University of Ibadan
Journal of Science and Logics in ICT Research

Framework for a Stimulated Predictive Distributed Learning Method

[1]Igodan E.C. and [2]Iyekowa J.O.

[1]*Department of Computer Science, University of Benin*
[2]*University of Sunderland, UK*
[1]*charles.igodan@uniben.edu*,[2]*johniyekowa@gmail.com*

Abstract

Due to the intrinsic properties of high-dimensional microarray datasets, most feature selection approaches do not scale well, which makes these models inapplicable and impairs the performance of most classifiers. This study used data complexity and stability measures to maintain class distribution and reduce features variability while proposing a novel predictive distributed FS model through horizontal partitioning. Brain tumour microarray benchmark was employed for implementation. Six classifiers as well as feature selection methods were employed along with their ensemble learning techniques. The study observed the proposed distributed model with an average accuracy of 98.54% and 99.67% obtained from both the single and ensemble models respectively.

Keywords: *Fisher's discriminant ratio, Stability index, Ensemble Learning Methods, Supervised algorithm, Distributed Feature Selection method*

1. Introduction

According to the literature, the increasing size of datasets poses a great challenge in the field of data mining. As previously stated, as of 2005 alone, the amount of dataset increased to over 600 terabytes. Studies also show that terabytes of data are collected online every second from surveys, more than experts or researchers could possibly mine [1, 2, 65]. Because of this, machine learning algorithms may be limited in a number of ways according to the volume, complexity, variety, and authenticity of the data [3, 4]. The majority of machine learning (ML) algorithms struggle to scale-up due to the nascent and possible difficulties posed by these computational complexities. This results in inefficiency, instability, and sub-optimal performance in terms of time and space complexity. These scaling-up issues affect the majority of classical or statistical methods, particularly when data sizes surpass their capability. This leads to poor model performance, a decrease in generalization ability, and an increase in computing complexity in terms of time and space [4, 5]. The literature uses Feature Selection (FS) techniques to address this problem when dealing with high dimensionality.

In order to retrieve tiny feature subsets with minimum loss or maximum performance increase, feature selection is the crucial procedure that entails the detection and removal of irrelevant, redundant, and noisy data [6,7,8]. Filter, wrapper, and embedding are the three main categories of FS approaches. By relying on the generic properties of the training data, the filter model carries out the FS method as a pre-processing step separately from the induction algorithm. Wrappers, on the other hand, incorporate a prediction optimization step into the selection process. The embedded approach, which is usually customized for each learning machine, falls in between these two models and performs FS during training. Because of feature-to-feature or feature-to-class interactions, the wrapper and embedded approaches outperform the filter method in terms of prediction accuracy; nevertheless, they come at a greater computational cost.

Igodan E. C. and Iyekowa J. O. (2025). Framework for a Stimulated Predictive Distributed Learning Method. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)*, Vol. 13 No. 1, pp. 24-37

Research on the genetic causes of diseases is becoming easier thanks to high-throughput technologies like next-generation sequencing, mass spectrometry, DNA and RNA microarray profiling, and others [7, 8]. However, filtering techniques are better when working with high-dimensional data [2,9,10] as in this study. The issue of feature interaction (i.e., feature-class correlation or non-correlation) can result in the loss of crucial information and, as a result, poor categorization outcomes that impact the filter method's performance. This is because the intrinsic characteristic of feature interaction is its reducibility, that is, a feature could lose its relevance due to the absence of its interacting features resulting in the unstable feature subsets used for the prediction model. Whereas, the goal of FS stability is to boost domain experts' confidence in the analysis of results obtained by carefully choosing features that are comparatively resilient to changes in input data [8,11].

Traditionally, FS methods have been used in a centralized manner, in that, a single learning algorithm is used to address a particular problem in a standalone environment without the need for coupling, or combining, or utilizing an ensemble of several algorithms for optimality. The need to address this limitation with the traditional or centralized method introduced the distributed learning method [1-3,12-14]. There are two basic ways of partitioning data in distributed learning methods: horizontal and vertical partitioning [2,11,12,14-18].

The practice of creating parallel processes by way of splitting the original training dataset into smaller sets is referred to as partitioning. While horizontal or homogeneous partitioning splits the dataset into many packets having the same attributes as the original set, the vertical or heterogeneous method splits the dataset into multiple packets having the same number of instances as the original but with different attributes [4,16]. A combination of both approaches [4], forms what is referred to as the mixed or horizo-vertical distributed FS methods described in [16].

Overall, these aforementioned approaches did not address the complexity, nor the stability issues within their model as a result, they affected their performance with regard to time and space complexity in achieving high accuracy. In this study, the distributed FS method is adopted by applying the FS algorithms to the partitioned dataset, and finally, the results obtained from the FS methods are aggregated to obtain optimal results for prediction models based on support vector machine (SVM), K-nearest neighbour (KNN), classification and regression (CART) Decision tree, Logistic regression (LR), Naive Bayes (NB) and Multilayer perceptron (MLP) algorithms respectively. Furthermore, we decided to improve our previous work described in Igodan *et al.,* [19] by the introduction of more key FS and classification algorithms, aiming to improve their generalization capability and accuracy, thereby reducing the time and space complexity of the models.

In the following sections, related works are discussed. We provide the datasets and the methods used for our study. We briefly introduced the proposed architecture, various FS methods, and ensemble methods of the study. We then present the results of our analysis and comparative performance analysis against established approaches in the field. Finally, we discuss the implications of our findings and suggest potential directions for future research.

2. Related Works

In Tan and Gilbert [20], an ensemble machine learning approach using bagged and boosted decision trees, k-fold CV, and C4.5 decision trees for cancer classification was described. Duan *et al.,* [23] designed a multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data using a backward elimination procedure (Multiple-SVM-RFE) and SVM classifier. The study did not provide FS stability.

In Wang *et al.,* [24] gene selection from microarray data for cancer classification was used on three learning algorithms and four FS selection methods: chi-squared, Information gain, symmetric uncertainty, ReliefF with Decision trees, naive Bayes and SVM classifiers. However, the complexity of their model was high which affected classification performance.

A Novel Ensemble Approaches for Cancer Data Classification was researched in Zhao [25]. The ensemble method was based on correlation analysis methods only, which includes cosine coefficient, Pearson correlation, Spearman correlation, Euclidean distance. Dogan *et al.,* [26] developed a feature correlation algorithm with application to biological sequence classification. Chi-square, Information gain, Mutual information, and KL-distance were used as FS methods and C-modified least squares (CMLS) as classifier yielding a reduced diversity accuracy. In Kim and Cho [27], a novel approach towards optimal ensemble classifiers for DNA microarray data analysis was presented using evolutionary algorithms. The MLP, SASOM, SVM, and K-NN classifiers, and the following statistical measures were used: Pearson correlation, Spearman coefficient, Euclidean distance, cosine coefficient, information gain, mutual information, signal-to-noise ratio, and real value genetic algorithm.

However, their study was characterized by poor efforts to optimize classifiers, whereas details of the combination of feature subsets used was not discussed, and the lack of diversity in both the classification and feature selectors adopted. Their work was limited to binary-based classification problem. A hybrid of both filter and wrapper FS method used for microarray classification was designed in Chuang *et al.,* [28] using information gain, binary PSO and GA with SVM and K-NN classifiers. Hameed *et al.,* [36] proposed Filter-Wrapper Combination and Embedded FS for Gene Expression Data using ReliefF, Least Absolute Shrinkage and Selection Operator (LASSO), and WrapperSubsetEval (with greedy stepwise search) on Bayes Net, SVM, Naive Bayes, and K-NN classifiers. However, their study was limited as filter-based model with low diversity. In Potharaju and Sreedevi [17], a distributed FS strategy for microarray gene expression data to improve the classification performance was proposed.

The study used Symmetric Uncertainty (SU), Correlation-based feature subset selection (CFS), Synthetic minority oversampling technique (SMOTE) on MLP, K-NN, SVM, SC (tree) and Ridor (Rule-based) classifiers. The study was characterized by low generalization, increased runtime complexity in using MLP. In Li *et al.,* [37] Random value-based oversampling (RVOS), Recursive feature elimination, Variable-step size RFE, Linear-SVM, Large-scale LSVM, and L2 regularized logistic regression, an effective FS and classification method for microarray data was proposed. However, the risk of over-fitting, classifier-dependent selection and computationally expensiveness characterized their study. Tuysuzoglu *et al.,* [38] proposed an ensemble method in environmental data mining by applying standard single classifier involving decision tree (C4.5), naive Bayes, SVM, and K-NN. The single ensemble strategy used was a limitation to the environment engineering fields in the study, also the lack of use of oncology to extract semantic relations to improve accuracy and develop better decision support systems characterized the study.

A parallel FS for distributed-memory clusters was proposed by Gonzalez-Dominguez *et al.,* [39] applying fast-mRMR-multi-threaded parallelization (MPI), and Open Message Passing (MP) threads. The magnitude of the acceleration (lowest runtime) depends on the characteristics of the dataset - number of samples less than the number of features, and the model was not flexible as it does not accept more formats for the input datasets.

An insight into distributed feature ranking was investigated by Bolon-Canedo *et al.,* [40] while the authors in McConnell & Skillicorn [22], built predictors from vertically distributed data by applying a decision trees (J48) ensemble approach for vertically partitioned data with replacement. Their approach was limited being an unstable and only model-based. Abeel *et al.,* [29] built a robust biomarker identification for cancer diagnosis with ensemble FS Methods. In the study, an IQR-normalization and an ensemble FS using SVM-RFE and linear SVM was adopted. However, no evaluation was carried out. Hernandez *et al.,* [30] implemented a multiple-filter-GA-SVM method for dimension reduction and classification of DNA microarray data.

In the study, Between and within sum of squares (BSS/WSS), Wilcoxon test and T-statistical filters GA/SVM on a LOOCV was used. Their study was computationally

expensive and characterized with the risk of over-fitting. The works of the authors in Nagi and Bhattacharyya [31], conducted experimental comparison of J48, NB, IBK on nine microarray cancer datasets and also analyzed their performance with Bagging, Boosting and Stack Generalization. Bolon-Candeo *et al.,* [32] suggested a parallel FS technique with C4.5, naive Bayes, IB1, SVM, and correlation-based, consistency-based, INTERACT, ReliefF, and information gain classifiers from vertically partitioned data. Bolon-Canedo *et al.,* [12] suggested application of distributed FS to microarray data classification applying the following classification algorithms of C4.5, naive Bayes, SVM, k-NN on consistency-based filter, INTERACT, correlation-based FS for classification problem. However, the variability associated with FS method was not addressed and their model was based on the vertical partitioning and distribution of binary data only and limited by the poor class distribution and lacks class representation.

Hodge *et al.,* [33] suggested a Hadoop neural network for parallel and distributed FS using Mutual information, Chi-squared, gain ratio, odd ratio (OR), and correlation-based features subset selection (CFS) on K-NN and Associative memory (binary) neural network applied on an advanced uncertainty reasoning architecture (AURA) framework, and Apache Hadoop. Sun *et al.,* [34] proposed a hybrid gene selection method based on ReliefF and ant colony optimization algorithm for tumor classification. Their research was hampered by the algorithm's inability to balance the size of the chosen gene subset and classification accuracy in all high-dimensional gene expression datasets, as well as experienced an inadequate biological explanations of the gene chosen for cancer classification.

In Mandell and Mukhopadhyay [35], an improved minimum redundancy maximum relevance (mRMR) approach for FS on gene expression data was proposed. However, the study did not consider the scalability and stability of the feature subsets and lacks generalization ability. Bolon-Canedo *et al.,* [1] used information gain on a vertically partitioning of data without replacement to design a distributed wrapper model for FS using SVM, IB1, naive Bayes, and C4.5 classifiers. The study, though experienced over-fitting, is classifier dependent and computationally expensive. Seijo-Pardo *et al.,* [13] proposed an ensemble FS using Chi-squared, mRMR, information gain, SVM-RFE, ReliefF, and FS-P for rankings of feature whilst SVM-Rank is used as the combination method. However, the diversity and stability of the models were not addressed in their study.

A distributed FS approach based on a complexity measure was proposed by Bolon-Canedo *et al.,* [10]. In the work, naive Bayes, SVM, k-NN, C4.5 classifiers with Correlation-based, Consistency-based, INTERACT, ReliefF, and information gain FS were modeled in a horizontally partitioning method. However, there was no sensitivity analysis done. A comparison of distributed and centralized FS techniques based on data complexity measures was done in Bekkerman *et al.,* [2].

For the comparative examination of four classifiers - naive Bayes, C4.5, k-NN, and SVM - the study employed Fishers' discriminant ratio (D-F1), length of the overlapping region (D-F2), and ratio of average intra/inter class using nearest Neighbour distance (D-N2). Their analysis was limited by the difficulty to detect redundancy between features as they were distributed across the packets of features, also the feature selectors scalability issues; and the combination of partial results for both distributed learning were not addressed.

Furthermore, most recently, the authors in Duarte, [65] proposed a distributed fuzzy cognitive maps for feature selection in big data based on wrapper method. The experimental results obtained from a classification task shows that the features selected helped to expedite the classification process using the random forest classifier with an average accuracy above 90% as opposed to 85% when no feature selection strategy was adopted. Also Zerhari *et al.,* [16], proposed a new horizo-vertical distributed features selection approach by applying Chi-squared, information gain, gain ratio, reliefF, CFS, INTERACT, mRMR, and Consistency-based FS filters on Naive Bayes, SVM, C4.5, and K-NN as classifiers. The study did not report time complexity and

did not address the overlap between features as well as the features stability.

In Brankovic and Piroddi, [41] a distributed FS scheme with partial information sharing by vertical data partitioning and a distributed searching architecture was proposed. The Neural network was applied on features selected through the use of Sequential FS, ReliefF, and Randomized FS classifier. No a priori filtering and parameter optimization in their study which affects the optimal performance of the study. In Singh [42], a hybrid meta-heuristic strategy for the classification and selection of gene expression data was developed. The research was characterized with model overfitting. Singh and Kavith [43], carried out an analysis of microarray gene expression data using various FS and classification techniques. Information gain, Analysis of Variance (ANOVA), and Correlation (FCBF), and AdaBoost, neural networks and Random Forest on k-fold cross validation was adopted in their study. Their work was limited by overfitting, low generalization and classifier-dependent.

Thiyagupriyadharsan and Suja [44], proposed Classification of Brain MRI Tumor Images using Fuzzy C Means Clustering with Firefly Algorithms Optimized Support Vector Machine. Colombelli *et al.*, [45] proposed a Hybrid Ensemble FS model for Candidate Biomarkers Discovery from Transcriptome Profiles using five FS methods on a SVM, random forest, and gradient boosting classifiers. The consistency and Kuncheva stability index were adopted using Border count voting approach. However, the study only concentrated in binary problems. The stability of different aggregation techniques in Ensemble FS was proposed in Salman *et al.*, [46]. The within aggregation method (WAM) using the Spearman and correlation coefficient, and the averages of canberra's distance, and Jaccard's index similarity measures were adopted in their study. The features were selected using information gain, symmetric uncertainty (SU), Chi-square, and mRMR approaches.

Additionally employed as aggregating approaches were the geometric mean, arithmetic mean, L2 Norm, robust rank aggregation (RRA), and Stuart aggregation. Last, but not the least, the authors in Al-Shalabi [47], proposed a new FS algorithm based on feature stability and correlation to select the effective minimum subset of appropriate features. The study showed high predictive accuracy through the pioneering of significant reduction of a given dataset, and the importance of stability measure.

The purpose of most related works proposed so far concentrated on finding optimal subset of features while maintaining the physical meanings of the original feature sets for better classification performance best satisfying the objective set, accuracy, precision, computational and storage complexity in the learning process. However, as different FS algorithms vary as to how they perform on a given dataset when used in a distributed learning approach - a limitation we observed in the literature - their stability is influenced by the quantity and quality of the selected feature subsets and the dataset's underlying characteristics. This gap is what this research study wants to address in our proposed horizontally distributed learning approach through the use of Kuncheva and Jaccard stability measures [11,48,45,46,49] and the Fisher's Discriminant Ratio (FDR) in mitigating the features' complexity problem as described in [2,9,10], and the imbalanced feature sample size distribution [62].

3. Methodology

3.1 Datasets And Attributes

The DNA microarray brain tumor dataset were used in this study to assess the efficacy and performance of the proposed distributed framework and described in Table 1. The description shows the original and the SMOTE-based datasets divided into 2/3 and 1/3 train and test sets respectively maintaining the class distribution. The URL is at https://file.biolab.si/biolab/supp/bi-cancer/projections/info/brai ntumor.html.

3.2 Data Preprocessing

The three primary data preprocessing stages used in this study are mean imputation, normalization, and SMOTE as described in [19, 59].

3.3 Classification algorithms

Six popular classifiers from two different classification families were employed [55, 56, 60] . These include two linear classifiers: Naive Bayes [8] and KNN [57], as well as four non-linear classifiers: Logistic regression [58], SVM, Decision tree (CART) [57], and MLP [17] respectively.

The following literature [1, 2, 12, 14] provides details about the centralized and distributed systems, while the former compared with our proposed approach in terms of accuracy, time complexity, number of selected features, and classification accuracy. For this research, an HP Intel Core i5 vPro 7th Gen processor running at 2.9 GHz was used, utilizing a Jupyter-notebook API running on the Google Collab platform and using Python as the implementation language on a Windows 10 PC.

3.4 Proposed Distributed Feature Selection Algorithm

The premise, that integrating the work of several experts is superior to the production of any one expert is the foundation for the idea of distributing the data horizontally. The proposed distributed FS learning architecture is depicted in Figure 1.

The stages of the suggested methodology of this study were described in our previous work in [19,59] and as follows:

A. The training datasets are divided into multiple packets (by samples);
B. The FS techniques are applied to the subsets in multiple rounds to choose pertinent characteristics.
C. Computation of stability index for consistency measure
D. Computation of the Data Complexity measure to determine good candidate features
E. Combining several results for each distribution approach into a single features subset based on complexity measure
F. Build classifiers based on the selected stable subsets to evaluate the selected features

This study consists of three main parts, the first part divides the datasets into packets, applies the FS techniques, computes the consistency index (1) and (2), and complexity measure using Kuncheva (3) and Jaccard index (4), and finally aggregates the selected features using majority voting scheme to selects the final features for modelling. Finally, the modellingis done using individual classifiers and their ensemble methods.

Table 1: Brain tumor datasets characteristics

Dataset (Brain)	Features	Training	Test	Packets	Classes
Non-smote	7,130	30	10	2	5
SMOTE	7,130	1,000	236	6	5

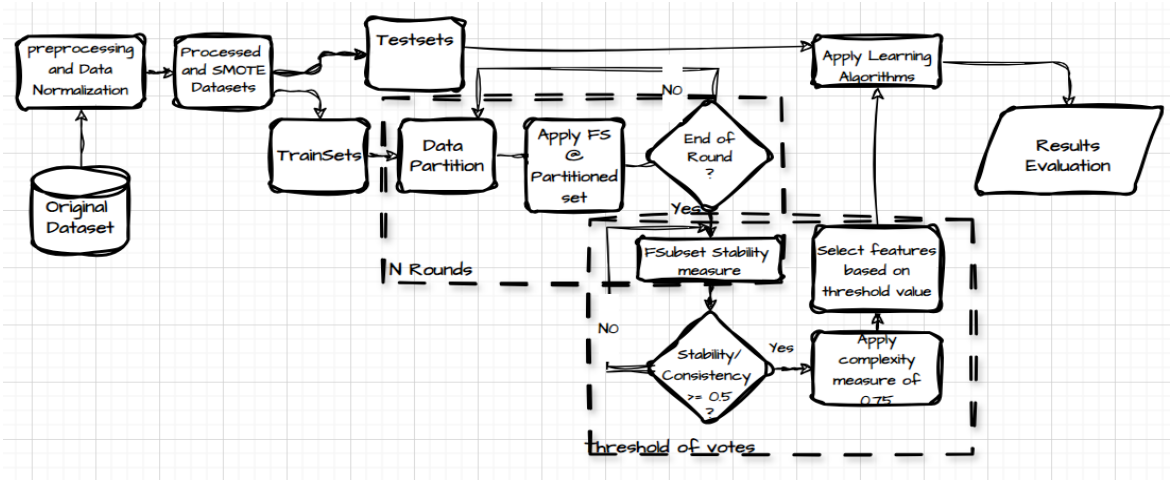

Figure 1: The proposed distributed FS learning architecture

$$\boldsymbol{\varphi(S)} = \frac{\mathbf{2}}{\boldsymbol{r(r-1)}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{r-1}} \sum_{\boldsymbol{j=i+1}}^{\boldsymbol{r}} \boldsymbol{\varphi(S_i, S_j)} \qquad (\mathbf{1})$$

Where $\boldsymbol{\varphi(S_i, S_j)}$ is a predefined stability measure. In our work, we present the different metrics after $\mathbf{r = 5}$ runs as:

$$S = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_l \end{pmatrix} = \begin{pmatrix} x_{11} \& x_{12} \ldots \& x_{1n} \\ x_{21} \& x_{22} \ldots \& x_{2n} \\ x_{l1} \& x_{l2} \ldots \& x_{rn} \end{pmatrix} \qquad (2)$$

$$\varphi Kuncheva(S_i, S_j) = \frac{|S_i \cap S_j| * n - k^2}{k * (n-k)} \qquad (3)$$

$$\varphi Jaccard(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \qquad (4)$$

This study design entails executing multiple fast FS methods on different partitioned packets of training data by sample while maintaining the original class distributions in order to prevent bias in the learning process. Specifically, there are two basic partitioning schemes involved in this study: 2 and 6 partitions, and each partitioning scheme is applied over 5 rounds, producing 10 and 30 feature subsets respectively after applying the FS methods. At the end of each round, features are removed based on votes as described in [19, 59]. Then, the stability index is calculated using both the Kuncheva [49] and Jaccard [61] index methods. The subset with a stability index that meets the predetermined value of 0.5 is selected while others are discarded.

This stage involves selecting the most recurring and similar features from among the features generated after each round or iteration. This is to reduce the size of the subset of features. Subsequently, the complexity index is calculated using a weight factor of 0.75 as suggested in [9, 19, 59]. The purpose of applying the complexity measure CM using the Fisher's Discriminant Ratio (FDR) in (5) is to find the adequate number of features determined by a weight factor of 0.75 in obtaining a more robust feature subset that is stable and can improve the performance of any model while handling both binary and multi-class datasets.

In the literature, it is assumed that using a complexity measure suggests that good candidate features contribute to decreasing the complexity of the model, which must be maintained, while bad candidate features contribute to increasing the model's complexity and must be discarded [1, 2, 6, 7, 62]. Although the weight factor ranges from 0.25, 0.5, and 0.75, this study applied 0.75 in calculating the fitness criterion to ascertain classification performance with respect to accuracy, number of features selected, and time complexity of the various models.

$$CM = F1 = \frac{1}{f} \qquad (5)$$

$$f = \frac{\sum_{i=1, j=1 i \neq j}^{C} P_i P_j (\mu_i - \mu_j)^2}{\sum_{i=1}^{C} P_i Q_i^2} \qquad (6)$$

where μ_i, σ_i^2, p_i are the mean, variance, and proportion of the ith class respectively. To access the experimental results both for the datasets with SMOTE technique [51] were used, and evaluated using standard metrics [19].

4. Results and Discussion

In this section, the experimentation of the proposed model's performance and their results are discussed with regard to their accuracy, number of selected features used, and runtime complexity. This study used the default parameter setting for all cases.

The proposed horizontal distributed FS approach, based on the stability and complexity measures, was implemented and compared with the centralized approach on a SMOTE-based microarray dataset. Table 2 depicts the performance of the centralized approach, which indicates significant improvement when the centralized approach was applied on the SMOTE-based dataset in terms of the runtime, number of selected features, and the classification accuracy respectively. Furthermore, the symmetric uncertainty, correlation and consistency-based FS methods selected 250 features out of the 7140 features within a runtime of 11.62, 9.25, and 9.50 minutes respectively. These improvements were obtained as a result of the use of the SMOTE technique, in achieving both the class imbalance to mitigate the over-fitting problems associated in the models [12].

The corresponding classification accuracy obtained by the centralized approach is depicted in Table 3. The highest accuracy of 97% was obtained by Naive Bayes and the use of information gain while the least minimum of 70% was obtained by Logistic Regression (LR) and correlation-based FS method. The visualization of Table 3 is represented in Figure 2 showing the average accuracy of each model after applying all six FS methods.

Table 4 summarizes the results of the corresponding ensemble methods in a centralized environment. Both the boosted Naive Bayes with symmetric uncertainty and bagged SVM with symmetric uncertainty and correlation-based FS methods obtained the highest accuracy of 98% respectively. Overall, on the average accuracy, boosted naive Bayes achieved 96.5% as depicted in Figure 3.

Table 5 depicts both the number of features selected and the respective time complexity of the different FS methods applied on the Non-SMOTE and SMOTE datasets respectively. The minimum number of 190 features was obtained using the consistency-based FS method when applied on the SMOTE dataset while 2.26 minutes was achieved as the minimum run time on the SMOTE dataset respectively.

Table 6 captures the corresponding accuracy for the proposed model. Naive Bayes obtained an accuracy of 98.98 as the highest as a based model with Chi-squared FS method and also obtained 98.54 on average as shown in Figure 4 respectively. The scalability of the existing FS methods in the proposed distributed manner showed improvement in time complexity, the number of selected features, and increased classification accuracy achieved. The general idea, as originally proposed in Bolon-Canedo *et. al.* [12] and later modified by Moran-Fernandez [2] and Ho and Basu [63], was applied in this study. To overcome the drawback of overfitting, the SMOTE technique was used to obtain a balanced class distribution and feature/sample ratio and was used with a horizontal partitioning approach.

Furthermore, the computational burden experienced in Bolon-Canedo *et. al.* [12] by using their classification error was mitigated by updating our final feature subset according to their theoretical complexity measures using the FDI [2] instead. Hence, our study is independent of the classifier chosen, which reduced the time complexity significantly compared to the centralized approach as reported in Moran-Fernandez [62].

In light of these achieved results, since the running time and number of features selected were consistently reduced while classification accuracy did not drop to inadmissible values but rather increased, the ensemble learning method was also introduced in our study. The idea of ensemble learning [12] was introduced along with some selected state-of-the-art FS algorithms in this study to further improve the performance of other machine learning algorithms from our previous studies [19, 59].

With the benefits of applying data complexity measures in Haritha *et. al.* [64] and the distributed FS proposed in Moran-Fernandez *et. al.* [2], we further reduced the problem of variability through the diversity in the selection of features using ensemble learning methods as indicated in Table 7 and Figure 5 respectively. The boosted DT ensemble model obtained the highest accuracy with Chi-squared, as 99.95%. Overall, an average accuracy of 99.67% was obtained using the boosted DT, while the least accuracy obtained was from bagged KNN at 96.35% respectively.

5. Conclusion

The paper proposes a distributed FS method based on data complexity and feature stability measures in a distributed environment using brain tumor microarray benchmark dataset. The method consists of three main parts: partitioning of the dataset, calculating the data complexity using the Fisher's discriminant ratio (FDR) to find an adequate number of features, and measuring the feature stability index so as to reduce feature variability in the presence of perturbation using the Kuncheva and Jaccard measures. In addition, this paper, provided an empirical study to evaluate six FS methods, base classifiers, and selected ensemble learning methods so as to reliably ascertain the models classification performance and applicability when combined with the FS methods in the distributed learning framework. The results obtained clearly confirm our expectation of an improved performance with appreciable stability and still maintaining a reduced number of features with a corresponding reduced run time complexity. An accuracy of 98.98% and 99.94% was achieved in applying both base Naive Bayes classifier with Chi-squared FS method, and an ensemble bagged MLP classifier with a consistency-based FS respectively.

Table 2: Number of features and time complexity for the centralized approach

Feature Selector	No. of Features Selected		Time complexity (Minutes)	
	Non-SMOTE	SMOTE	Non-SMOTE	SMOTE
Info Gain	3520	1000	23	18.30
Gain Ratio	3550	1000	20	16.98
Chi-Squared	3600	1000	21	15.80
Symmetric uncertainty	2500	250	25	11.62
Correlation	2520	250	23	9.25
Consistency	2555	250	24	9.50
Average time			**22.67**	**13.58**

Table 3: Centralized approach (single classifiers)

Feature Selector	NB	KNN	LR	SVM	DT	MLP
Info Gain	97	90	90	80	85	90
Gain Ratio	95	85	85	82	85	85
Chi-Squared	87	75	87	87	85	87
Symmetric	92	75	85	75	95	85
Correlation	86	80	70	90	87	87
Consistency	86	75	80	90	90	79
Avg Accuracy	**90.5**	**80**	**82.83**	**84**	**87.83**	**85.5**

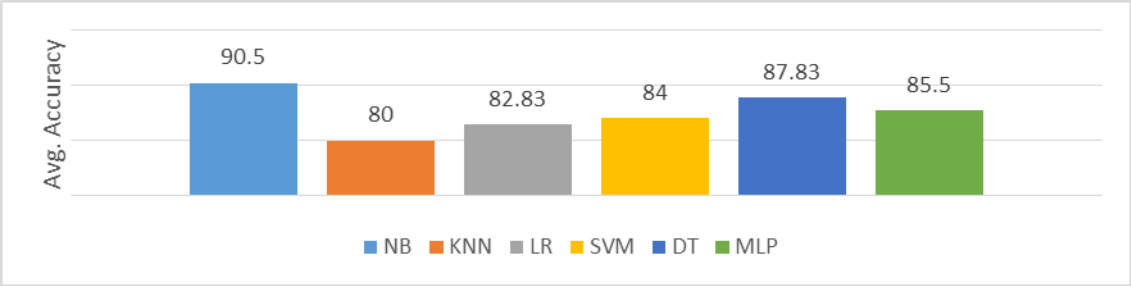

Figure 2: Average Accuracy for Base Classifiers of Centralized Approach

Table 4: Centralized approach (ensemble classifiers)

Feature Selector	Boost NB	Bag NB	Boost DT	Bag DT	Boost SVM	Bag SVM	Bag KNN	Bag MLP
Info Gain	95	90	95	90	90	90	97	90
Gain Ratio	97	95	90	90	90	90	95	90
Chi-Squared	97	87	86	95	87	87	95	97
Symmetric	98	95	97	94	95	98	97	90
Correlation	97	87	94	96	95	98	93	97
Consistency	95	95	95	97	94	96	87	98
Avg. Accuracy	**96.5**	**91.5**	**92.83**	**93.67**	**91.83**	**93.17**	**94**	**93.67**

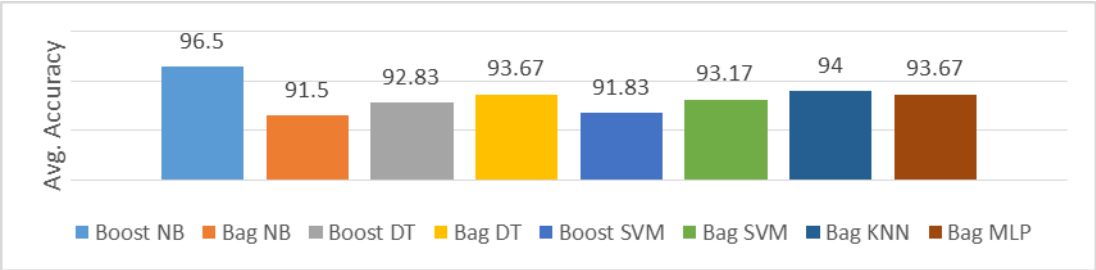

Figure 3: Average accuracy for ensemble of centralized approach

Table 5: Number of features and time complexity for the distributed approach

	No. of Features Selected		Time complexity (Min)	
Feature Selector	**Non-SMOTE**	**SMOTE**	**Non-SMOTE**	**SMOTE**
Info Gain	501	200	1.01	3.94
Gain Ratio	650	250	1.50	4.54
Chi-Squared	750	300	1.40	2.97
Symmetric	350	220	0.95sec	5.68
Correlation	340	200	0.32sec	2.26
Consistency	341	190	0.25sec	5.33

Table 6: Distributed approach (single classifiers)

Feature Selector	**NB**	**KNN**	**LR**	**SVM**	**DT**	**MLP**
Info Gain	98.47	98.80	94.80	97.75	97.00	98.00
Gain Ratio	98.90	97.80	94.80	97.90	98.47	97.90
Chi-Squared	98.98	96.80	94.47	98.80	98.47	97.82
Symmetric	98.29	95.60	95.64	98.50	98.00	98.47
Correlation	98.47	95.90	95.82	89.80	98.29	98.29
Consistency	98.11	96.80	96.29	98.90	98.82	95.94
Avg. Accuracy	**98.54**	96.95	95.30	96.94	98.17	97.74

Table 7: Distributed approach (ensemble classifiers)

Feature Selector	**Boost NB**	**Bag NB**	**Boost DT**	**Bag DT**	**Boost SVM**	**Bag SVM**	**Bag KNN**	**Bag MLP**
Info Gain	99.70	99.47	99.82	99.50	97.26	99.50	93.82	98.20
Gain Ratio	99.70	99.47	99.47	99.64	98.26	98.20	97.50	99.00
Chi-Squared	99.47	99.47	99.95	99.47	96.73	98.49	95.82	98.82
Symmetric	99.64	99.29	99.64	99.64	97.32	99.47	97.82	98.47
Correlation	99.47	99.47	99.29	99.64	98.29	99.10	94.90	98.50
Consistency	99.47	99.11	99.82	99.82	98.82	98.94	97.23	99.94
Avg. Accuracy	99.58	99.38	**99.67**	99.62	97.78	98.95	96.35	98.82

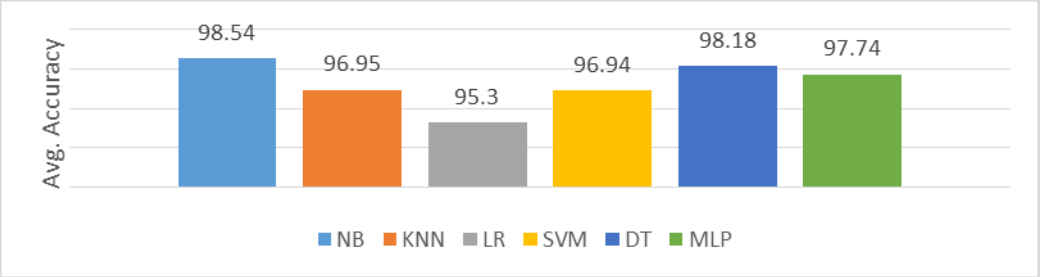

Figure 4: Average Accuracy for Base Classifiers of Proposed Distributed Approach

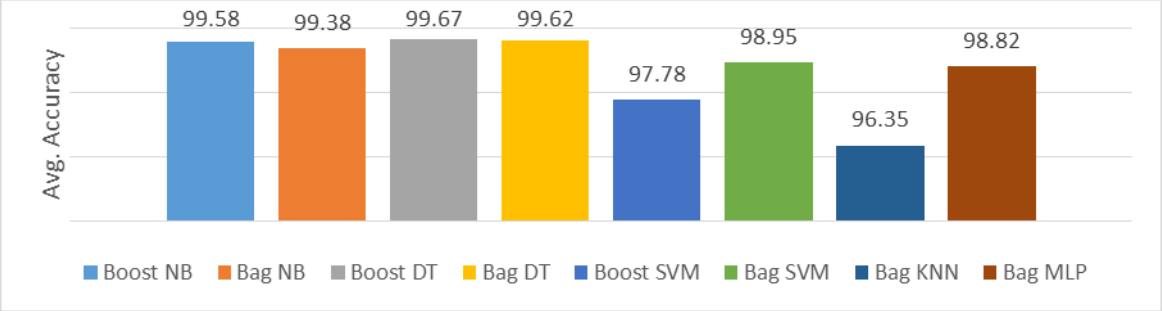

Figure 5: Avg. Accuracy for Ensemble of Proposed Distributed Approach

References

[1] Bolon-Canedo, V., Sanchez-Marono, N., & Alonoao-Beranzos, A. (2013a). A distributed wrapper approach for feature selection. ESANN 2013 proceedings, European symposium on artificial neural networks, computational intelligence and machine learning, pp. 173-178.

[2] Moran-Fernandez, L., Bolon-Canedo, V., & Alonso-Betanzos, A. (2016). Centralized vs. distributed feature selection methods based on data complexity measures. Knowledge-based systems, 000:1-19.

[3] Bekkerman, R., Bilenko, M., & Langford, J. (2012). Scaling up machine learning: Parallel and Distributed Approaches, Cambridge.

[4] Peteiro-Barral, D., & Guijarro-Berdinas, B. (2013). A Survey of Methods for Distributed Machine Learning. Prog. Artif. Intell., 2:1-11. Springer.

[5] Zhang, Y. (2016). Distributed machine learning with communication constraint, Ph.D. Thesis. University of California, Berkely.

[6] Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2015c). Feature Selection for High-Dimensional Data. Springer.

[7] Igodan, C.E. & Ukaoha, K.C. (2019d). Using Multi-layer Perceptron and Deep Neural Networks for the Diagnosis of Breast Cancer Classification," IEEE Africon, pp. 1-7, doi:101109/AFRICON46755.2019.9133873.

[8] Sahu, B., & Mishra, D. (2012). A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data. Procedia Engineering, 38: 27-31, 2012. https://doi.10.1016/j.proeng.2012.06.005

[9] Bolon-Canedo, V., Sanchez-Marono, N., & Cervino-Rabunal, J. (2013b). Scaling up feature selection: a distributed filter approach. In: Bielza, C., Salmeron, A., Alonso-Betanoz, A. Hildalgo, J.I., Martinez, L., Tronocoso, A., Carchado, E., Corchado, J.M. (eds) CAEPIA 2013. LNCS, Vol. 8109, pp.121-130. Springer, Heidelberg.

[10] Bolon-Canedo, V., Sanchez-Marono, N,m & Alonso-Betanzos, A. (2015a). A distributed feature selection approach based on a complexity measure. In international work-conference on artificial neural networks, pp. 15-28. Spain: Palma de Mailorea

[11] Maalej, Z., Rejab, F.B., & Nouira, K. (2022). Risk Factors of Breast Cancer Determination: a Comparative Study on Different Feature Selection Techniques. Research Square; 2022. DOI: 10.21203/rs.3.rs-2120645/v1.

[12] Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2015b). Distributed feature selection: an application to microarray data classification". Applied soft computing, 30:136-150. https://doi.org/10.1016/j.asoc.2015.01.035

[13] Seijo-Pardo, B., Bolon-Canedo, V., &Porto-Diaz, I. (2015). Ensemble Feature Selection for Rankings of Features. Springer, pp. 29-42. DOI: 10.1007/978-3-319-19222-2-3.

[14] Hoque, N., Singh, M., & Bhattacharyya, D.K. (2017). EFS-MI-an ensemble feature selection method for classification. Complex Inteli. Syst. Springer, 2017. DOI 10.1007/s40747-017-0060-x.

[15] McConnell, S., & Skillicorn, D.B. (2004). Building predictors from vertically distributed data. In: Proc. Of the Conference of the Centre for advanced studies on collaborative research," IBM press, 2004, pp. 150-162.

[16] Zerhari, B., Ait Lehcen, A., & Mouline, S. (2018). A New Horizo-Vertical Distributed Feature Selection Approach". Cybernetics and Information Technologies, volume 18, No 4, pp. 15-28. doi: 10.2478/cait-2018-0045.

[17] Potharaju, S.P., & Sreedevi, M. (2018). Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. Clinical epidemiology and global health, no. 7, 171-176. https://doi.org/10.1016/j.cegh.2018.04.001

[18] Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2014d). Towards parallel feature selection from vertically partitioned data. ESANN proceedings, European symposium on artificial neural networks, computational intelligence and machine learning, pp. 395-400.

[19] Igodan, E.C., Obe, O.O., Thompson, A-F., Owolafe, O., Usiosefe, L.O., & Katyo, P. (2024). Predictive Distributed Learning based on Stability and Complexity Measures. Presented at the 2nd EEE International conference on Science, Engineering and Business for driving Sustainable Development Goals. Omu-Aran, Kwara State, Nigeria, 2nd - 4th April, 2024.

[20] Tan, A.C., & Gilbert, D. (2003). Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics, 2(3 suppl), pp. 1-9, 2003.

[21] Duan, K-B. & Rajapakse, J.C. (2004). A Variant of SVM-RFE for Gene Selection in Cancer Classification with Expression Data, pp. 49-55.

[22] McConnell, S., & Skillicorn, D.B. (2004). Building predictors from vertically distributed data. In: Proc. Of the Conference of the Centre for advanced studies on collaborative research. IBM press, 2004, pp. 150-162.

[23] Duan, K.B., Rajapakse, J.C, & Azuaje, F. (2005). Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data. IEEE Transactions on Nano-bioscience. Vol. 4, No 3, pp. 228-234.

[24] Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, E., Mayer, K.F.X., & Mewes, H.W., (2005). Gene Selection from microarray data cancer classification - a machine learning approach. Computational Biology and Chemistry, 29, 37-46.

[25] Zhao, Z (2007). Searching for Interacting Features, In: IJCAI, 7, pp. 1156-1161.

[26] Dogan, R.I., Getoor, L., & Wilbur, W.J. (2007). Chapter 18: A feature generation algorithm with applications to biological sequence classification. In Computational methods of feature selection. University of minnesota, pp. 355-375. ISBN 978-1-58488-878-9

[27] Kim, K-J., & Cho, S-B. (2008). An evolutionary algorithm approach to optimal ensemble classifiers for DNA microarray data analysis. IEEE transactions on evolutionary computation, Vol. 12, No. 3, pp. 377-388.

[28] Chuang, L-Y., Ke, C-H., & Yang, C-H. (2008). A hybrid both filter and wrapper feature selection method for microarray classification. IMECS. 146-150.

[29] Abeel, T., Helleputte, T., de Peer, Y.V., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, Vol. 26, no. 3, pp. 392-298, doi:10.1093/bioinformatics/btp630.

[30] Hernandez, M.L.A., Bonilla, H.E., & Morales, C.R. (2011). A multiple-filter-GA-SVM method for dimension reduction and classification of DNA microarray data," Revista Mexicana de Ingenieria Biomedica, vol. XXXII, num. 1, pp. 32-39.

[31] Nagi, S., & Bhattacharyya, Kr. (2013). Classification of microarray cancer data using ensemble approach. Netw Model and Health Inform Bioinformatics. Springer, 2:159-173. https://doi.10.1007/s13721-013-0034-x.

[32] Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2014a). Novel feature selection methods for high dimensional data. https://www.scitepress.org/papers/2014/49301/49301.pdf

[33] Hodge, J.V., O'Keefe, S.O., & Austin, J. (2016). HADOOP neural network for parallel and distributed feature selection. Neural network learning in big data, https://doi.org/10.1016/j.neunet.2015.08.01.

[34] Sun, L. Kong, X., Xu, J., Xue, Z., Zhai, R., & Zhang, S. (2019). A hybrid gene selection method based on ReliefF and Ant colony optimization Algorithm for Tumor Classification," Scientific Reports, 9:8979, pp1-14. https://doi.og/10.1038/s41598-019-45223

[35] Mandell, M., & Mukhopadhyay, A. (2013). An improved minimum redundancy maximum relevance approach for feature selection in gene expression data. International Conference on Computational Intelligence: Modeling, techniques and applications. Procedia Technology, 10:20-27. https://doi:10.1016/j.protcy.2013.12.332

[36] Hameed, S., Petinrin, O.O., Hashi, A.O. & Saeed, F. (2018). Filter-wrapper combination and embedded feature selection for gene expression data. International Journal of Advanced Soft Computation Application. Vol. 10, No. 1.

[37] Li, Z., Xie, W., & Liu, T. (2018). Efficient feature selection and classification for microarray data. PLoS ONE, 13(8): e0202167. https://doi.org/10.1371/journal.pone.0202167

[38] Tuysuzoglu, G., Birant, D., & Pala, A. (2018). Ensemble Methods in Environmental Data Mining. Data Mining. DOI: 10.5772/intechopen.74393.

[39] Gonzalez-Dominguez, J., Bolon-Canedo, V., Freire, B., & Tourino, J. (2019). Parallel Feature Selection for Distributed-Memory Clusters. Information Sciences, Volume 496, Pages 399-409, https://doi.org/10.1016/j.ins.2019.01.050.

[40] Bolón-Canedo, V., Sechidis, K., Sánchez-Maroño, N., Alonso-Betanzos, A., & Brown, G. (2019). Insights into distributed feature ranking. Information Sciences, 496, 378-398. https://doi.org/10.1016/j.ins.2018.09.045

[41] Brankovic, A., & Piroddi, L. (2019). A distributed feature selection scheme with partial information sharing. Machine Learning. 108:2009-2034. https://doi.org/10.1007/s.10994-019-05809-y

[42] Singh, R. (2020). A gene expression data classification and selection method using hybrid meta-heuristic technique. EAI endorsed transactions on scalable information systems, volume 7, issue 25, pp. 1-8.

[43] Singh, W.J., & Kavith, R.K. (2020) Analysis of microarray gene expression data using various feature selection and classification techniques. Biosc. Biotech. Res. Comm. Special Issue Vol. 13, No. 11, pp. 105-108.

[44] Thiyagupriyadharsan, M.R. & Suja, S. (2021). Classification of brain MRI tumor images using fuzzy C means clustering with firefly algorithms optimized support vector machine. Journal of university of Shanghai for science and technology, Vol23, Issue 11, pp 70-77.

[45] Colombelli, F., Kowalski, T.W. & Recamonde-Mendoza, M. (2022). A hybrid ensemble feature selection design for candidate biomarkers discovery from transcriptome profiles. Knowledge-Based Systems, Volume 254, 109655, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2022.109655.

[46] Salman, R., Alzaatreh, A., & Sulieman, H. (2021). The stability of different aggregation techniques in ensemble feature selection. Journal of Big Data. Volume 9, pp. 1-23.

[47] Al-Shalabi, L. (2022). New feature selection algorithm based on feature stability and correlation. IEEE Access. Volume 10, pp. 4699-4713.

[48] Khaire, U.M. & Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A Review," Journal of King Saud University computer and information sciences. https://doi.org/10.1016/j.jksuci.2019.06.012

[49] Kuncheva, L.I. (2007). A Stability index for feature selection. In: Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications, ACTA Press, pp. 390-395.

[50] Kalousis, A., Prados, J., & Hilario, M. (2005). Stability of feature selection algorithms: a study on high dimensional spaces. In proc. 5th IEEE international conference on data mining (ICD' 05), pages 218-225. Springer. Doi: 10.1007/51015-006-004-8.

[51] Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2022). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[52] Rish, L. (2001). An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, 3, IBM New York, 2001, pp. 41-46.

[53] Hall, M.T. (1999). Correlation-based feature selection for machine learning. Ph.D Thesis, University of Waikato.

[54] Dash, M., & Liu, H. Consistency-based search in feature selection. Artifificial Intelligence, 151(1-2):155–176. doi:10.1016/S0004-3702(03)00079-1

[55] Igodan, E.C., Obe, O.O., Thompson, A.F-B., & Owolafe, O. (2022a). Erythemato Squamous Disease Prediction using Ensemble Multi-Feature Selection Approach. International Journal of Computer Science and Information Security (IJCSIS), Vol. 20, No. 2, pp. 95-106.

[56] Igodan, E.C., Obe, O., Thompson, A.F-B. & Owolafe, O. (2022b). Prediction of erythemato Squamous-disease using ensemble learning framework. The Institute of Engineering and Technology. In Explainable Artificial Intelligence in Medical Decision Systems, pp.197-228.

[57] Wu, J., & Hicks, C. (2021). Breast Cancer Type Classification Using Machine Learning. JPM, 11, 61.

[58] Srivenkatesh, M. (2020). Prediction of Prostate Cancer using Machine Learning Algorithms. International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-5, pp. 5353-5362.

[59] Igodan, E.C. (2024). Development of a Predictive Distributed Feature Selection Model Using Machine Learning Techniques. Ph. D Thesis.

[60] Rajesh, M.M., Chandrasekar, B.S., & Swamy, S.S. (2022). Prostate Cancer Detection Using Radiomics-based Feature Analysis with ML

Algorithms and MR Images. International Journal of Engineering Trends and Technology. Volume 70, Issue 12, 42-58.

[61] Budhraja, S., Doborjeh, M., Balaran, S., Tan, S., Doborjeh, Z., Edmund, L., Alexander, M., Lee, J., Goh, W., & Kasabov, N. (2023). Filter and Wrapper Stacking Ensemble (FWSE): a Robust Approach for reliable Biomarker Discovery in High-Dimensional Omics Data. Briefings in Bioinformatics, 24(6), 1-17. https//doi.org/1093/bibad382.

[61] Alonso-Betanzos, A., olon-Canedo, V., Moran-Fernandex, L., & Seijo-Pardo, B. (2019). Feature Selection Applied to Microarray Data. Microarray Bioinformatics. Methods in Molecular Biology, Vol. 1986, 2019. https//doi.org/10.1007/978-1-4939-9442-7_7. Springer Science.

[62] Moran-Fernandez, L., Bolon-Canedo, V., Alonso-Betanzos, A. (2015). A Time Efficient Approach for Distributed Feature Selection Partitioning by Features. In Conference of the Spanish Association for Artificial Intelligence. Springer, Cham, pp. 245-254.

[63] Ho, T.K. &Basu, M. (2002). Complexity Measures of Supervised Classification Problems. IEEE Trans Pattern Anal Mach Intell., 24(3):289-300.

[64] Haritha, E., Judy, M.V., Papageorgiou, K., & Georgiannis, V.C. (2024). Distributed fuzzy cognitive maps for feature selection in big data classification. Algorithms, 15, 383. https://doi.orgg/10.3390/a15100383

[65] Duarte, F. (2024). Amount of Data Created Dail. https://explodingtopics.com/blog/data-generated-per-day