ISSN: 0974-5823 Vol. 7 No. 6 June, 2022

*International Journal of Mechanical Engineering*

# An Efficient Genetic Boruta(GenBoruta) Algorithm Based Feature Selection on Brain Tumor Dataset

**R.Vinayaga Moorthy[1], Dr.R.Balasubramanian[2]**

Research Scholar[1], Professor[2],

Department of Computer Science & Engineering,

Manonmaniam Sundaranar University, Tirunelveli-627 012

## Abstract

The World Health Organization revealed that the brain tumor is one of the most severe sicknesses since it affects most people, including kids, worldwide. Developing a system to identify brain tumors at a beginning stage would assist in saving the existence of many people. Much exploration has been made around here to develop a system for distinguishing brain tumors; however, this system should be improved, its exactness upgraded. Consequently, feature selection methods are expected to improve the system. The main intention of the feature selection techniques in machine learning (ML) is to select a suitable set of features. Wrapper methods are used to filter. These methods are classified into four categories: forward selection, backward elimination, exhaustive feature selection, and recursive feature elimination. In recent years, brain tumor disease affected more people. Brain tumor disease affects the brain, sometimes sprit into some other parts. Besides, there are 55 features concentrated on, like the image roughness, consistency or energy, and nearby homogeneity removed to show the quality distinction between methods. The goal is to search for the possibility of features that structure a large problem with feature selection techniques, which is resolved using bruta and genetics. Boruta feature selection algorithm based on random forest. In this paper, we introduced a hybrid feature selection technique called GenBoruta. GenBoruta is a hybrid feature selection algorithm for finding all relevant variables. It iteratively eliminates the features which are demonstrated by a measurable test to be less significant than random probes. The proposed techniques performed well compared to existing techniques like forwarding Selection, Backward Elimination, Boruta, and Genetic.

**Keywords:** Forward Selection, Backward Elimination, Recursive Feature Elimination, Genetic, Boruta, GenBoruta

## 1. Introduction

Brain tumors are a very dangerous disease because of their effect on the brain. The feature selection techniques are used to select the particular features. Brain tumor features are a brain tumor attribute that is helpful in the solution, and selecting the most important features for the techniques is called feature selection. Image processing needs human interference.

## 2. Feature selections method

There are two types of feature selection techniques such as supervised techniques and unsupervised techniques. Supervised techniques are classified into three categories: filter methods, wrapper methods, and embedded methods.

The following Table 1 shows different types of feature selection method

| Filter | Wrapper | Embedded |
|---|---|---|
| Non-exclusive arrangement of techniques that don't incorporate a particular Machine Learning (ML) algorithm. | Evaluates a particular ML algorithm to track down optimal features. | Embeds {fix) features during the model structure process. Include determination is finished by noticing every emphasis of the model training stage. |
| A lot quicker contrasted with wrapper techniques as far as time complexity. | High computation time for a<br>dataset with numerous features. | Sits between filter techniques and wrapper techniques as far as time complexity. |
| Less inclined to over-fitting | Highest possibilities of over-fitting since it includes preparing of ML models with various mixes of features. | For the most part, used to lessen over-<br>Fitting by punishing the coefficients of a model being excessively large. |

Table 1 Feature Selection Method

### 2.1Filter Method

The filter method is used to assess the feature selection. This approach is a statistical measure. Feature Selection has become progressively significant for machine learning, data mining, and data analysis. Particularly for high-layered data sets, it is important to filter out the insignificant and redundant features by selecting an appropriate subset of important features to order to over-fitting and tackling the scourge of dimensionality. Concerning data sets from the medical area, including feature selection permits identifying significant features for medical processes of interest [3].

### 2.2 Wrapper Method

Wrapper techniques depend on greedy search algorithms. They evaluate all potential combinations of the features and select the combination that delivers the best outcome for a particular machine learning algorithm. The wrapper method was used to assess the feature subsection. The interaction between the feature subset and regression model in the wrapper method is considered.

### 2.3 Embedded Method

In Embedded Methods, the feature selection algorithm is coordinated as a feature of the learning algorithm. Embedded techniques combine the characteristics of filter and wrapper methods. Algorithms carry it out with their feature selection techniques. A learning algorithm exploits its variable selection process and simultaneously performs feature selection and classification/regression.

In [20], embedded Methods are used to assess the lower computational cost. The embedded method approach is a fast processing method. Embedded methods within algorithms that select features inductively, that is, the classification function and selection of subsets of features are jointly learned. Generally, these methods optimize objective functions that point to classification accuracy and fine-tune the use of additional features simultaneously. In [24], embedded methods assess the feature subset by inserting the feature selection in the cycle of classifier construction. These feature selection methods are intended to perform the same assignments with different design methodologies. Each has its benefits and weaknesses in various perspectives, such as model complexity, computational effectiveness, and time productivity.

## 3. Literature Review

### 3.1 Wrapper Method

The wrapper method approach is decision making. An optimization approach based on hybrid wrapper-filter feature selection and ensemble classification are used to minimize the impact of an imbalanced brain tumor dataset. It is used to assess the accuracy [21]. In [10] order to identify the most suitable subset of features and assess each subset's accuracy in predicting the target, it uses a learning algorithm.

### 3.2 Forward Selection

In [8], the forward selection method process begins with no shifting & adds separately. Each adds one variable that reduces the error until some further variable addition doesn't notably reduce the error. The forward selection method is a simple ranking-based feature selection. The forward feature selection method is restricted to certain features with huge contribution rates; be that as it may, in a few practical cases, the optimal combination of features is typically not made out of these features. Subsequently, the feature subset got by the forward feature selection technique is probably going to prompt lower-order accuracy. In [16], the forward selection (FS) algorithm adds a new variable to the existing model, starting from a null model with no caveats, and at each step based on some criterion such as the decrease in the residual sum of squares (RSS).

### 3.3 Backward Elimination

S.Vanaja et al. [19] Select a feature using the backward eliminations method to select and eliminate the appropriate feature. The feature selection is a three-stage process in a particular search, evaluation, and stop. The wrapper method uses backward elimination to eliminate the irrelevant features from the subset. In [15], backward elimination can detect models where all the elements are genuine. If no noise variables are present, the model is found to be valid.   A second unfounded assumption is that classic least squares estimation results can be applied to the final model.

### 3.4 Recursive Feature Elimination

B.Sathees Kumar [4] Applying the recursive feature elimination method, one can identify the important features before even conducting the classification task. The technique is based on its iteration count and size. It uses the random forest principle and k-fold validation. In [23], a recursive feature elimination algorithm is used to remove a single feature at a time. It is used to solve the multiclass classification. The weight vector for recursive removal is computed on each subsample and put into every individual classifier.

### 3.5 Genetic Algorithm

Sourabh Katoch et al. [22] select a feature using a genetic algorithm used to assess the great features. In addition to providing information about each component of GAs, this paper provides a source of recent research on GAs. Researchers will be encouraged to understand GA fundamentals and use the knowledge in their research. In addition, it expands the range of possible users.GA consists of encoding, selection, crossover, and mutation. In [13], Genetic has been efficiently used in the feature selection problem to redact high-dimensional datasets. One of the disadvantages of this technique is that it doesn't think about the associations among the features when selecting the final features. Accordingly, the probability of choosing a subset with redundancy will increment.

L. Haldurai et al. [11] Selection operation involve selecting elite individuals that can produce offspring from the current population. Individuals are judged on their fitness values to determine whether they are elitists. In [1], Crossover in GA involves exchanging information between good solutions to form new and hopefully better solutions. Crossover operators facilitate rapid convergence towards a good solution. The crossover consists of a single point, two-point, k-point, uniform, partially matched, order, precedence preserving crossover, shuffle, reduced surrogate, and cycle. In [6], the mutation is the process of changing (growing) chromosome genomes, resulting in the flipping of bits (genes) of chromosomes. Mutation consists of displacement, simple inversion, and scramble mutation.

### 3.6 Boruta Algorithm

In [25] Boruta algorithm is intended to be a wrapper algorithm that performs robust, statistically relevant feature selection. To select the optimal feature groups based on HCR and DLR features, they applied a random forest (RF)-based Boruta algorithm. Johannes Haubold et al. [9] select a feature using the Boruta algorithm to reduce the noise added by redundant features; a subset of features with the highest predictive power needs to be selected from among the extensive features. To achieve this, the Boruta algorithm used a wrapper method combining Gradient Boosting XGBoost. In [17], the Boruta package empowers the user to select the most important features of an information system based on unbiased criteria, while manual methods are more likely to render errors. The boruta wrapper algorithm is used to select features in an information system unbiased to important and other attributes. Supervised learning is essential to train raw data after feature engineering.

### 3.7 Hybrid Genetic Algorithm

Ahmed Kharrat and Mahmoud Neji [2] proposed a hybrid SA-GA algorithm for selecting the optimal feature subsets from countless features capable of acting as an instrument for PC-supported findings in the separation of brain tumors on magnetic resonance images. Bain Khusnul Khotimah et al. [5] select a feature using a hybrid algorithm that produces an effective and less feature set and the weakest feature enhancements. Since the NB classification results in higher accuracy in several heterogeneous datasets. GA worked well in finding optimal feature values, achieving imputation results with an error rate10%.

Joans and Sandhiya [18] proposed a genetic Algorithm consisting of a feature selection technique to classify the MRI images using the random forest as a classifier. For a mixture medical images recovery system, a GA approach is introduced to choose a dimensionality decreased set of features. This framework has been in three stages. In the first stage, three particular algorithms are utilized to extricate the imperative elements from the pictures. The algorithm conceived for extracting the features is the Texton-based intrinsic pattern extraction algorithm, contour gradient extraction algorithm, and shift-invariant feature transformation algorithm. The second stage to distinguish the potential element vector GA-based highlight choice is done, utilizing a hybrid approach of "Branch and Bound Algorithm" and "Fake Bee Colony Algorithm" utilizing the brain tumor, thyroid images, and breast cancer. The Chi-Square distance estimation is utilized to evaluate the likeness between inquiry pictures and data set images. A fitness function with regard Minimum depiction length rule was utilized as a beginning necessity for the GA algorithm. In the third stage, the assorted thickness-based pertinence criticism technique is utilized to work on the presentation of the mixture content-based clinical picture recovery framework. The term mixture is utilized as this framework can recover any sort of clinical picture like brain tumor, breast cancer, thyroid disease, lung cancer, etc. [7].

## 4. Methodology

Characterizing pictorial information is expected to distinguish significant features present in pictures that lead to the arrangement. Such features could be gathered into basic and complex features. The selection of features can also improve classifier performance. In machine learning and measurements, highlight determination, otherwise called variable choice, characteristic choice, or variable subset choice, is the most common way of choosing a subset of pertinent elements (factors, indicators) for model development.

The wrapper method utilizes a prescient model to score highlight subsets. Every new subset is utilized to prepare a model, which is tried on a hold-out set. As the wrapper method train another model for every subset, they are computationally escalated yet ordinarily give the best performing highlight set for that specific sort of model or regular issue.

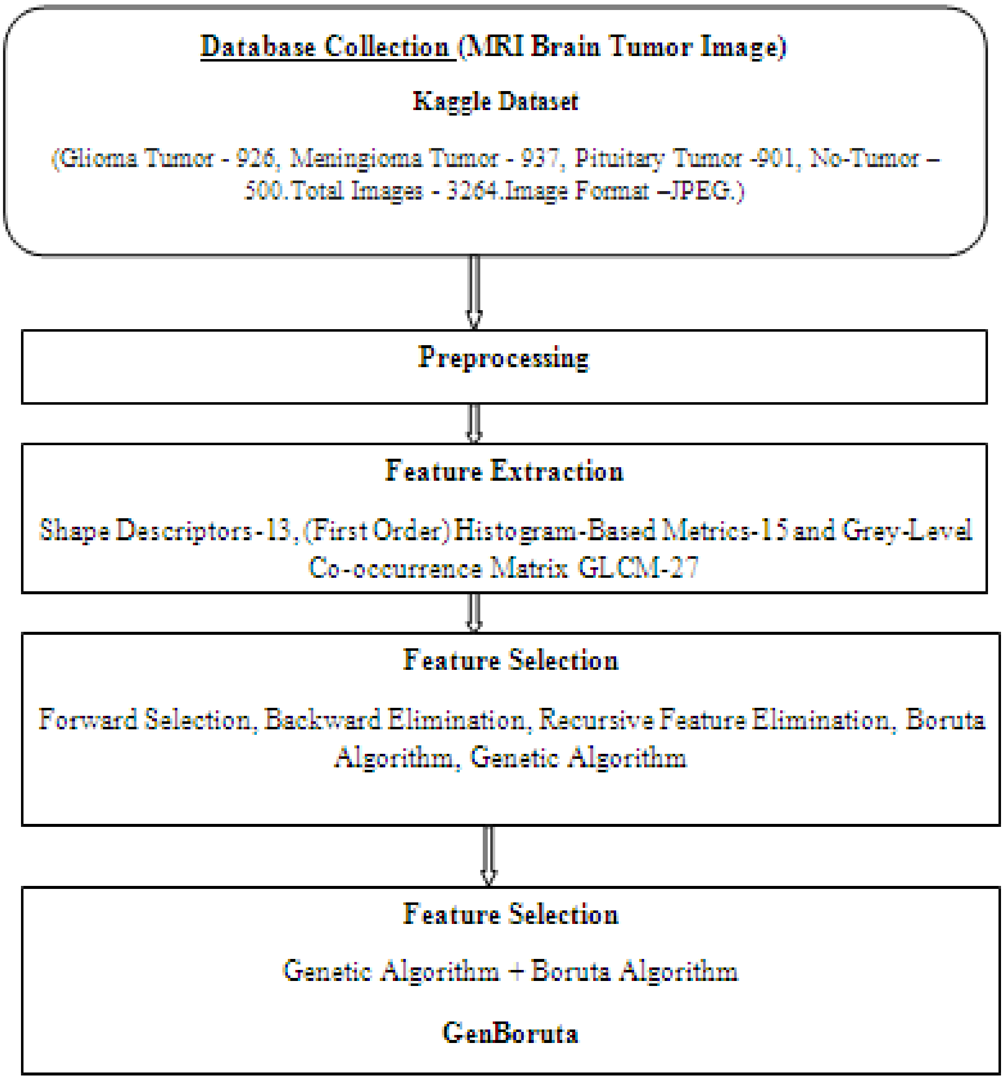

**Fig 1- Feature Selection Proposed Architecture**

### 4.1 Feature Extraction

A start to finish guide on the most proficient method to diminish a dataset dimensionality utilizing Feature Extraction Techniques, Such as Shape Descriptors, (First Order) Histogram-Based Metrics, and Grey-level co-occurrence matrix GLCM.

The following Tables 2, 3, and 4 represent types of features and formula

| Shape Descriptors | | |
|---|---|---|
| 1. | Mesh Volume | $$V_i = \frac{Oa_i \cdot (Ob_i \times Oc_i)}{6} \quad (1)$$ $$V = \sum_{i=1}^{N_f} V_i \quad (2)$$ |

| | | |
|---|---|---|
| 2. | Voxel Volume | $V_{voxel} = \sum_{k=1}^{N_v} V_k$ |
| 3. | Surface Area | $A_i = \frac{1}{2}\lvert \mathbf{a}_i\mathbf{b}_i \times \mathbf{a}_i\mathbf{c}_i \rvert \ (1)$<br>$A = \sum_{i=1}^{N_f} A_i \ (2)$ |
| 4. | Surface Area to Volume ratio | $surface\ to\ volume\ ratio = \frac{A}{V}$ |
| 5. | Sphericity | $sphericity = \frac{\sqrt[3]{36\pi V^2}}{A}$ |
| 6. | Compactness 1 | $compactness\ 1 = \frac{V}{\sqrt{\pi A^3}}$ |
| 7. | Compactness 2 | $compactness\ 2 = 36\pi\frac{V^2}{A^3}$ |
| 8. | Spherical Disproportion | $spherical\ disproportion = \frac{A}{4\pi R^2} = \frac{A}{\sqrt[3]{36\pi V^2}}$ |
| 9. | Major Axis Length | $major\ axis = 4\sqrt{\lambda_{major}}$ |
| 10. | Minor Axis Length | $minor\ axis = 4\sqrt{\lambda_{minor}}$ |
| 11. | Least Axis Length | $least\ axis = 4\sqrt{\lambda_{least}}$ |
| 12. | Elongation | $elongation = \sqrt{\frac{\lambda_{minor}}{\lambda_{major}}}$ |
| 13. | Flatness | $flatness = \sqrt{\frac{\lambda_{least}}{\lambda_{major}}}$ |

Table 2 Shape Descriptors features and formula

| (First Order) Histogram-Based Metrics | | |
|---|---|---|
| 1 | Energy | $energy = \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$ |
| 2. | Total Energy | $total\ energy = V_{voxel} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2$ |
| 3 | Entropy | $entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \big(p(i) + \epsilon\big)$ |
| 4 | Max Intensity | $minimum = \min(\mathbf{X})$ |
| 5 | Min Intensity | $maximum = \max(\mathbf{X})$ |
| 6 | Mean Value | $mean = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{X}(i)$ |
| 7 | Mean absolute deviation | $MAD = \frac{1}{N_p} \sum_{i=1}^{N_p} \lvert \mathbf{X}(i) - \bar{X} \rvert$ |
| 8 | Robust Mean Absolute Deviation (rMAD) | $rMAD = \frac{1}{N_{10-90}} \sum_{i=1}^{N_{10-90}} \lvert \mathbf{X}_{10-90}(i) - \bar{X}_{10-90} \rvert$ |
| 9 | Range | $range = \max(\mathbf{X}) - \min(\mathbf{X})$ |
| 10 | Root mean square (RMS) | $RMS = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2}$ |
| 11 | Standard deviation | $standard\ deviation = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (\mathbf{X}(i) - \bar{X})^2}$ |
| 12 | Uniformity | $uniformity = \sum_{i=1}^{N_g} p(i)^2$ |

| | | |
|---|---|---|
| 13 | Variance | $variance = \frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) - \bar{X})^2$ |
| 14 | Skewness | $skewness = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) - \bar{X})^3}{\left(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) - \bar{X})^2}\right)^3}$ |
| 15 | Kurtosis | $kurtosis = \frac{\mu_4}{\sigma^4} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) - \bar{X})^4}{\left(\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) - \bar{X})^2\right)^2}$ |

Table 3 (First Order) Histogram-Based Metrics features and formula

| Grey-level co-occurrence matrix GLCM | | |
|---|---|---|
| 1. | Autocorrelation | $autocorrelation = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)ij$ |
| 2. | Joint Average | $joint\ average = \mu_x = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)i$ |
| 3. | Cluster Prominence | $cluster\ prominence = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(i + j - \mu_x - \mu_y\right)^4 p(i,j)$ |
| 4. | Cluster shade | $cluster\ shade = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(i + j - \mu_x - \mu_y\right)^3 p(i,j)$ |
| 5. | Cluster Tendency | $cluster\ tendency = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(i + j - \mu_x - \mu_y\right)^2 p(i,j)$ |
| 6. | Contrast | $contrast = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i - j)^2 p(i,j)$ |

| | | |
|---|---|---|
| 7. | Correlation | $\textit{correlation} = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)ij - \mu_x \mu_y}{\sigma_x(i)\sigma_y(j)}$ |
| 8. | Difference Average | $\textit{difference average} = \sum_{k=0}^{N_g-1} k p_{x-y}(k)$ |
| 9. | Difference entropy | $\textit{difference entropy} = \sum_{k=0}^{N_g-1} p_{x-y}(k) \log_2 \left(p_{x-y}(k) + \epsilon\right)$ |
| 10. | Difference variance | $\textit{difference variance} = \sum_{k=0}^{N_g-1} (k - DA)^2 p_{x-y}(k)$ |
| 11. | Dissimilarity | $\textit{dissimilarity} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \lvert i - j \rvert p(i,j)$ |
| 12. | Joint Energy | $\textit{joint energy} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left(p(i,j)\right)^2$ |
| 13. | Joint Entropy | $\textit{joint entropy} = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \log_2 \left(p(i,j) + \epsilon\right)$ |
| 14. | Homogeneity | $\textit{homogeneity} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{p(i,j)}{1 + \lvert i - j \rvert}$ |
| 15. | Information measure of correlation 1 | $\textit{IMC 1} = \frac{HXY - HXY1}{\max\{HX, HY\}}$ |
| 16. | Information measure of correlation 2 | $\textit{IMC 2} = \sqrt{1 - e^{-2(HXY2 - HXY)}}$ |
| 17. | Inverse difference Moment | $IDM = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1 + k^2}$ |

| | | |
|---|---|---|
| 18. | Maximum Correlation Coefficient | $MCC = \sqrt{\text{second largest eigenvalue of Q}}$<br>$Q(i,j) = \sum_{k=0}^{N_g} \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$ |
| 19. | Inverse Difference Moment Normalized (IDMN) | $IDMN = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\left(\frac{k^2}{N_g^2}\right)}$ |
| 20. | Inverse Difference (ID) | $ID = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+k}$ |
| 21. | Inverse Difference Normalized (IDN) | $IDN = \sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+\left(\frac{k}{N_g}\right)}$ |
| 22. | Inverse Variance | $\mathit{inverse\ variance} = \sum_{k=1}^{N_g-1} \frac{p_{x-y}(k)}{k^2}$ |
| 23. | Maximum Probability | $\mathit{maximum\ probability} = \max\left(p(i,j)\right)$ |
| 24. | Sum Average | $\mathit{sum\ average} = \sum_{k=2}^{2N_g} p_{x+y}(k)k$ |
| 25. | Sum variance | $\mathit{sum\ variance} = \sum_{k=2}^{2N_g} (k-SA)^2 p_{x+y}(k)$ |
| 26. | Sum entropy | $\mathit{sum\ entropy} = \sum_{k=2}^{2N_g} p_{x+y}(k)\log_2\left(p_{x+y}(k)+\epsilon\right)$ |
| 27. | Sum of squares | $\mathit{sum\ squares} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i-\mu_x)^2 p(i,j)$ |

Table 4 GLCM features and formula

### 4.2 Forward Selection Algorithm

Forward selection is an iterative technique wherein we initialize with no elements in the model. In every cycle, we continue to add a component that best works on our model till an option of another variable doesn't work with the presentation of the model. To execute an optimization, it is important to lessen the levels of opportunity for the framework in request to make the issue viable with the most widely recognized optimization.

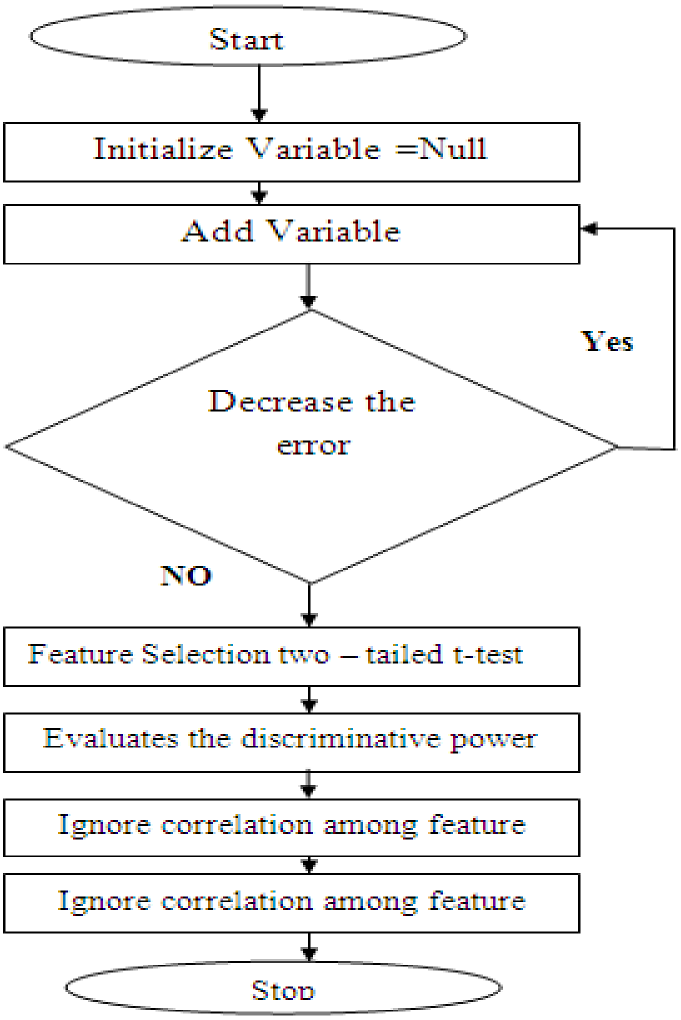

**Fig 2- Forward Selection Architecture**

### 4.3 Backward Elimination Algorithm

As an iterative approach, backward elimination is also an option, but it is the opposite to forward selection. Machine learning models are built using backward elimination as a feature selection technique. The process starts by considering all the features and then removing the least significant features. It continues until removing the features and does not improve the model's performance.

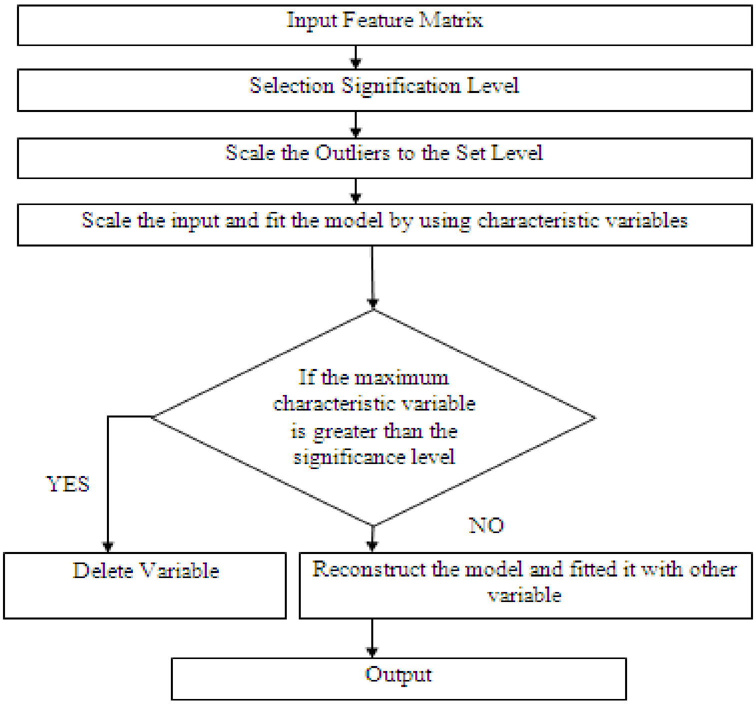

**Fig 3- Backward Elimination Architecture**

### 4.4 Recursive Feature Elimination (RFE)

RFE is an FS technique that fits a model and eliminates the most fragile feature (or highlight) until the predefined number of features is reached. Recursive Feature Elimination with Cross-Validation shows the features which are significant with significant positioning. This empowers us to fabricate the model with ideal aspects.

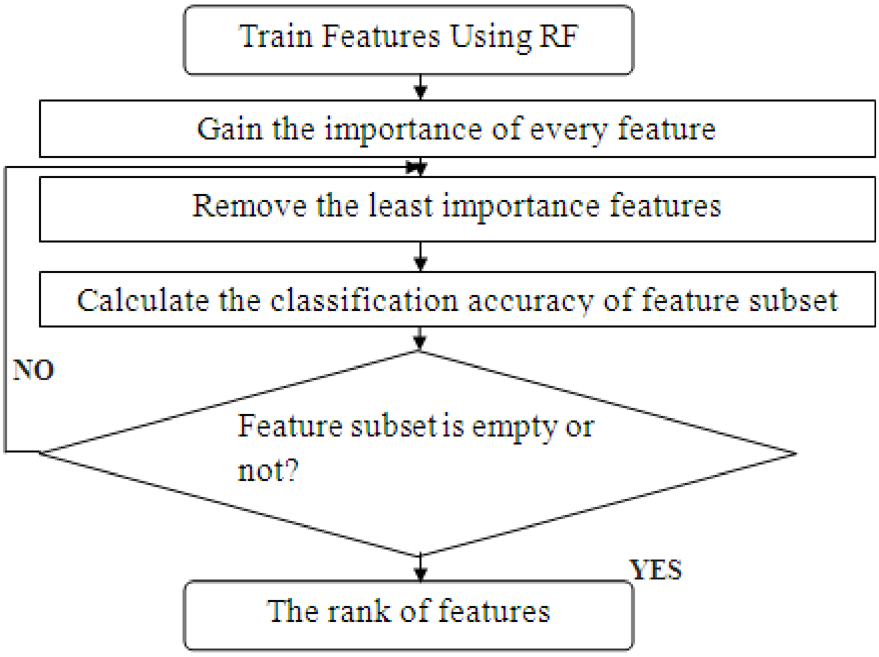

**Fig 4- Recursive Feature Elimination Architecture**

### 4.5 Genetic Algorithm

GA is a heuristic hunt technique utilized in artificial intelligence (AI) and figuring. It is utilized for tracking down enhanced answers for search issues in view of the hypothesis of normal determination and developmental science. GA is superb for looking through the enormous and complex dataset.

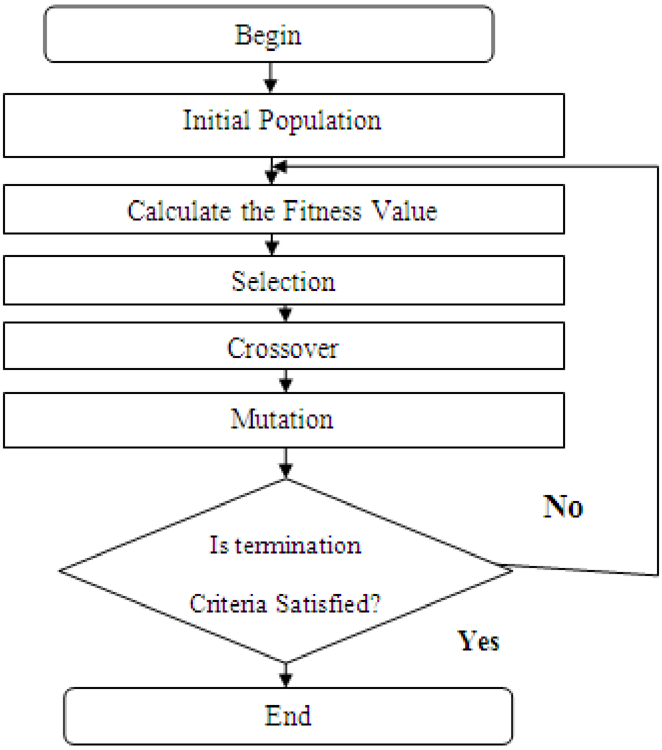

**Fig 5- Genetic Algorithm Architecture**

### 4.6 Boruta Algorithm

Perform rearranging of predictors qualities, go along with them with the first predictors, and afterward construct a random forest on the consolidated dataset. Then, at that point, make the correlation of unique factors with the randomized factors to quantify variable significance. Just factors having higher significance than that of the randomized factors are viewed as significant.

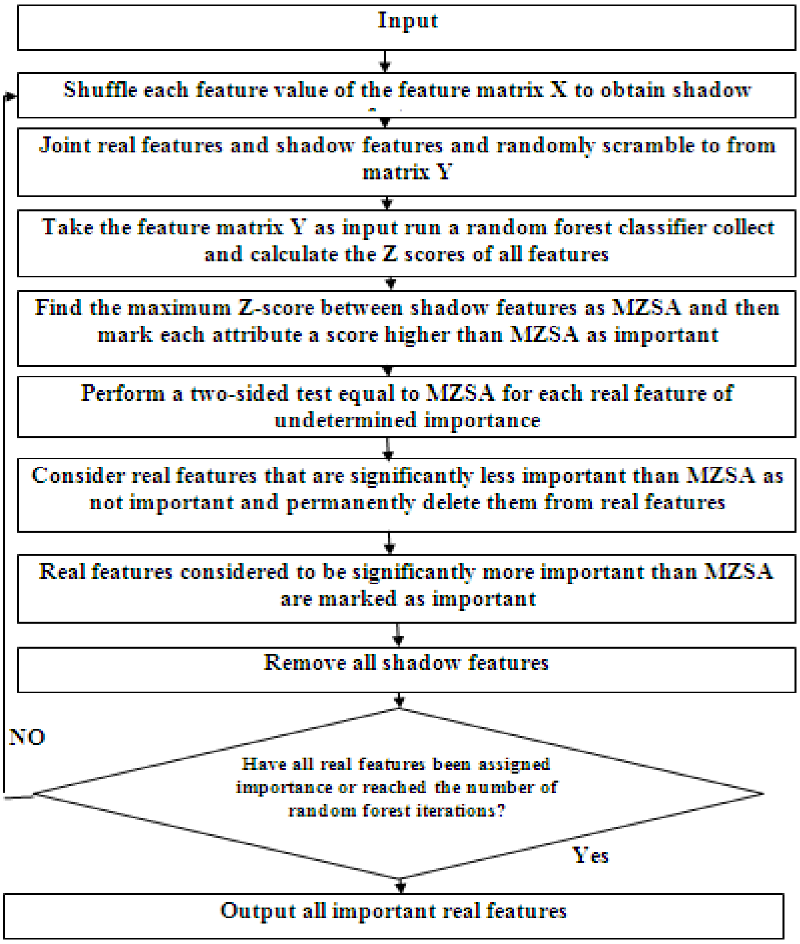

**Fig 6- Boruta Algorithm Architecture**

### 4.7 Proposed (GenBoruta) Feature Selection Technique:

The techniques proposed in this work depend on making a hybrid model that joins a GA and boruta, fully intent on grouping tests before selecting the small number of important variables. The proposed hybrid technique is used to overcome the disadvantage of each one.

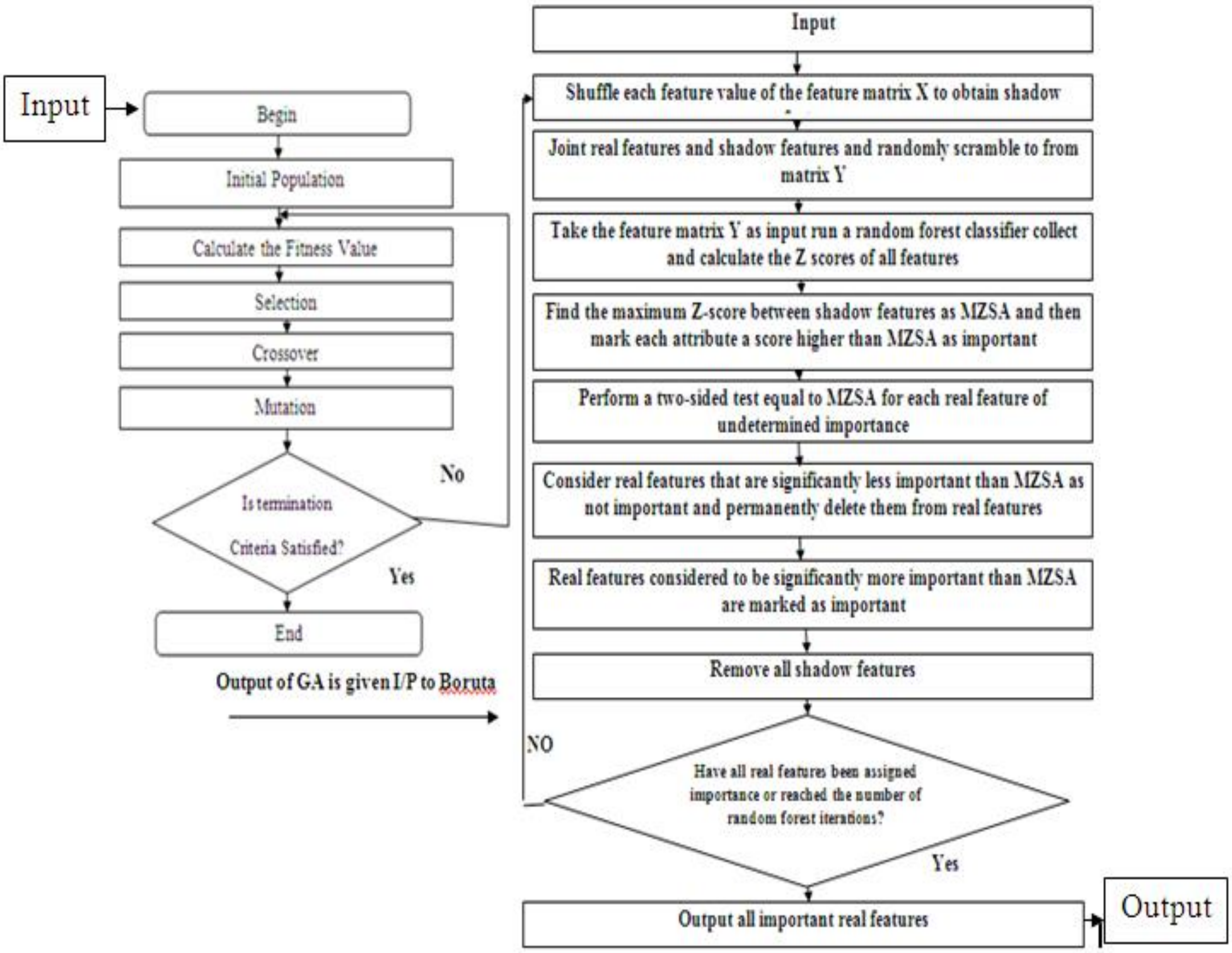

**Fig 7- Architecture of proposed GenBoruta Feature Selection Technique**

## 5. Experimental Setup

### 5.1. Dataset Description

A brain tumor dataset with imitated acoustic features has been utilized for training and assessing our approach. Dataset is from the Kaggle website. The dataset comprises 3264 brain MRI images classified into four classes: glioma, meningioma, pituitary, and no tumor.

The following Table 5 represents the types of features.

| **S.No** | **Attribute** | **S.No** | **Attribute** | **S.No** | **Attribute** | **S.No** | **Attribute** | **S.No** | **Attribute** |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Mesh Volume | 12 | Elongation | 23 | Root mean square (RMS) | 34 | Contrast | 45 | Inverse difference Moment |
| 2 | Voxel Volume | 13 | Flatness | 24 | Standard deviation | 35 | Correlation | 46 | Maximum probability coefficient |
| 3 | Surface Area | 14 | Energy | 25 | Uniformity | 36 | Difference Average | 47 | Inverse Difference Moment Normalized (IDMN) |
| 4 | Surface Area to Volume ratio | 15 | Total Energy | 26 | Variance | 37 | Difference entropy | 48 | Inverse Difference (ID) |
| 5 | Sphericity | 16 | Entropy | 27 | Skewness | 38 | Difference variance | 49 | Inverse Difference Normalized (IDN) |
| 6 | Compactness 1 | 17 | Max Intensity | 28 | Kurtosis | 39 | Dissimilarity | 50 | Inverse Variance |
| 7 | Compactness 2 | 18 | Min Intensity | 29 | Autocorrelation | 40 | Joint Energy | 51 | Maximum Probability |
| 8 | Spherical Disproportion | 19 | Mean Value | 30 | Joint Average | 41 | Joint Entropy | 52 | Sum Average |
| 9 | Major Axis Length | 20 | Mean absolute deviation | 31 | Cluster Prominence | 42 | Homogeneity | 53 | Sum variance |
| 10 | Minor Axis Length | 21 | Robust Mean Absolute Deviation (rMAD) | 32 | Cluster shade | 43 | Information measure of correlation 1 | 54 | Sum entropy |
| 11 | Least Axis Length | 22 | Range | 33 | Cluster Tendency | 44 | Information measure of correlation 2 | 55 | Sum of squares |

Table 5 Types of features

### 5.2 Results

Performance is evaluated based on features extracted from Shape Descriptors, (First Order) Histogram-Based Metrics, and Grey-level co-occurrence matrix (GLCM). This part shows the experimental results of the proposed GenBoruta framework: 1) this method enables the selection of more essential and relevant functions than existing methods of detecting brain tumors.2) Compared to existing methods, the proposed method produces better results. The following Table 6 represents selected features.

<table>
<tr><th>S.No</th><th>Features Selection Algorithms</th><th colspan="2">Features</th></tr>
<tr><td>1</td><td>Forward Selection</td><td colspan="2">25 features are Surface Area to Volume ratio, Least Axis Length, Robust Mean Absolute Deviation (rMAD), Cluster Tendency, Joint Average, Joint Entropy, Joint Energy, Kurtosis, Skewness, Uniformity, Difference entropy, Contrast, Standard deviation, Inverse Difference Normalized (IDN), Difference variance, Sum Average, Sum of squares, Least Axis Length, Minor Axis Length, Sphericity, Mesh Volume, Maximum correlation coefficient, Maximum Probability, Mean Value, Min Intensity.</td></tr>
<tr><td>2</td><td>Backward Elimination</td><td colspan="2">22 features are Mesh Volume, Surface Area, Sphericity, Spherical Disproportion, Minor Axis Length, Elongation, Energy, Total Energy, Max Intensity, Mean Value, Robust Mean Absolute Deviation (rMAD), Root mean square (RMS), Uniformity, Skewness, Autocorrelation, Cluster Prominence, Cluster Tendency, Correlation, Difference entropy, Dissimilarity, Joint Entropy, Information measure of correlation 1.</td></tr>
<tr><td>3</td><td>Recursive Feature Elimination</td><td colspan="2">21Features are Mesh Volume, Voxel Volume, Surface Area, Surface Area to Volume ratio, Sphericity, Uniformity, Variance, Skewness, Kurtosis, Autocorrelation, Maximum Probability, Sum Average, Sum variance, Sum entropy, Sum of squares, Max Intensity, Min Intensity, Mean Value, Mean absolute deviation, Robust Mean Absolute Deviation (rMAD), Cluster shade.</td></tr>
<tr><td>4</td><td>Boruta</td><td colspan="2">14 Features are Voxel Volume, Surface Area to Volume ratio, Uniformity, Variance, Kurtosis, Autocorrelation, Maximum Probability, Sum Average, Sum variance, Sum of squares, Max Intensity, Min Intensity, Mean Value, Cluster shade.</td></tr>
<tr><td>5</td><td>Genetic</td><td colspan="2">15 Features are Mesh Volume, Surface Area, Surface Area to Volume ratio, Sphericity, Uniformity, Variance, Skewness, Kurtosis, Autocorrelation, Maximum Probability, Sum Average, Sum entropy, Max Intensity, Mean Value, Robust Mean Absolute Deviation (rMAD).</td></tr>
<tr><td rowspan="7"><b>6</b></td><td rowspan="7"><b>Genboruta</b></td><td colspan="2"><b>12 Features are</b></td></tr>
<tr><td><b>Surface Area</b></td><td><b>Flatness</b></td></tr>
<tr><td><b>Skewness</b></td><td><b>Uniformity</b></td></tr>
<tr><td><b>Contrast</b></td><td><b>Correlation</b></td></tr>
<tr><td><b>Range</b></td><td><b>Cluster shade</b></td></tr>
<tr><td><b>Dissimilarity</b></td><td><b>Maximum Probability</b></td></tr>
<tr><td><b>Variance</b></td><td><b>Max Intensity</b></td></tr>
</table>

Table 6 Selected features

Table 7 Performance Analysis of Various FS techniques and proposed technique

| Feature Section Techniques | Feature Selection | Accuracy | Error Rate |
|---|---|---|---|
| Forward Selection | 25/55 | 94 | 11 |
| Backward Elimination | 22/55 | 96 | 9 |
| Recursive Feature Elimination | 21/55 | 95.5 | 8.2 |
| Boruta | 14/55 | 96 | 9 |
| Genetic | 15/55 | 96.5 | 8 |
| **Proposed (GenBoruta)** | **12/55** | **97.5** | **7** |

Table 7 Feature selection performance analysis

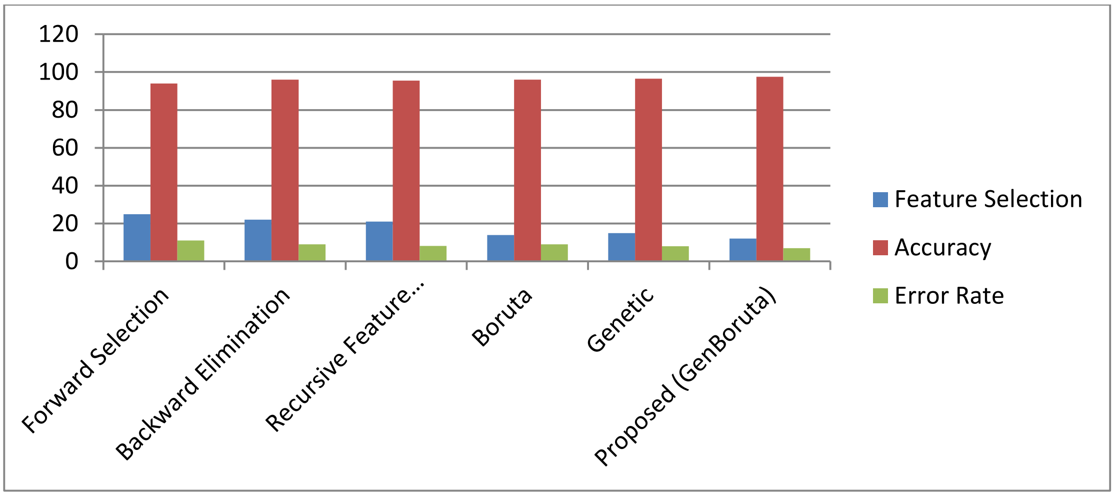

Fig 8- Feature Selection Graphical Representation

## 6. Conclusion

The proposed hybrid algorithm combines the Genetic algorithm and the Boruta algorithm. A hybrid algorithm Genboruta has several advantages from the combination of benefits of existing algorithms gave to the choice of optimal feature subsets from a small number of features. To select the most appropriate features from the feature extraction data, the feature selection method is applied. Crucial features are specified using Genboruta. Genboruta obtained the best performance among these methods and overcame all other methods. So we limit the error rate as well as increase the accuracy. The performance of the proposed approach gives a significant improvement against four closely related techniques, accomplishing an accuracy of 97.5% and an error rate7. Our approach proved its efficiency in feature selection on the brain tumor dataset. For the implementation of the study, Jupiter notebooks version 6.3.0 was used, and Python for coding. Our model is chosen because of its high predictive accuracy.

## Reference

1. Ahmed Kharrat and Mahmoud Neji, "A Hybrid Feature Selection for MRI Brain Tumor Classification", In International Conference on Innovations in Bio-Inspired Computing and Applications, Springer, pp. 329-338, March 2018.
2. Ahmed Kharrat, and Mahmoud Neji, “Feature selection based on hybrid optimization for magnetic resonance imaging brain tumor classification and segmentation”, Applied Medical Informatics, Vol. 41, Issue. 1, pp. 9-23, March 2019.
3. Andrea Bommert, Xudong Sun, BerndBischl, Jörg Rahnenführer, Michel Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data", Vol. 143, pp.1-22, March 2020.
4. B.Sathees Kumar, "Identification and Classification of Brain Tumor Images Using Efficient Classifier", International Journal of Engineering and Advanced Technology, Vol-8, Issue-6, pp. 3677 - 3683, August 2019.
5. Bain Khusnul Khotimah, Miswanto Miswanto, Herry Suprajitno, “Optimization of Feature Selection Using Genetic Algorithm in Naïve Bayes Classification for Incomplete Data”, International Journal of Intelligent Engineering and Systems, Vol.13, Issue.1, pp.334-343, December 2020.
6. Dr. S. Mary Joans, J. Sandhiya, "A Genetic Algorithm Based Feature Selection for Classification of Brain MRI Scan Images Using Random Forest Classifier", International Journal of Advanced Engineering Research and Science, Vol. 4, Issue-5, pp. 124- 130, May 2017.
7. G.Nagarajana, R.I.Minub , B Muthukumar c, V.Vedanarayanan d & S.D.Sundarsinghe, “Hybrid Genetic Algorithm for Medical Image Feature Extraction and selection”, Procedia Computer Science, vol. 85,pp. 455 – 462,2016.
8. Jianli Ding, Liyang Fu, “A Hybrid Feature Selection Algorithm Based on Information Gain and Sequential Forward Floating Search”, Journal of Intelligent Computing Volume 9, Issue 3, pp. 93-103, September 2018.
9. Johannes Haubold, René Hosch, Vicky Parmar, Martin Glas, Nika Guberina,Onofrio Antonio Catalano, Daniela Pierscianek, Karsten Wrede , Cornelius Deuschl , Michael Forsting , Felix Nensa , Nils Flaschel and Lale Umutlu , “Fully Automated MR Based Virtual Biopsy of Cerebral Gliomas”, cancers, MDPI, Vol. 13,Issue 24, pp. 1-13, December 2021.
10. Katkoori Arun Kumar1, Ravi Boda2, Dr. Praveen Kumar3, Dr. Vijendar Amgothu, "Feature Selection using Multi-Verse Optimization for Brain Tumor Classification", Annals of the Romanian Society for Cell Biology, Vol. 25, Issue 6, pp. 3970 – 3982, May 2021.
11. L. Haldurai1, T. Madhubala and R. Rajalakshmi “A Study on Genetic Algorithm and its Applications”, International Journal of Computer Sciences and Engineering, Vol. 4, Issue 10, pp. 139-143, Oct 2016.
12. Maryam Bahojb Imani, Mohammad Reza Keyvanpour, and Reza Azmi, “A Novel Embedded Feature Selection Method: A Comparative Study In The Application Of Text Categorization”, Applied Artificial Intelligence, Vol. 27, pp.408–427, 2013.
13. Mehrdad Rostami, Kamal Berahmand, and Saman Forouzandeh, “A novel community detection based genetic algorithm for feature selection” Journal of Big Data, Vol. 8, Issue 2, pp. 2-27, Dec 2021.
14. Methaq Kadhum, Saher Manaseer, and Abdel Latif Abu Dalhoum, "Evaluation Feature Selection Technique on Classification by Using Evolutionary ELM Wrapper Method with Features Priorities", Journal of Advances in Information Technology Vol. 12, issue. 1, pp-21-28, February 2021.
15. Milan Bašta, "Properties of Backward Elimination and Forward Selection In Linear Regression", International Days of Statistics and Economics, Prague, pp.114-124, September 2018.
16. Naveen Naidu Narisetty, “Bayesian model selection for high-dimensional data” In Handbook of Statistics, Elsevier, Vol-43, pp. 207-248, 2020.
17. Nazmun Nahar, Ferdous Ara, Md. Arif Istiek Neloy, Anik Biswas, Mohammad Shahadat Hossain, and Karl Andersson, "Feature Selection Based Machine Learning to Improve Prediction of Parkinson Disease", In International Conference on Brain Informatics, Springer, pp. 496-508, September 2021.
18. Nilesh Bhaskarrao Bahadure,1 Arun Kumar Ray,1 and Har Pal Thethi, "Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM", International Journal of Biomedical Imaging, Vol. 2017, pp.1-12, March 2017.
19. S.Vanaja, K.Ramesh Kumar, “Analysis of Feature Selection Algorithms on Classification: A Survey", International Journal of Computer Applications, Volume 96– No.17, pp.0975 – 8887, June 2014.

20. Samaneh Liaghat and Eghbal G. Mansoori, "Filter-based unsupervised feature selection using Hilbert–Schmidt independence criterion", International Journal of Machine Learning and Cybernetics, Vol-10, Issue-9, pp. 2313-2328, August 2018.
21. Shamsul Huda, John Yearwood, Herbert F. Jelinek, Mohammad Mehedi Hassan, Giancarlo Fortino, Michael Buckland "A Hybrid Feature Selection with Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis", IEEE ACCESS, Vol. 4, pp.1-13, January 2017.
22. Sourabh Katoch, Sumit Singh Chauhan1 & Vijay Kumar, "A review on genetic algorithm: past, present, and future", Multimedia Tools and Applications, Springer, Vol. 80, Issue 5, pp. 8091–8126, October 2020.
23. V. Panca and Z. Rustama, "Application of machine learning on brain cancer multiclass classification", AIP Conference Proceedings, Vol. 1862, Issue 1, pp. 030133-030139, July 2017.
24. Yee Ching Saw, Zeratul Izzah Mohd Yusoh, Azah Kamilah Muda and Ajith Abraham, "Ensemble Filter-Embedded Feature Ranking Technique (FEFR) for 3D ATS Drug Molecular Structure", International Journal of Computer Information Systems and Industrial Management Application, Vol. 9, pp. 124-134, 2017.
25. Zhiyuan Liu, Zekun Jiang, Li Meng, Jun Yang, Ying Liu, Yingying Zhang, Haiqin Peng, Jiahui Li, Gang Xiao, Zijian Zhang, and Rongrong Zhou, "Handcrafted and Deep Learning-Based Radiomic Models Can Distinguish GBM from Brain Metastasis", Journal of Oncology, Hindawi, Vol. 2021, pp.1-10, June 2021.