Appli. **81**(29), 41785–41805 (2022). https://doi.org/10.1007/s11042-021-11653-x
29. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Med. **5**(1), 1–8 (2022). https://doi.org/10.1038/s41746-022-00592-y
30. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst. Man Cybernet., 408–421 (1972)
31. Wolberg, W.H., Street, W.N., Mangasarian, O.L.: Breast cancer wisconsin (diagnostic) data set [uci machine learning repository] (1992)
32. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: Beyond Empirical Risk Minimization (Apr 2018). https://doi.org/10.48550/arXiv.1710.09412