An enhanced Genetic Folding algorithm for prostate and breast cancer detection

Mohammad A. Mezher[1], Almothana Altamimi[2], Ruhaifa Altamimi[3],

[1] Computer Science Department, Fahd Bin Sultan University, Tabuk, SA
[2] Department of Clinical Medicine and Surgery, Università degli Studi di Napoli Federico II, IT
[3] Department of Business and Data Analytics, University of Huddersfield, UK

Corresponding Author:
Mohammad A. Mezher[1]
Jordan Str., Tabuk, 15700, Saudi Arabia
Email address: mmezher@fbsu.edu.sa

Abstract

Cancer's genomic complexity is gradually increasing as we learn more about it. Genomic classification of various cancers is crucial in providing oncologists with vital information for targeted therapy. Thus, issues of patient genomic classification become more pertinent to address. Prostate cancer is a cancer subtype that exhibits extreme heterogeneity. Prostate cancer contributes to 7.3% of new cancer cases worldwide, with a high prevalence in males. Breast cancer is the most common type of cancer in women and the second most significant cause of death from cancer in women. Breast cancer is caused by abnormal cell growth in the breast tissue, generally referred to as a tumour. Tumours are not synonymous with cancer; they can be benign (noncancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Fine Needle Aspiration (FNA) tests are used to biopsy the breast to diagnose breast cancer. Artificial Intelligence (AI) and Machine Learning (ML) models are used to diagnose with varying accuracy. In light of this, we used the Genetic Folding (GF) algorithm to predict prostate cancer status in a given dataset. An accuracy of 96% was obtained, thus being the current highest accuracy in prostate cancer diagnosis. The model was also used in breast cancer classification with a proposed pipeline that used Exploratory Data Analysis (EDA), label encoding, feature standardization, feature decomposition, log transformation, detect and remove the outliers with Z-score, and the BAGGINGSVM approach attained a 95.96% accuracy. The accuracy of this model was then assessed using the rate of change of PSA, age, BMI, and filtration by race. We discovered that integrating the rate of change of PSA

34   and age in our model raised the model's area under the curve (AUC) by 6.8%, whereas BMI and

35   race had no effect. As for breast cancer classification, no features were removed.

37   **Keywords:** Genetic Folding algorithm, evolutionary algorithms, prostate cancer, classification,

38   genetic programming, support vector machine.

40   Introduction

41   Cancer is, at times, a severe disease or set of diseases that have historically been the most prevalent

42   and challenging to treat (Adjiri, 2016). It is commonly defined as the abnormal proliferation of

43   various human cells, thus resulting in its etiological heterogeneity (Cooper, 2000). This abnormal

44   growth can be classified into two subsets (a) Malignant and (b) Benign. Benign tumours stay

45   localized to their original site, whereas malignancies can invade and spread throughout the body

46   (metastasize) (Cooper, 2000). Breast cancer recently eclipsed lung cancer as the most prominent

47   cancer subtype worldwide, equalling 11.7% (2.3 million) of the new cancer cases (Sung et al.,

48   2021).

49   Prostate cancer is one of the most prevalent male malignancies globally, contributing 7.3% of the

50   estimated incidence in 2020 (Sung et al., 2021). This amounted to 1,414,259 new cases and

51   375,304 deaths from this disease (Sung et al., 2021). The prostate is a dense fibromuscular gland

52   shaped like an upside-down cone. It is located around the neck of the urinary bladder, external to

53   the urethral sphincter and functions in a supportive role in the male reproductive system. The

54   alkaline fluid it secretes into semen protects sperm cells from the acidic vaginal environment

55   (Singh & Bolla, 2021).

56   Although early prostate cancer is typically asymptomatic, it may manifest itself in the form of

57   excessive urination, nocturia, haematuria, or dysuria (Leslie et al., 2021). Classically, prostate

58   cancer is detected with a Digital Rectal Examination (DRE) and a blood test for prostate-specific

59   antigen (PSA) (Descotes, 2019). TRUS-guided biopsy continues to be the gold standard for

60   confirming diagnoses, despite its 15–46% false-negative rate and up to 38% tumour under grading

61   rate compared to the Gleason score (Descotes, 2019; Kvåle et al., 2009).

62   However, prostate cancer development's aetiology and mechanisms are still determined (Howard

63   et al., 2019). The different mechanisms that develop prostate cancer ultimately affect the therapy

64   proposed. Hence, patient stratification and tumour classification are vital. Especially true

65 considering that prostate cancer is heterogeneous in a clinical, spatial and morphological aspect
66 (Tolkach & Kristiansen, 2018). The following prostate cancer stages of development are currently
67 posited: intraepithelial-neoplasia, androgen-dependent adenocarcinoma, and androgen-
68 independent or Castration Resistant Cancer (CRC) (Howard et al., 2019). Prostate cancer
69 heterogeneity is further highlighted in a recent study shedding light on mRNA expressional
70 variations from a normal prostate to full metastatic disease (Marzec et al., 2021). This means that
71 Metastatic CRC (mCRC) tumours are more complicated than primary prostate tumours. An
72 association that is further aggravated when genomics come into play.

73 Prostate cancer development and progression have been heavily linked to Androgen Receptor
74 (AR) signalling pathway; thus, Androgen Deprivation Therapy (ADT) has been used for patients
75 who have advanced prostate cancer (Hatano & Nonomura, 2021). Although primarily effective, a
76 large proportion of patients develop androgen-independent or CRC. Thus, pharmacological
77 therapies are considered, including abiraterone, enzalutamide, docetaxel and radium-223 (Howard
78 et al., 2019).

79 As mentioned above, breast cancer has become the most common cancer diagnosed, eclipsing lung
80 cancer. The etiology of breast cancer is multi-faceted; many risk factors play a role in the
81 likelihood of a diagnosis; these risk factors can be sub-divided into seven groups. (1) age (2) gender
82 (3) previous diagnosis of breast cancer (4) histology (5) family history of breast cancer, (6)
83 reproduction-related risk factors and (7) exogenous hormone use (Alkabban & Ferguson, 2022).
84 Race may also indicate a higher prevalence in non-Hispanic white individuals than African
85 Americans, Hispanics, Native Americans, and Asian Americans (Alkabban & Ferguson, 2022).

86 Many modalities of screening exist with varying specificities and sensitivities to breast cancer.
87 Mammography is one of the front-line tests to screen for breast cancer, with its sensitivity ranging
88 from 75-90% and its specificity ranging from 90-95% (Bhushan et al., 2021). Screening is done at
89 a macro level to determine whether the growth is malignant or benign. However, classification at
90 the micro-level is required to identify the molecular basis behind breast cancer progression to target
91 specific therapies. Luminal-A tumours (58.5%) are the most prevalent subtype of breast cancer
92 tumours, then triple-negative (16%), luminal-B (14%) and HER-2 positive (11.5%) being the least
93 prevalent (Al-thoubaity, 2019).

94 Breast cancer tumours are classified using the TNM classification system in which the primary
95 tumour is denoted as T, the regional lymph nodes as N, and distant metastases as M. Breast cancer

96 can also be classified about its invasiveness, with lobular carcinoma in situ (LCIS) and ductal
97 carcinoma in situ (DCIS) being non-invasive. Invasive ductal cancer accounts for 50-70% of
98 invasive cancers, while invasive ductal cancer accounts for 10% (Alkabban & Ferguson, 2022).
99 Once the molecular basis of breast cancer is identified, treatment can begin. Many
100 chemotherapeutic agents are used in the treatment of breast cancer, including Tamoxifen (ER-
101 positive breast cancer), Pertuzumab (HER2-overexpressing breast cancer) and Voxtalisib
102 (HR+/HER- advanced breast cancer) (Bhushan et al., 2021).
103 This paper aims to develop a proper kernel function using GF to separate benign prostate and
104 breast cells from tumour-risk genetic cells using SVM, compared to six machine learning
105 algorithms. In addition, the proposed GF-SVM classifier can classify breast and prostate cancer
106 cells better than the six different classifiers. We have also included all features found in the datasets
107 to test the ability of the proposed modelling to predict the best accuracy. The proposed GF-SVM
108 implementation is reliable, scalable, and portable and can be an effective alternative to other
109 existing evolutionary algorithms (M. A. Mezher, 2022).
110
111 **Literature Review**
112 Artificial intelligence (AI) is revolutionizing healthcare, incredibly patient stratification and
113 tumour classification, which are pivotal components of targeted oncology. Training a Machine
114 Learning (ML) model to analyze large datasets means more accurate prostate cancer diagnoses
115 (Tătaru et al., 2021). Artificial Neural Networks (ANN) is a tool that has been used in advanced
116 prognostic models for prostate cancer (Jović et al., 2017). Other models employed in the
117 classification of cancers based on gene expression include K-Nearest Neighbor (KNN) and
118 Support Vector Machines (SVM). The accuracy of these models varies between 70% and 95%,
119 depending on the number of genes analyzed. Previous studies have used these and other ML
120 models to classify prostate cancer (Bouazza et al., 2015) thus, we aim to use the Genetic Folding
121 (GF) algorithm in classifying patients with prostate cancer.
122 (Alba et al., 2007) used a hybrid technique for gene selection and classification. They presented
123 data of high dimensional DNA Microarray. They found it initiates a set of suitable solutions early
124 in their development phase. Similarly (Tahir & Bouridane, 2006), found that using a hybrid
125 algorithm significantly improved their results in classification accuracy. However, they concluded
126 their algorithm was generic and fit for diagnosing other diseases such as lung or breast cancer.

127   (Bouatmane et al., 2011) used an RR-SFS method to find a 99.9% accuracy. This was alongside

128   other classification methods such as bagging/boosting with a decision tree. (Lorenz et al., 1997)

129   compared the results from two commonly used classifiers, KNN and Bayes classifier. They found

130   78% and 79%, respectively. They had the opportunity to improve these results for ultrasonic tissue

131   characterisation significantly.

132   Developing a quantitative CADx system that can detect and stratify the extent of breast cancer

133   (BC) histopathology images was the primary clinical objective for (Basavanhally et al., 2010), as

134   they demonstrated in their paper the ability to detect the extent of BC using architectural features

135   automatically. BC samples with a progressive increase in Lymphocytic Infiltration (LI) were

136   arranged in a continuum. The region-growing algorithm and subsequent MRF-based refinement

137   allow LI to isolate from the surrounding BC nuclei, stroma, and baseline level of lymphocytes

138   (Basavanhally et al., 2010). The ability of this image analysis to classify the extent of LI into low,

139   medium and high categories can show promising translation into prognostic testing.

140   A meta-analysis identified several trends concerning the different types of machine learning

141   methods in predicting cancer susceptibility and outcomes. It was found that a growing number of

142   machine learning methods usually improve the performance or prediction accuracy of the

143   prognosis, particularly when compared to conventional statistical or expert-based systems (Cruz

144   & Wishart, 2006). There is no doubt that improvements in experimental design alongside

145   biological validation would enhance many machine-based classifiers' overall quality and

146   reproducibility (Cruz & Wishart, 2006). As the quality of studies into machine learning classifiers

147   improves, there is no doubt that they will become more routine use in clinics and hospitals.

148

149   **The Proposed Model**

150   *A. Dataset*

151   The prostate dataset of 100 patients (Shown in Table 1) was used to test the proposed GF algorithm

152   and analyze the outcomes. The dataset consists of 100 observations and nine variables; Radius

153   (Rad.), Texture (Text.), Perimeter (Perim.), Area, Smoothness (Smooth), Compactness (Compact),

154   Symmetry, Fractal_dimension (Fractal_dim.), and Diagnosis.

155   The breast cancer dataset is a repository maintained by the University of California. The dataset

156   contains 569 samples of malignant and benign tumor cells. Columns 1-30 contain real-value

157   features that have been computed from digitized images of the cell nuclei. Table 2 shows the first

158   seven columns found in the breast dataset (D. & C., 2019).

159   All the prostate cancer dataset observations included in this experiment have been collected from

160   Kaggle.com (Saifi, 2018). Prostate and Breast cancer patients were labeled with M, whereas those

161   without the cancer were labeled with B, as seen in the following tables.

162

163   ***B. Support Vector Machine***

164   This section will go through the basic SVM concepts in the context of two-class classification

165   problems, either linearly or non-linearly separable. SVM is based on the Vapnik–Chervonenkis

166   (VC) theory and the Structural Risk Minimization (SRM) principle (Shawe-taylor et al., 1996).

167   The goal is to identify the optimum trade-off between lowering training set error and increasing

168   the margin to get the highest generalization ability while resisting overfitting. Another significant

169   benefit of SVM is convex quadratic programming, which generates only global minima,

170   preventing the algorithm from being trapped in local minima. (Cristianini & Shawe-Taylor, 2000)

171   and (Cristianini et al., 2001) provide in-depth explanations of SVM theory. The remaining part of

172   this section will go through the fundamental SVM ideas and apply them to classic linear and

173   nonlinear binary classification tasks.

174   **1.   Linear Margin Kernel Classifier**

175   Suppose a binary classification problem is given as $\{x_i, y_i\}$, $x_i \in \Re N$ and a set of corresponding

176   labels are notated $y_i \in \{-1,1\}$ for i=1, 2, …, N, where $\Re N$ denotes vectors in a d-dimensional

177   feature space. A hyperplane provided in SVM separates the data points using equation (1):

178                          $$(x_i) = w^T \cdot X + b = 0 \tag{1}$$

179   Where **w** is an n-dimensional coefficient vector that is normal to the hyperplane and *b* is the offset

180   from the origin and X is the features in a sample.

181   **2.   Nonlinear Margin Kernel Classifier**

182   It has been shown that the input space can be mapped onto a higher-dimensional feature space

183   where the training set cannot be separated. The input vectors are mapped into a new, higher-

184   dimensional feature space, indicated as $F{:}\Re N {\to} H^K$, where N < K, to produce the SVM model:

185

186
$$F(x) = w^T \cdot K(x_i, x_j) + b \tag{2}$$

187   The mapping's functional (x), is implicitly determined by the kernel trick $K(x_i, x_j) = \varphi(x_i) \cdot \varphi($

188   x_j), which is the dot product of two feature vectors in extended feature space. In SVMs, an

189   extensive range of kernel tricks are accessible for application. The most common SVM kernels

190   are shown in Table 3. In the table, γ and p are predefined user parameters.

191   ***C. Genetic Folding Algorithm***

192   The GF algorithm was created by (M. A. Mezher & Abbod, 2011). The GF algorithm's

193   fundamental principle is to group math equations using floating numbers. Floating numbers are

194   created by taking random samples of operands and operators. In GF, the linear chromosomes

195   serve as the genotype, while the parse trees serve as the phenotype. This genotype/phenotype

196   technique works well for encoding a lengthy parse tree in each chromosome. The GF approach

197   has shown to be effective in various computer problems, including binary, multi-classification,

198   and regression datasets. For example, GF has been proven to outperform other members of the

199   evolutionary algorithm family in binary classification, multi-classification (M. Mezher & Abbod,

200   2014), and regression (M. Mezher & Abbod, 2012).

201   On the other hand, GF may yield kernels employed in SVMs. Kernels are created by combining

202   operands and operators to generate appropriate pairings, such as $(3 + 4)$. At each pair, the pair is

203   indexed at random in a GF array cell (known as GF kernel). Correlation between pairings is

204   boosted by picking pairs at random that boost the strength of the generated GF kernels. Each GF

205   kernel chromosome was divided into a head segment that only carries functions and a tail segment

206   containing terminals. However, the size of the head segment must be determined ahead of time,

207   but the size of the tails segment does not need to be determined since the GF algorithm predicts

208   the number of genes needed based on the pairs required for the drawn functions (M. Mezher &

209   Abbod, 2017). Furthermore, the GF algorithm predicts the number of operands (terminals)

210   necessary each time the GF algorithm generates different operators (function) at random. The

211   proposed modified GF algorithm has the pseudo-code that can be seen in Figure 1:

212   The following operands and operators were used to forecast malignant and benign cells in the

213   prostate dataset, as shown in Table 4:

214

215   The GF genome comprises a continuous, symbolic string or chromosome of equal length. Table

216   5 shows an example of a chromosome with varying pairs that may be used to create a valid GF

217   kernel.

218

219   **D. Model Evaluations**

220   The classification performance of each model is evaluated using statistical classification accuracy.

221   The equation (3) is used to determine the accuracy of the GF algorithm's correctly classified

222   instances:

223    $$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$ (3)

224   The True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) define

225   this accuracy measurement. A TP is made when the algorithm predicts malignant (positive), and

226   the actual result is malignant (positive). A TN is made when the algorithm predicts benign

227   (negative), and the actual result is benign (negative). FP occurs when the algorithm predicts a

228   benign (negative) instance as malignant (positive). Finally, when the GF algorithm classifies a

229   malignant (positive) instance as benign (negative), the result is FN. The accuracy performance

230   metrics are compared in Table 5.

231   We have also included another evaluating estimator to measure the quality of the proposed mode

232   to choose the best. Some of the estimators tolerate the small samples better than others, while other

233   measures the quality of the estimator with large samples. Mean Square Error (MSE) (Schluchter,

234   2014) has been used as a standard metric to measure model performance in medical, engineering,

235   and educational studies. Assume E' is the predicted classes E'={(pred(y1), pred(y2), pred(y3),

236   … } of the observed classes E={y1, y2, y3,..}, then the MSE is defined as the expectation of the

237   squared deviation of E' from E:

238    $$MSE(E') = \sum_{i=1}^{n} \frac{(E' - E)^2}{n}$$ (4)

239   MSE measures the estimator's bias (accuracy), which shows how much its predicted value deviates

240   consistently from the actual value and the estimator's variance (precision), which indicates how

241   much its expected value fluctuates due to sampling variability. Squaring the errors magnifies the

242   effect of more significant errors. These calculations disproportionately penalize larger errors more

243   than more minor errors. This attribute is critical if we want the model to be as accurate as possible.

244

Results

246 We are preprocessing the dataset using min-max standardization, without removing any features

247 that prevented incorrect relevance assignments. The elimination of the features showed either

248 weakness or strength correlation and reduced the accuracy values, as shown in Table 5. In

249 statistics, correlation is any connection between two variables but does not indicate causation,

250 implying that the variables are dependent on one another. The closer a variable is to 1, the more

251 significant the positive correlation; the closer a variable is to -1, the stronger the negative

252 correlation; and the closer a variable is to 0, the weaker the negative correlation. Each malignant

253 datapoint was subjected to analysis. We implemented the proposed method using Visual Studio

254 Code by Python.

255 This paper conducted our experiments on 100 patients of 8 features collected from kaggle.com

256 (Saifi, 2018)for prostate cancer and 596 patients of 30-features for the breast cancer dataset. The

257 instances were classified as malignant or benign in the experiments. The following are the steps

258 involved in conducting the GF algorithm to produce a valid GF kernel. GF starts by correctly

259 generating the initial GF genes (operators, operands) (operand, operator, operand). Then, the GF

260 algorithm generates a valid GF chromosome containing 50 genes. Based on a 5-fold cross-

261 validation approach, the dataset was divided into a training set and a testing set. The detailed GF

262 algorithm is explained in Figure 1 (Mohammad et al., 2016).

263 Figures 2 and 3 (a-e) depict the complexity, diversity, ROC curve, accuracy, and best GF

264 chromosome results for prostate and breast datasets. Figures 2 and 3 (a) show the GF algorithm's

265 performance concerning the complexity of the generic kernel. Figures 2 and 3 (b) show the

266 population's diversity in each generation, where GF maintains the best fitness values with a

267 chance of accepting weak kernels. Figures 2 and 3 (c) shows that GF generic kernel was the best

268 Area Under Curve (AUC) value compared with conventional SVM classifiers; linear, RBF, and

269 polynomial kernel functions. Based on the MSE, the proposed model shows (Figure 2, 3 (e))

270 minor differences between the observed values and the predicted classes. In Figure 2 (e), the

271 proposed model beats the other kernels with at least the average of errors was better than the

272 minimum box plot of the best model (RBF kernel). Figure 3 (e) shows a perfect model with no

273 error produced of almost zero variance. In order to show the performance of the GF algorithm in

274 each generic kernel, median, average, and standard deviation are calculated in Figure 2.3 (d). The

275   best GF kernel represented in the Tree structure is shown in Figures 2, 3 (f) on the scaled prostate

276   dataset and with a setting hyperparameter.

Discussion

278   The proposed GF was deployed for the first time in the prostate/breast cancer detection dataset,

279   and it demonstrated a significant performance improvement over the existing models in this

280   domain. Experiments were carried out with a broad set of features from the conducted datasets.

281   Table 6 gives a comparative description of several ML algorithms based on the same preprocessing

282   conditions. The accuracy performance of the proposed GF model was superior compared to the

283   six ML approaches in the prostate cancer dataset. Furthermore, we expanded this comparison with

284   the suggested hybrid model by using the SVM classifier with several conventional kernels such as

285   linear, polynomial, and RBF kernels.

286   The proposed model achieved 96.0% accuracy in the prostate cancer dataset, which is better than

287   the ANN by 16% and the LR by 6%. The proposed GF beat the best-preset kernel by 8%, 20%,

288   and 16% compared to linear, RBF, and polynomial kernels, respectively. The best kernel found

289   for prostate cancer dataset (shown Figure 2(f)) is:

290   ['Plus_s', 'Multi_s', 'Multi_s', 'Plus_s', 'Minus_s', 'Minus_v', 'Minus_v', 'Plus_s', 'Plus_s', 'x',

291   'Minus_s', 'Plus_v', 'Plus_v', 'Plus_v', 'x', 'Plus_s', 'Minus_s', 'Minus_v', 'Minus_v', 'Minus_v', 'x',

292   'x', 'x', 'x', 'x', 'x', 'x']

293   The folding indicis of the best GF chromosome found for the produced kernel were:

294   ['1.2', '3.4', '5.6', '7.8', '9.10', '0.5', '0.6', '11.12', '13.14', '0.9', '15.16', '17.18', '19.20', '21.22', '0.14',

295   '23.24', '25.26', '0.17', '0.18', '0.19', '0.20', '0.21', '0.22', '0.23', '0.24', '0.25', '0.26']

296   In the same side, the best kernel string found for breast cancer dataset was shownin Figure 3(f):

297   ['Plus_s', 'Plus_s', 'x', 'Plus_s', 'Multi_s', 'Multi_s', 'Minus_v', 'Plus_s', 'Plus_s', 'x', 'Minus_v', 'x',

298   'x', 'x', 'Minus_v']

299   Where the folding indicis of the best GF chromosome found for breast cancer dataset were:

300   ['1.2', '3.4', '0.2', '5.6', '7.8', '9.10', '0.6', '11.12', '13.14', '0.9', '0.10', '0.11', '0.12', '0.13', '0.14']

301

302

303

304

305

306   As seen in Figure 4, the proposed GF model performed the best for prostate cancer classification.

307   After using 5-Fold, the absolute average accuracy on the prostate cancer dataset was 96.0%.

308   Additionally, when evaluated on a prostate cancer dataset, dimensionality reduction had an

309   enhancing impact.

310   Consequently, the RF model achieved an accuracy of 96.0 %, more significant than the proposed

311   model by 0.09%, which was the best result. The accuracy comparison for breast cancer is shown

312   in Figure 5, where all features are included. The proposed model was validated using just 5-folds,

313   while the RF model-averaged accuracies ranged from 0-fold to 20-folds.

314   Although our system does not get the highest accuracy results in the breast cancer dataset, it is a

315   close second. To emphasize the relevance of our findings, we compare the P-values of the

316   hypothesis tests for both predefined SVM kernels (Figure 6 (a and b)) and ML models (Figure 6

317   (c and d)) to the reported P-value results for the proposed models in Figures 4 and 5. Except for

318   the prostate cancer dataset, all results are significant compared to a standard $P = 0.05$ criterion.

319   Except for RBF, the SVM kernels in both datasets have an insignificant structure to GF.

320   Furthermore, in both SVM kernels and ML models, the dataset size played the most critical role

321   in GF's ability to forecast all negative samples and a representative sample for each positive

322   sample. Figure 6 displays the P-value results for both the prostate and cancer datasets compared

323   to the predefined kernel-SVM and ML models.

324 **Conclusions**

325   This paper demonstrated that using a GF algorithm to classify patients with prostate cancer

326   provides better accuracy than KNN, SVM, DT, and LR models. The GF algorithm achieved an

327   average accuracy of 96% without eliminating any features from the dataset. This enables

328   healthcare professionals to classify prostate cancers more precisely and provide more targeted

329   therapies. Further improvements can be introduced to the model's accuracy based on our results.

330   Using multidimensional data whilst choosing a range of feature selection/classification algorithms

331   can be a promising tool for early-onset prostate cancer classification. We plan to work on the

332   concept further and apply it to other types of cancer.

333

334 **References**

335   Adjiri, A. (2016). Identifying and Targeting the Cause of Cancer is Needed to Cure Cancer.

336   *Oncology and Therapy*, *4*(1), 17–33. https://doi.org/10.1007/s40487-015-0015-6

337   Alba, E., Garcia-Nieto, J., Jourdan, L., & Talbi, E.-G. (2007). Gene selection in cancer

338   classification using PSO/SVM and GA/SVM hybrid algorithms. *2007 IEEE Congress on*

339   *Evolutionary Computation*, 284–290. https://doi.org/10.1109/CEC.2007.4424483

340   Alkabban, F. M., & Ferguson, T. (2022). Breast Cancer. In *StatPearls*. StatPearls Publishing.

341   http://www.ncbi.nlm.nih.gov/books/NBK482286/

342   Al-thoubaity, F. K. (2019). Molecular classification of breast cancer: A retrospective cohort

343   study. *Annals of Medicine and Surgery*, *49*, 44–48.

344   https://doi.org/10.1016/j.amsu.2019.11.021

345   Anjelito, D., & Mezher, M. (2022). *MMezher_14_4_2022_ML:Breast Cancer Classification*.

346   https://www.kaggle.com/code/mozmezher/mmezher-14-4-2021-ml-breast-cancer-

347   classification/edit/run/92993553

348   Basavanhally, A. N., Ganesan, S., Agner, S., Monaco, J. P., Feldman, M. D., Tomaszewski, J.

349   E., Bhanot, G., & Madabhushi, A. (2010). Computerized Image-Based Detection and

350   Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology. *IEEE*

351   *Transactions on Biomedical Engineering*, *57*(3), 642–653.

352   https://doi.org/10.1109/TBME.2009.2035305

353   Bhushan, A., Gonsalves, A., & Menon, J. U. (2021). Current State of Breast Cancer Diagnosis,

354   Treatment, and Theranostics. *Pharmaceutics*, *13*(5), 723.

355   https://doi.org/10.3390/pharmaceutics13050723

356   Bouatmane, S., Roula, M. A., Bouridane, A., & Al-Maadeed, S. (2011). Round-Robin sequential

357   forward selection algorithm for prostate cancer classification and diagnosis using

358   multispectral imagery. *Machine Vision and Applications*, *22*(5), 865–878.

359   https://doi.org/10.1007/s00138-010-0292-x

360   Bouazza, S. H., Hamdi, N., Zeroual, A., & Auhmani, K. (2015). Gene-expression-based cancer

361   classification through feature selection with KNN and SVM classifiers. *2015 Intelligent*

362   *Systems and Computer Vision (ISCV)*, 1–6.

363   https://doi.org/10.1109/ISACV.2015.7106168

364   Cijov, A. (2021). *Prostate Cancer ⧠ Adaboost from Scratch*.

365   https://kaggle.com/alincijov/prostate-cancer-adaboost-from-scratch

366   Cooper, G. M. (2000). The Development and Causes of Cancer. *The Cell: A Molecular*

367   *Approach. 2nd Edition*. https://www.ncbi.nlm.nih.gov/books/NBK9963/

368   Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and*

369   *Other Kernel-based Learning Methods*. Cambridge University Press.

370   https://doi.org/10.1017/CBO9780511801389

371   Cristianini, Nello, & Shawe-Taylor, J. (2001). An introduction to support vector machines and

372   other kernel-based learning methods. Repr. *Introduction to Support Vector Machines*

373   *and Other Kernel-Based Learning Methods*, 22.

374   https://doi.org/10.1017/CBO9780511801389

375   Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and

376   Prognosis. *Cancer Informatics*, *2*, 117693510600200030.

377   https://doi.org/10.1177/117693510600200030

378   D., D., & C., G. (2019). *UCI Machine Learning Repository*. [http://archive.ics.uci.edu/ml]

379   Descotes, J.-L. (2019). Diagnosis of prostate cancer. *Asian Journal of Urology*, *6*(2), 129–136.

380   https://doi.org/10.1016/j.ajur.2018.11.007

381   Gomes, S. (2021). *Prostate cancer prediction model*. https://kaggle.com/smogomes/prostate-

382   cancer-prediction-model

383   Hatano, K., & Nonomura, N. (2021). Genomic Profiling of Prostate Cancer: An Updated Review.

384   *The World Journal of Men's Health*, *39*. https://doi.org/10.5534/wjmh.210072

385   Howard, N., Clementino, M., Kim, D., Wang, L., Verma, A., Shi, X., Zhang, Z., & DiPaola, R. S.

386   (2019). New developments in mechanisms of prostate cancer progression. *Seminars in*

387   *Cancer Biology*, *57*, 111–116. https://doi.org/10.1016/j.semcancer.2018.09.003

388   Jović, S., Miljković, M., Ivanović, M., Šaranović, M., & Arsić, M. (2017). Prostate Cancer

389   Probability Prediction By Machine Learning Technique. *Cancer Investigation*, *35*(10),

390   647–651. https://doi.org/10.1080/07357907.2017.1406496

391   Kvåle, R., Møller, B., Wahlqvist, R., Fosså, S. D., Berner, A., Busch, C., Kyrdalen, A. E.,

392   Svindland, A., Viset, T., & Halvorsen, O. J. (2009). Concordance between Gleason

393   scores of needle biopsies and radical prostatectomy specimens: A population-based

394   study. *BJU International*, *103*(12), 1647–1654. https://doi.org/10.1111/j.1464-

395   410X.2008.08255.x

396   Leslie, S. W., Soon-Sutton, T. L., Sajjad, H., & Siref, L. E. (2021). Prostate Cancer. In

397   *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK470550/

398   Lorenz, A., Blum, M., Ermert, H., & Senge, T. (1997). Comparison of different Neuro-Fuzzy

399   classification systems for the detection of prostate cancer in ultrasonic images. *1997*

400   *IEEE Ultrasonics Symposium Proceedings. An International Symposium (Cat.*

401   *No.97CH36118)*, *2*, 1201–1204 vol.2. https://doi.org/10.1109/ULTSYM.1997.661794

402   Marzec, J., Ross-Adams, H., Pirrò, S., Wang, J., Zhu, Y., Mao, X., Gadaleta, E., Ahmad, A. S.,

403   North, B. V., Kammerer-Jacquet, S.-F., Stankiewicz, E., Kudahetti, S. C., Beltran, L.,

404   Ren, G., Berney, D. M., Lu, Y.-J., & Chelala, C. (2021). The Transcriptomic Landscape

405   of Prostate Cancer Development and Progression: An Integrative Analysis. *Cancers*,

406   *13*(2), 345. https://doi.org/10.3390/cancers13020345

407   Mezher, M. A. (2022). GFLibPy: An Open-Source Python Toolbox for Genetic Folding

408   Algorithm. In A. M. A. Musleh Al-Sartawi (Ed.), *Artificial Intelligence for Sustainable*

409   *Finance and Sustainable Technology* (pp. 423–437). Springer International Publishing.

410   https://doi.org/10.1007/978-3-030-93464-4_42

411   Mezher, M. A., & Abbod, M. F. (2011). Genetic Folding: A New Class of Evolutionary

412   Algorithms. In M. Bramer, M. Petridis, & A. Hopgood (Eds.), *Research and Development*

413   *in Intelligent Systems XXVII* (pp. 279–284). Springer London.

414   Mezher, M., & Abbod, M. (2014). Genetic folding for solving multiclass SVM problems. *Applied*

415   *Intelligence*, *41*, 464–472. https://doi.org/10.1007/s10489-014-0533-1

416   Mezher, M., & Abbod, M. (2017). A Novel Genetic Operator for Genetic Folding Algorithm: A

417   Refolding Operator and a New Genotype. *International Journal of Engineering and*

418   *Manufacturing*, *7*, 1–11. https://doi.org/10.5815/ijem.2017.06.01

419   Mezher, M., & Abbod, M. F. (2012). A New Genetic Folding Algorithm for Regression Problems.

420   *2012 UKSim 14th International Conference on Computer Modelling and Simulation*, 46–

421   51. https://doi.org/10.1109/UKSim.2012.107

422   Ribeiro, W. (2021). *Logistic Regression Classifier Prostate Cancer*.

423   https://kaggle.com/willribeiro/logistic-regression-classifier-prostate-cancer

424   Saifi, S. (2018). *Prostate Cancer*. https://www.kaggle.com/sajidsaifi/prostate-cancer

425   Schluchter, M. D. (2014). Mean Square Error. In *Wiley StatsRef: Statistics Reference Online*.

426   John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat05906

427   Şenol, C. (2021). *ANN Prediction*. https://kaggle.com/cemhansenol98/ann-prediction

428   Shawe-taylor, J., Holloway, R., Bartlett, P. L., Williamson, R. C., & Anthony, M. (1996).

429   *Structural Risk Minimization over Data-Dependent Hierarchies*.

430   Singh, O., & Bolla, S. R. (2021). Anatomy, Abdomen and Pelvis, Prostate. In *StatPearls*.

431   StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK540987/

432   Smogomes, & Mezher, M. (2022). *MMezher_14_4_2022_Prostate cancer prediction model*.

433   https://kaggle.com/mozmezher/mmezher-14-4-2022-prostate-cancer-prediction-model

434   Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F.

435   (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality

436   Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, *71*(3),

437   209–249. https://doi.org/10.3322/caac.21660

438   Tahir, M. A., & Bouridane, A. (2006). Novel Round-Robin Tabu Search Algorithm for Prostate

439   Cancer Classification and Diagnosis Using Multispectral Imagery. *IEEE Transactions on*

440   *Information Technology in Biomedicine*, *10*(4), 782–793.

441   https://doi.org/10.1109/TITB.2006.879596

442   Tătaru, O. S., Vartolomei, M. D., Rassweiler, J. J., Virgil, O., Lucarelli, G., Porpiglia, F.,

443   Amparore, D., Manfredi, M., Carrieri, G., Falagario, U., Terracciano, D., de Cobelli, O.,

444   Busetto, G. M., Giudice, F. D., & Ferro, M. (2021). Artificial Intelligence and Machine

445   Learning in Prostate Cancer Patient Management—Current Trends and Future

446   Perspectives. *Diagnostics*, *11*(2), 354. https://doi.org/10.3390/diagnostics11020354

447   Tolkach, Y., & Kristiansen, G. (2018). The Heterogeneity of Prostate Cancer: A Practical

448   Approach. *Pathobiology: Journal of Immunopathology, Molecular and Cellular Biology*,

449   *85*(1–2), 108–116. https://doi.org/10.1159/000477852

450   Will Ribeiro. https://www.kaggle.com/willribeiro/logistic-regression-classifier-prostate-cancer.

451   [Accessed 13.11.2021]

Figure 1

Pseudo-code of GF Algorithm

#Initialization

Set the operators and operands needed along with the size of GF chromosome

input $Op_k = \{plus\_v, plus\_s, minus\_v, minus\_s, multi\_s,\}$, $Operand_n = \{1,.., n-1\}$,

Len = GF chromosome length

Initialize the pool of GF pairs

set $S = (indx_0, operand_1, Op_1, operand_2), …, (indx_{n-1}, operand_n, Op_k, operand_n)$

GF chromosome/population generation

Algorithm GFs

    1.   For pop= 1 to number_of_generation

    2.   For i=1 to len

    3.   M= Select (Op_i, Operands) from S

    4.   GF_chromosome = concatenate (GF_chromosome, M)

    5.   End for

    6.   Construct pop = pop + GF_chromosome

    7.   Fitness = SVM(GF_chromosome)

    8.   pop = pop + 1

    9.   End for

    10. New kernels are generated to the most fitness

    11. Result: Optimum GF kernel that have highest accuracy

Figure 2

GF algorithm results for Prostate cancer dataset

(a) accuracy versus complexity (b) population diversity (c) the area under the curve (AUC) (d)

Accuracy values (e) MSE values for each SVM kernel (f) GF kernel function

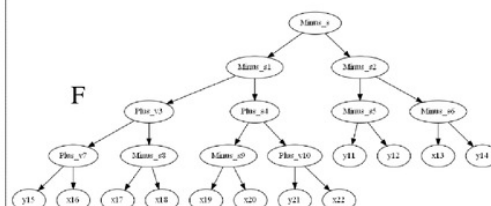
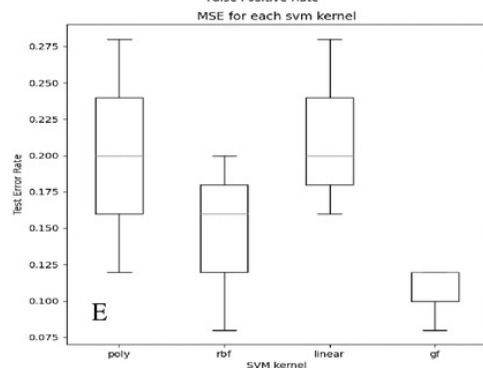
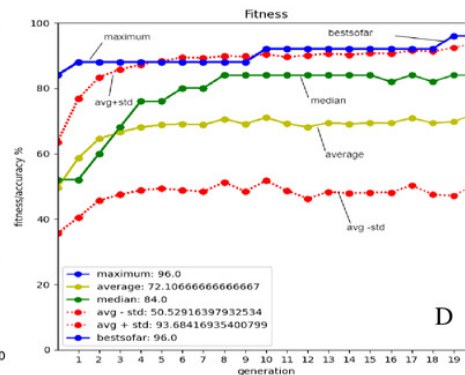
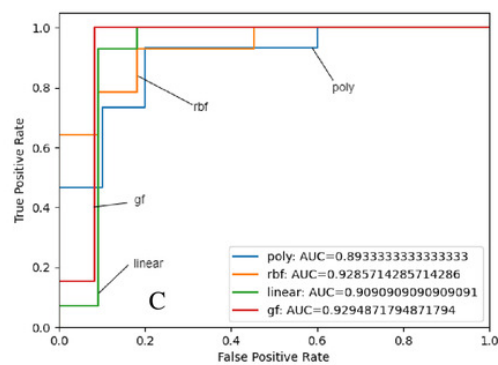
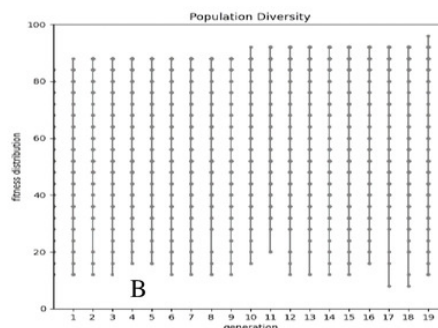
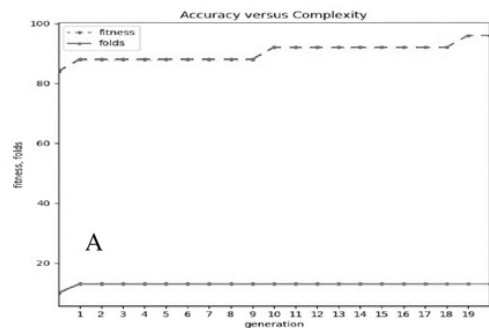



Figure 3

GF algorithm results for Breast cancer dataset

(a) accuracy versus complexity (b) population diversity (c) the area under the curve (AUC) (d)

Accuracy values (e) MSE values for each SVM kernel (f) GF kernel function

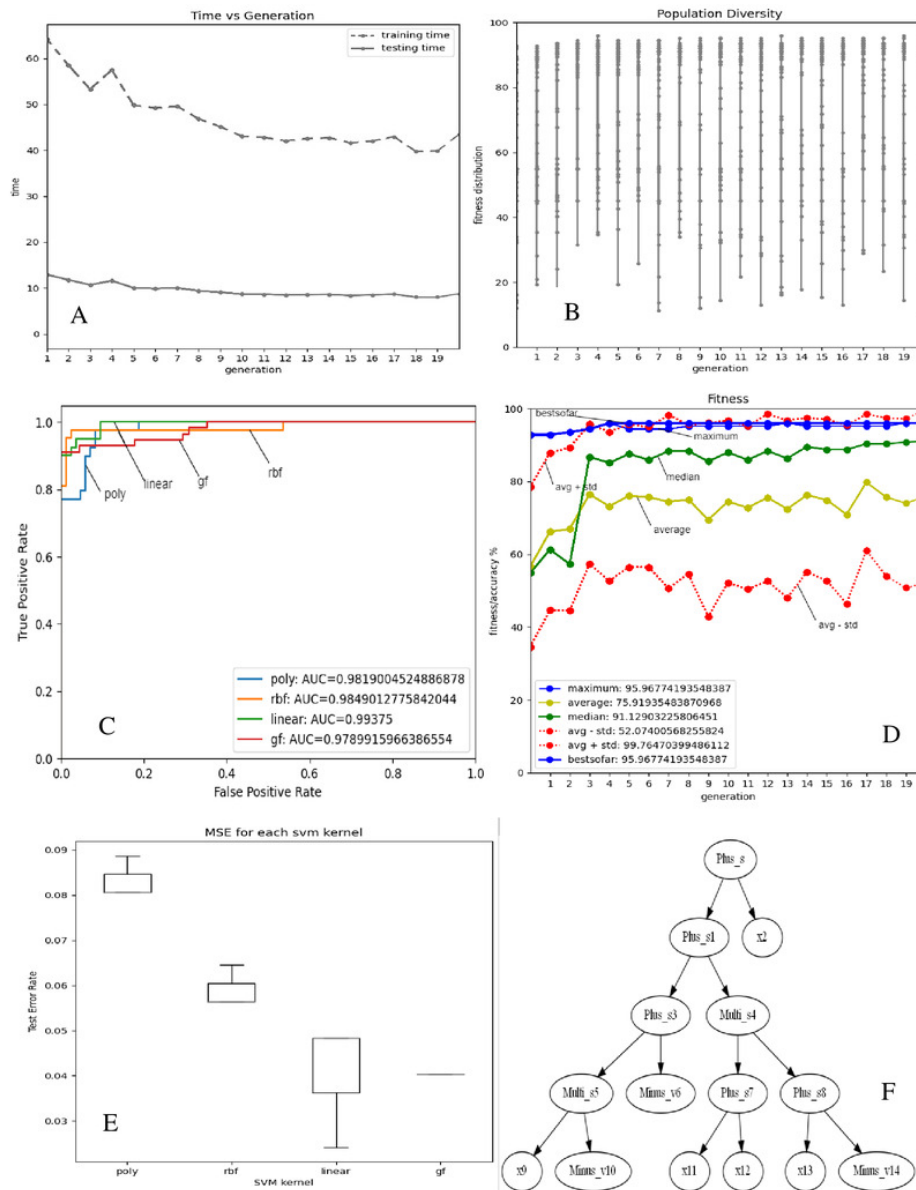



Figure 4

Accuracies comparisons for prostate cancer dataset

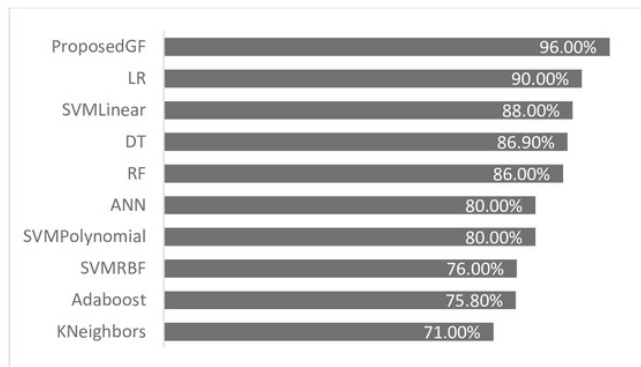



Figure 5

Accuracies comparisons for breast cancer dataset

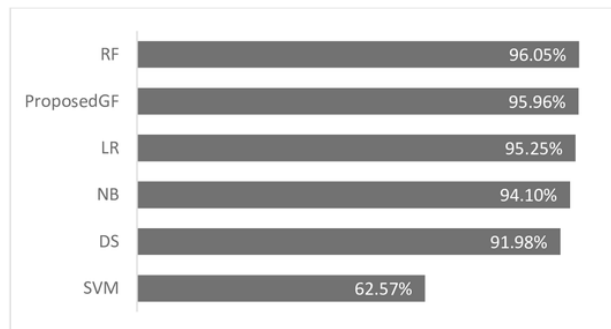
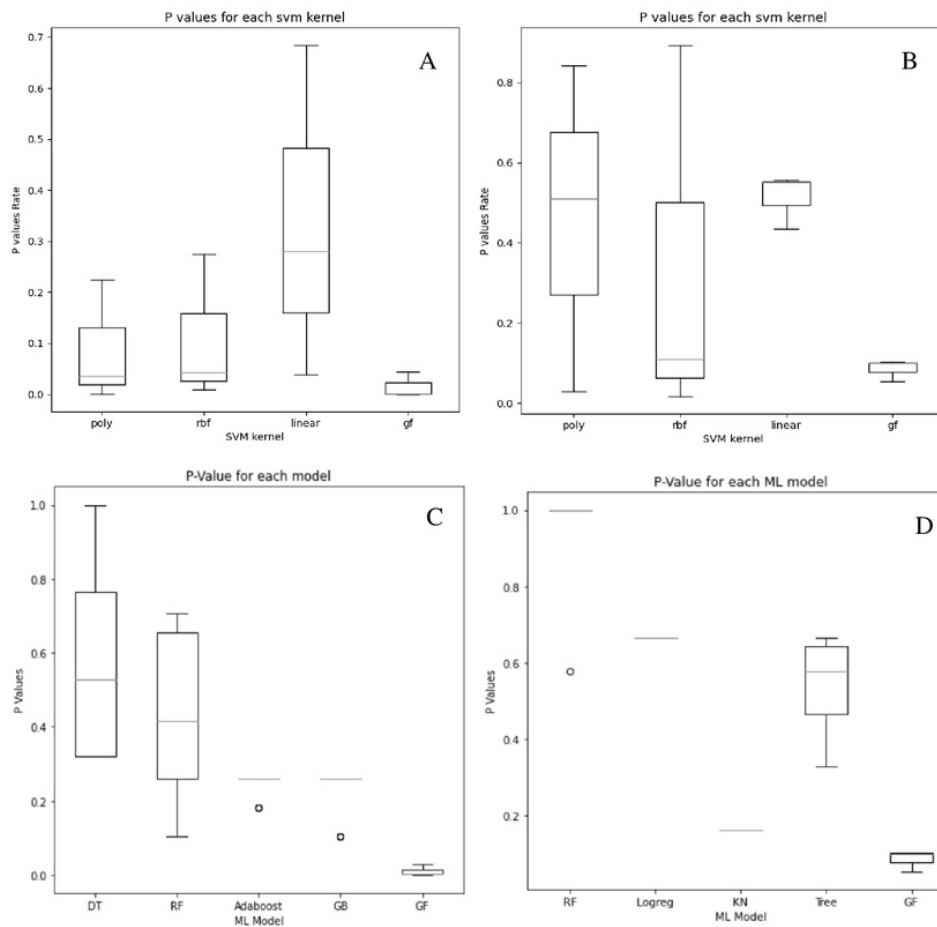



Figure 6

P-values comparisons for prostate cancer (a, b) and breast cancer (c, d) dataset

(a) GF breast cancer vs. (M. A. Mezher, 2022) (b) GF prostate cancer vs. (M. A. Mezher, 2022)

(c) GF breast cancer vs. (Anjelito & Mezher, 2022) (d) GF prostate cancer vs. (Smogomes & Mezher, 2022)

Samples of Prostate Cancer Dataset

Rad.	Text.	Perim.	Area	Smooth	Compact	Symmetry	Fractal_dim.	Diagnosis
23	12	151	954	0.143	0.278	0.242	0.079	M
9	13	133	1326	0.143	0.079	0.181	0.057	B
21	27	130	1203	0.125	0.16	0.207	0.06	M
14	16	78	386	0.07	0.284	0.26	0.097	M
9	19	135	1297	0.141	0.133	0.181	0.059	M
25	25	83	477	0.128	0.17	0.209	0.076	B
16	26	120	1040	0.095	0.109	0.179	0.057	M
15	18	90	578	0.119	0.165	0.22	0.075	M
19	24	88	520	0.127	0.193	0.235	0.074	M
25	11	84	476	0.119	0.24	0.203	0.082	M

Samples of Breast Cancer Dataset

Radius	Texture	Perimeter	Area	Smoothness	Compactness	Concavity
20.57	17.77	132.90	1326.0	0.08	0.08	0.09
19.69	21.25	130.00	1203.0	0.11	0.16	0.20
20.29	14.34	135.10	1297.0	0.10	0.13	0.20
12.45	15.70	82.57	477.10	0.13	0.17	0.16
18.25	19.98	119.60	1040.0	0.09	0.11	0.11
13.71	20.83	90.20	577.90	0.12	0.16	0.09

The traditional Kernel functions

Name	Kernel Function
Linear Kernel	$k(x_i, x_j) = x_i \cdot x_j$
Polynomial	$k(x_i, x_j) = (x_i \cdot x_j + 1)^p$
RBF	$k(x_i, x_j) = exp^{-\gamma(x_i - x_j)^2}$

The symbols used to formulate Kernel functions

Type of symbols	Name	No. of Arity
Operators	'Plus_s'	1
	'Minus_s'	1
	'Multi_s'	1
	'Plus_v'	2
	'Minus_v'	2
Operands	'x'	1
	'y'	1

The encoding/decoding used to formulate Kernel functions

Index	0	1	2	3	4
GF Decoding	'Minus_s',	'Plus_s',	'y'	'y'	'x'
GF Encoding	'1.2'	'3.4'	'0.2'	'0.3'	'0.4'

Comparisons between different Kernel functions and AI models Prostate cancer

Model	Reference	Eliminated Features	Accuracy (%)
Adaboost	(Cijov, 2021)	None	75.8%
KNeighbors	(Gomes, 2021)	['fractal_dimension', 'texture', 'perimeter']	71.0%
SVM (Linear)	GFLibPy	None	88.0%
SVM (RBF)	GFLibPy	None	76.0%
SVM (Polynomial)	GFLibPy	None	80.0%
Random Forest (RF)	(Gomes, 2021)	['fractal_dimension', 'texture', 'perimeter']	86.0%
Decision Tree (DT)	(Gomes, 2021)	['fractal_dimension', 'texture', 'perimeter']	86.9%
Logistic Regression (LR)	(Ribeiro, 2021)	['Area']	90.0%
Proposed GF	Our paper	None	**96.0%**
Artificial Neural Network (ANN)	(Şenol, 2021)	None	80.0%