INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Enhanced Filtrate Feature Selection Algorithm for Feature Subset Generation

P Usha[1]*, **J G R Sathiaseelan**[2]

1 Assistant Professor, Department of Computer Science, Bishop Heber College, Bharathidasan University, Tiruchirappalli, 620 024, Tamilnadu, India
2 Associate Professor, Department of Computer Science,, Bishop Heber College, Affiliated to Bharathidasan University, Tiruchirappalli, 620 024, Tamilnadu, India

OPEN ACCESS

Received: 30-06-2024

Accepted: 08-07-2024

Published: 30-07-2024

Citation: Usha P, Sathiaseelan JGR (2024) Enhanced Filtrate Feature Selection Algorithm for Feature Subset Generation. Indian Journal of Science and Technology 17(29): 3002-3011. https://doi.org/10.17485/IJST/v17i29.2127

* **Corresponding author**.

usha.it@bhc.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: In the bioinformatics field feature selection plays a vital role in selecting relevant features for making better decisions and assessment of disease diagnosis. Brain Tumour (BT) is the second leading disease in the world. Most of the BT detection techniques are based on Magnetic Resonance (MR) images. **Methods**: In this paper, medical reports are used in the detection of BT to increase the surveillance of patients. To improve the accuracy of predictive models, a new adaptive technique called Enhanced Filtrate Feature Selection (EFFS) algorithm for optimal feature selection is proposed. Initially, the EFFS algorithm finds the dependency of each attribute and feature score by using Mutual Information Gain, Chi-Square, Correlation, and Fishore score filter methods. Afterward, the occurrence rate of each top-ranked attribute is filtered by applying threshold value and obtaining the optimal feature by using the Pareto principle. **Findings**: The performance of the selected optimal features is evaluated by time complexity, number of features selected, and accuracy. The efficiency of the proposed algorithm is measured and analyzed in a high-quality optimal subset based on a Random Forest classifier integrated with the ranking of attributes. The EFFS algorithm selects 39 out of 46 significant and relevant features with minimum selection time and shows 99.31 % of accuracy for BT, 29 features with 99.47% of accuracy for Breast Cancer, 15 features with 94.61% of accuracy for Lung Cancer, 15 features with 98.84% of accuracy for Diabetes and 43 features with 90% of accuracy for Covid-19 dataset. **Novelty**: To decrease the processing time and improve the performance of a model feature selection process will be done at the initial stages for the betterment of the classification task. Thus, the proposed EFFS algorithm is applied to different datasets based on medical reports and EFFS outperforms with greater performance measurements and time. The appropriate feature selection techniques help to diagnose diseases in the prior phase and increase the survival of human beings.

Keywords: Bioinformatics; Brain Tumour; ChiSquare; Correlation; EFFS; Feature Selection; Fishore Score; Information Gain; Optimal Features; Random Forest

1 Introduction

Medical data collected from patients' reports have noisy and incomplete data and don't have a uniform format. Scrutinizing medical data is a sensitive and necessary activity in decision-making and analysis tools. Disease prediction must depend on medical data thus the system needs careful feature selection for decision suggestions. Feature selection is also called attribute or variable selection which is used to select data automatically that are more and most relevant to the predictive model. So many feature selection mechanisms [1,2] have evolved to get purified data to analyze and make decisions for future purposes. When analyzing the dataset with a large number of features leads to poor performance and accuracy. Feature selection is a technique which is not only selects the features, it also eliminates the least significant data.

Brain tumor is one of the most dangerous and aggressive diseases in human beings and it is very difficult to diagnose. According to the Hospital-Based Cancer Registry (HBCR), 32.5 % of males and 36.5 % of females are affected by brain tumors. Due to the sudden growth and replication of cancer tissues, manual identification of abnormal growth in the brain is difficult and tracking changes over time is very tedious. Feature selection is a process to detect the dependency between the target and other features in order to get significant features to improve the predictive accuracy of a model. To reduce the assessment time of disease diagnosis, an effective feature selection mechanism is demanded.

Huanhuan Gong et al. [3] proposed the CONMI feature selection function which combines mutual information and correlation coefficient methods to select features and evaluate 20 datasets using SVM, DT, and KNN algorithms. Feature selected with SVM shows 88.98% of accuracy. Munish Khanna et al. [4], propose an approach that combines soft computing techniques such as Teaching Learning Based Optimization (TLBO), Elephant Herding Optimization (EHO), and a hybrid of TLBO + EHO to select features for the identification of benign and malignant breast cancer types and achieved 97.96% of accuracy on KNN classifier. Rathna Sekhar and Sujaths [5], propose an independent feature selection method based on a genetic algorithm (ISG-GA-IC) to enhance classification accuracy on the high dimensional datasets and produce accuracy up to 97% on 30MB size of dataset.

Arjomandi Rad Mohammad et al. [6] proposed a correlation-based feature selection application in Computer Aided Design. Regression-based machine learning algorithms are used to extract the hidden features that are involved in prediction. The extracted features are used as input to the predictive model which combines multivariate linear regression with gradient descent method. The simulation results are demonstrated with a higher efficiency of 95%. Zahra Asghari Varzaneh et al. [7] proposed meta-heuristic algorithm for feature selection technique which combined with a decision tree model to predict the intubation risk of COVID patients who are hospitalized. This model is used to assess an illness of severity and identify high-risk patients up to 93%.

Babafemi Oluropo et al. [8] proposed MIG, CS, and correlation-based feature selection algorithms with a Random Forest (RF) classifier and Support Vector Machine (SVM) as a predictive model to evaluate the risk of breast cancer in Africans. 11 features are selected out of 25 risk factors. Compared to MIG and CS feature selection methods, correlation-based methods with RF classifiers show greater performance metrics.

Asif Newaz et al. [9] proposed a framework to identify the risk factors of heart failure patients based on laboratory test results and medical records. A familiar CS method tests the statistics to identify the dependability between target and normal features

which is combined with recursive feature elimination to select the features. The selected features are evaluated using a balanced random forest classifier which shows greater performance measures. The proposed framework is used to give guidance to clinicians to increase the surveillance of a patient.

Al Mehedi Hasan et al.[10] proposed recursive feature elimination and minimum redundancy and maximum relevance as a feature selection method to select the most relevant features. This combination shows 69.52 % accuracy using the decision tree classifier. Parnika Bhat et al.[11] proposed a multi-tier feature selection (MTFS) model, for finding the significant features in malware detection systems. In MTFS, features are removed based on information gain score and frequency of occurrences. The selected features are tested and analyzed with a random forest classifier, showing 96.28% accuracy in selecting features. Thus, the MTFS-RF model is used to classify and identify the maliciousness of Android applications.

Jung et al.[12] proposed the Gini index as a feature selection method in a static analysis approach in which API calls and permissions are used as a feature to detect malicious applications in smart devices. The Gini index with random forest algorithm shows 96.51% accuracy in malware detection. Linsun et al.[13] analyze various machine learning algorithms to predict breast cancer and show accuracy range from 52.63% to 98.24%. Anna Karen Garate-Escamila et al.[14] proposed a combined feature selection that includes CS- PCA (Principal Component Analysis) to predict cardiac disease. The CS method arranges the features highly depending on the target value. PCA is used to detect the significant components in which features with greater eigenvalue (>= 1.00) are retained as selected features. CHI-PCA with a random forest classification algorithm shows 98.7% accuracy.

Ahmed Abdullah Farid et al.[15], proposed a new model as CHFS-BOGA (Composite Hybrid Feature Selection Learning-Based Optimization of Genetic Algorithm) for predicting breast cancer. This new model combines the advantages of feature selection approaches such as filter (IG, Gain Ratio), wrapper (Genetic algorithm), and Embedded (C4.5) with Optimized Genetic Algorithm, PCA, and SVM. The new model shows the highest accuracy when it is combined with SVM. Muhammed Niyas et al.[16], proposed an efficient feature selection algorithm to detect Alzheimer's disease among aged people. Greedy and fisher score ranking methods are used as a feature selection mechanism in which patients are classified into different categories based on fisher score values. This model is evaluated using SVM and KNN (K Nearest Neighbour) classifiers with 90% and 91% accuracy. Srivenkatesh[17] proposed an improved fishore score model to decrease the complexity of the multi-label dataset. This model selects features efficiently and improves the classification performance of multi-labeled data.

Comparative analysis:

In BT detection, the pre-existing research work focuses on MR images. The features are based on MR image properties such as brightness, contrast, image quality, intensity, etc. For feature selection, the CNN method is used. The major challenge in the prediction of disease in healthcare is finding the most relevant and non-redundant features. The inclusion of all features in a dataset gives it a high dimension, but it is very hard to interpret. The related work[8–10] concentrates on one or two feature selection methods that are combined to select features with or without the use of pre-processing techniques. In the EFFS algorithm, after pre-processing of the dataset, the suitable feature subset is generated by calculating the feature score, ranking, and rate of occurrence of each attribute using various filter-based algorithms on the patient's medical reports. The EFFS algorithm produces an optimal and robust dataset to assess the suggestion for treatment options that improve the survival rate of a patient.

The contribution of this paper is:

1. This paper proposes the Enhanced Filtrate Feature Selection (EFFS) algorithm, which has three phases. In Phase I, the feature score for each attribute is evaluated using existing filter-based methods such as Mutual Information Gain (MIG), Chi-Squared (CS), Heat Map (HM), and Fishore-Score (F-score). Identify the top-ranked feature based on predefined threshold (less than or equal to 80% of the original feature) values.

2. The rate of occurrence of each top-ranked selected feature is calculated in Phase II. Depending upon the occurrences of each feature (occurrences must be greater than or equal to 3), the feature is selected. Thus, the "EFFS Feature" subset is generated.

3. The result of (EFFS feature is applied to the Diabetes, Breast Cancer, Lung Cancer, COVID-19, and Brain Tumor datasets, and performance is evaluated using a Random Forest classifier with a ranking method.

4. The proposed EFFS algorithm finds relevant features (39 out of 46 features) in the BT dataset with the shortest time to select features (456 milliseconds) and a higher accuracy of 99.31%. The efficiency of the proposed EFFS algorithm compared with other benchmark datasets

This paper is organized as follows: Section 1 provides a brief overview of the related work, including its strengths and weaknesses. In Section 2, the theoretical background, algorithm, and flow diagram of the proposed EFFS algorithm are discussed, and in Section 3, the datasets that are used in the proposed system are discussed. The result obtained from the EFFS algorithm is analyzed with existing techniques in Section 4. Finally, Section 5 concludes the summary and future direction of this work.

2 Methodology

EFFS algorithm workflow model

The workflow of the proposed EEFS algorithm to select the top-ranked base features is shown in Figure 1. Pre-processing techniques are used to check the consistency of the record using data from patients' medical reports. The pre-processed data enters various phases of the proposed EFFS algorithm, and performance measures are evaluated.

2.1. Feature score evaluation

Mutual Information Gain (MIG) is one of the filter-based feature selection algorithms used to measure the mutual dependency between random features and identify the information obtained from one another called entropy (ps), which is calculated by Equation (1).

$$Info(S) = \sum_{i=1}^{m} p_s \, log_2 p_s \tag{1}$$

where Info(S) is an information gain and ps is the entropy of sample S.

The conditional entropy of a subset A of sample S is obtained by Equation (2).

$$Info_A(S) = \sum_{i=1}^{m} \left(\left| \frac{S_i}{S} \right| \right) * Info\ (S) \tag{2}$$

The gain of subset and target feature is obtained by Equation (3).

$$Gain(A) = \ Info(S) - Info_A(S) \tag{3}$$

The feature that has zero IG is eliminated from further processing because there is no dependency between random features. The minimum value obtained from MIG is set as the threshold value, and up to 80% (80/20 Pareto principle) of total features are selected as the top feature. It is denoted by Equation (4).

$$Mutual\_Info(Select_{Feature}) \tag{4}$$

CS is a feature selection method (Syed Asim Ali Shah et al., 2020; 13 Ann Romalt et al., 2020) in which the degree of associativity between two categorical features is measured. In this algorithm, the CS score is calculated for each feature and target feature with the desired number of features based on score values. It is denoted by Equation (5).

$$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^{\,2}}{E_{ij}} \tag{5}$$

Select the features with a high chi value based on the threshold and select the Pareto feature as a top-ranked feature. It can be denoted by Equation (6).

$$Chi\_square(Select_{Feature}) \tag{6}$$

Heat map (Jundong Li et al., 2018) is a feature selection technique that uses a correlation matrix to determine the irrelevance of a feature from a target.In this method, attributes are ranked according to the heuristic function based on the correlation of attributes. It can be calculated using Pearson's correlation coefficient, shown in Equation (7).

$$PCC = \frac{\sum F_1 * F_2 \ - (\sum F_1) \ * (\sum F_2)}{n \sum F_1^2 - (\sum F_1)^2 * n \sum F_2^2 - n(\sum F_2)^2} \tag{7}$$

Depending on the target feature, 0 represents a weak correlation, +1 represents a strong +ve correlation, and -1 represents a strong -ve correlation in a heat map. Top-ranked features that are strongly positively correlated and based on the Pareto principle are denoted by Equation (8).

$$Heat\_Map(Select_{Feature}) \tag{8}$$

Fisher score algorithm used to determine the feature subset. Features are ranked and listed based on importance. F_score is calculated for each feature using by Equation (9).

$$F_{Score}(f_i) = \frac{s_b\ (f_i)}{\sum_{i=1}^{n} S_t^{(i)}\ (f_i)} \tag{9}$$

Features are arranged in descending order and ranked based on scores. Pareto-based features are selected as a top-ranked features and are denoted by Equation (10).

$$F\_Score\,(Select_{Feature}) \tag{10}$$

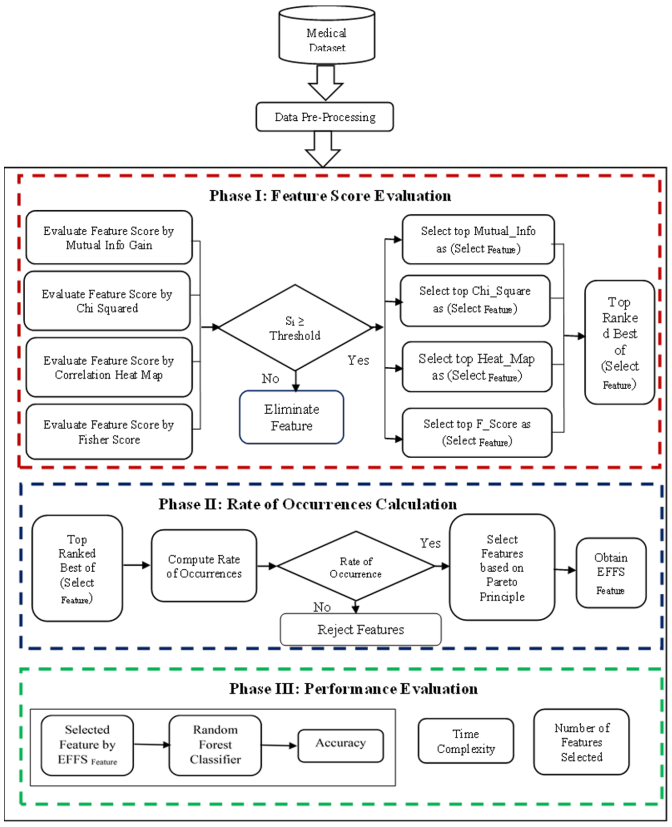

Fig 1. 'Proposed EFFS algorithm workflow model' are shown in which the Feature score is evaluated instep 1, then based on the rate of occurrences of each feature best features are selected. It can be evaluated using a random forest classifier

2.2. Rate of occurrences calculation

Combine all the selected top-ranked features from Equation (6), Equation (7), Equation (8), and Equation (10) is represented in Equation (11).

$$\begin{aligned} Best\,(Select_{Best}) = Mutual\_Info(Select_{Feature}) \bigcup Chi\_square\,(Select_{Feature}) \\ \bigcup Heat\_Map\,(Select_{Feature}) \bigcup F\_Score\,(Select_{Feature}) \end{aligned} \tag{11}$$

Here four feature selection methods are used. The occurrence of each attribute in each method is calculated, if the attribute occurs in three or all methods are selected. Thus rate of occurrence is set as threshold and has the value as 3. Based on the threshold value ($occurrence \geq .3$) is obtained as a feature subset based on number occurrences of each feature. It is represented as in Equation (12).

$$EFFS_{Feature} \;\; = \;\; Max\_Occurrence\,(Best\,(Select_{Best})) \tag{12}$$

2.3. Performance evaluation

2.3.1. Accuracy:

To evaluate the performance of a subset, Random Forest[14,15] is used as a classifier model. It is an ensemble classifier model that includes a number of decision trees and different subsets of a dataset and takes the average value to improve the predictive

accuracy of a given dataset. It can be defined in Equation (13).

$$\{h(m\ ,\ EFFS_{Feature}\ ),\ Feature=\ F_1, F_2 \ldots \ldots F_n,\} \tag{13}$$

Where m is a target feature and $EFFS_{Feature}$ input subset to a model. If a model contains more number of trees lead to higher accuracy and used to reduce over fitting problem in a datasets.

Steps:

Tree Creation:

1. Select m features randomly from $EFFS_{Feature}$ where $m \leq\ EFFS_{Feature}$
2. From m feature, calculate a best splitting node as a decision node d
3. Based on IG split d into number of nodes $(d_1\ ,\ d_2 \ldots\ d_n)$
4. Repeat step 1 to 3 until there is no node to split
5. Repeat step 1 to 4, n times to create n number of trees

Prediction:

1. Predict the outcome of each decision tree by running test data (80-20 % of train and test data)
2. Save the predicted target outcome
3. Perform voting on each target outcome
4. Print the highest voted target outcome as final predicted value.

2.3.2. Number of Features

Top ranked features are selected based on threshold and Pareto principle from existing and proposed algorithm.

2.3.3. Time Complexity

Time required to select features from existing and proposed features are calculated to show the performance.

Proposed Enhanced Filtrate Feature Selection (EFFS) algorithm:

The proposed EFFS algorithm uses multiple filtering methods (mutual information gain, Chi-square, correlation, and Fishore score filter methods). The proposed EFFS algorithm finds relevant features (based on feature score, rank, and rate of occurrences of each feature), which takes minimum time to select features with greater performance metrics.

3 Results and Discussion

The dataset details are given in Table 1, and downloaded from UCI, Kaggle, and Cancer imaging archives. The implementation of this work was done in Anaconda Python Jupyter Notebook.

Table 1. Dataset Details

S No	Dataset Name	No of Features	Sample Feature Names
1	Brain Tumour	46	id, age, gender, race, ethnicity, symptoms like seizures, vomiting, headache, head growth, motor movement, neurological sign etc.
2	Breast Cancer	31	id, radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, symmetry mean, and fractional dimension etc.
3	Lung Cancer	16	gender, age, smoking, yellow fingers, anxiety, peer pressure, chronic disease etc.
4	Diabetes	17	age, gender, polyuria, polydipsia, weakness, sudden weight loss etc.
5	Covid 19	61	gender, age, influenza vaccine, smoker, travel history, symptom duration, symptoms like sore throats, chills, cough, breath, sneeze etc.

Mutual_Info (S, A, Select $_{Feature}$) // Reduction of entropy
{
FOR each A in S
Calculate Expected Information
$Info(S) = -\sum_{i=1}^{m} p_s \log_2 p_s$ // p_s Probability, m distinct classes
Partition S into subset A, Calculate Information
Info $_A(S)$ = - $\sum_{i=1}^{m}$(|S_i / S|)*Info(S) // ($|S_i / S|$)weight of j th partition
Calculate Gain of A
$Gain(A) = Info(S) - Info_A(S)$
Select Optimal Feature
Neglect Gain (A) = = 0
$Select_{Feature} = (0.80 * Total\ No.of\ Feature)\ in\ (Gain(A))$
Return $Mutual\_Info(Select_{Feature})$
}
Chi-Square (S, A, Select $_{Feature}$) //Relationship between categorical attributes
{
State hypothesis H_0 and H_1
$H_0 \rightarrow Independent\ Features$
$H_1 \rightarrow Dependent\ Feature$
Calculate observed value by Contingency table
$DF = (R-1) * (C-1)$ //DF degrees of freedom, R rows, C columns
Calculate expected value E of a feature based on H_0
$E = N * P((A \cap B))$ //A,B independent features
Calculate Chi- Square Score
$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
$Select_{Feature} = (0.80 * Total\ No.of\ Feature)\ in\ X^2$
Return $Chi\_square(Select_{Feature})$
}
Heat_Map (S, A, Select $_{Feature}$) // Linear dependency
{
FOR $F_1, F_2 \ldots F_n$
Calculate Pearson's Correlation Coefficient
$PCC = \frac{\sum F_1 * F_2 - (\sum F_1) * (\sum F_2)}{n\sum F_1^2 - (\sum F_1)^2 * n\sum F_2^2 - n(\sum F_2)^2}$
$Sort(PCC)$
$Select_{Feature} = (0.80 * Total\ No.of\ Feature)\ in\ Sort(PCC)$
Return $Heat\_Map(Select_{Feature})$
}
F_Score (S, A, Select $_{Feature}$) // significance
{
FOR each A in S
Calculate Score for each feature
$F_{Score}(f_i) = \frac{s_b(f_i)}{\sum_{i=1}^{n} s_t^{(i)}(f_i)}$
Rank features by score
$Rank = Desc(F_{Score})$
$Select_{Feature} = (0.80 * Total\ No.of\ Feature)\ in$ Rank
Return $F\_Score(Select_{Feature})$
}
$Best(Select_{Best})$
$= Mutual\_Info(Select_{Feature}) \bigcup Chi\_square(Select_{Feature}) \bigcup Heat\_Map(Select_{Feature}) \bigcup F\_Score(Select_{Feature})$
$EFFS_{Feature} = Max\_Occurrence(Best(Select_{Best}))$ // EFFS Output

Fig 2.

3.1. Number of Features Selected:

Features are selected using the above Equation (6), Equation (7), Equation (8), and Equation (10)**.** The number of features is selected from various existing methods, and the proposed Enhanced Filtrate Feature Selection algorithms are executed in the Jupyter Notebook using Python and shown in Table 2. According to the results obtained, the proposed method, EFFS, selects the best optimal feature. Taking these selected optimal features as input to a predictive model for improving performance.

3.2. Time Complexity:

The time required to select features in MIG, CS, correlation heat maps, Fisher scores, and the EFFS algorithm is calculated in milliseconds for the datasets on breast cancer, lung cancer, brain tumors, COVID, and diabetes. The X axis represents algorithms,

Table 2. Enhanced Filtrate Feature Selection (EFFS)

Dataset	Total Number of Features	Existing Methods				EFFS
		Mutual Info	Chi-Squared	Heat Map	F-Score	
Brain Tumour	46	42	36	37	39	**39**
Breast Cancer	31	28	25	25	29	**29**
Lung Cancer	16	12	12	16	15	**15**
Diabetes	17	15	14	14	15	**15**
Covid 19	61	45	33	49	44	**43**

and the Y axis represents time to select features in milliseconds. Comparative time complexity analysis of the proposed EFFS shows minimal time to select the best optimal subset for the predictive model, as shown in Figure 3.

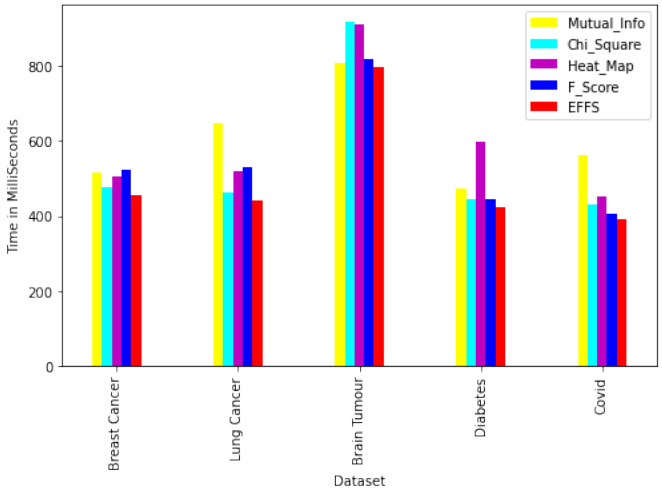

Fig 3. 'Comparative Result of Datasets based on existing and proposed Algorithms' The time required to select features in MIG, CS, correlation heat map, fisher score and EFFS algorithm is calculated in milli seconds and shown in Fig 4. For the datasets breast cancer, Lung cancer, Brain tumour, Covid and Diabetes. X axis represents Algorithms and Y axis represents time to select features in milli seconds

3.3. Accuracy:

Accuracy is one of the performance measures that is used to represent the number of hits based on overall values. A Random Forest classifier is used as a classifier model to predict the accuracy of the datasets taken. Equation (14) can be used to calculate accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + Fp + FN} \tag{14}$$

Whereas TP is correctly classified positive, TN is correctly classified negative, FP is incorrectly classified positive, and FN is incorrectly classified negative. Figure 4 represents a graphical view of accuracy as a performance metric for the datasets. In this, the X axis represents the dataset, and the Y axis shows the accuracy in percentage based on the features selected. The proposed EFFS method shows better accuracy than existing methods.

The previous work[18] proposed a Mixed Information Coefficient-based feature selection method to select optimal features in which IG was used and showed 80 % - 98% of accuracy on various datasets. In[19], filter–wrapper–based methods are used with KNN classifiers in which performance is evaluated using the number of features selected and obtains 89.06% of accuracy in filter methods. The framework for feature subset selection[20] uses a multi-step approach including chi square, Gini index, mutual information gain and symmetric uncertainty methods to select top ranked features and shows 98.31 % of accuracy using KNN classifier. In[21] demonstrate various filter-based approaches on the weka tool to select features and show 70 % to 98.34 % of accuracy on tree-based classifier models. Compared to other existing methods EFFS shows greater metrics shown in Table 3.

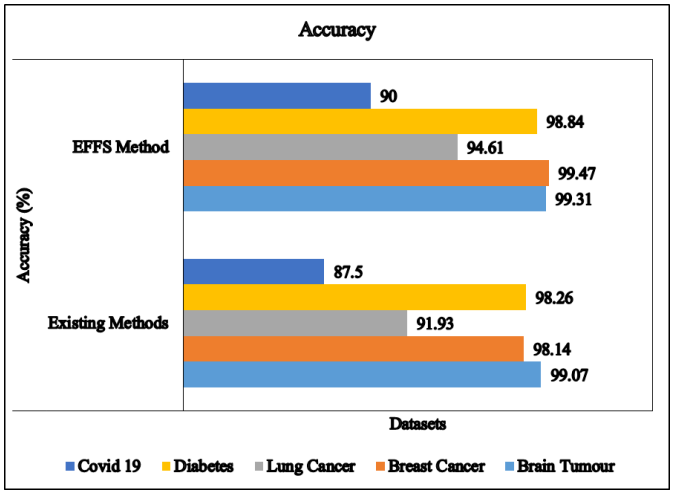

Fig 4. 'Accuracy of Existing Vs EFFS Algorithm' represents the graphical view of accuracy as a performance metric for the datasets. In this, the X axis represents the dataset and Y axis shows the accuracy in percentage based on features selected. The proposed EFFS method shows better accuracy than existing methods

Table 3. Comparative Analysis of EFFS Algorithm

Datasets	Total Number of Features	Number of Features Selected		Time taken to Select Features (in milliseconds)		Accuracy (%)	
		Existing Methods	EFFS Method	Existing Methods	EFFS Method	Existing Methods	EFFS Method
Brain Tumour	46	39-42	**39**	479-522	**456**	99.07	**99.31**
Breast Cancer	31	25-29	**29**	464-646	**442**	98.14	**99.47**
Lung Cancer	16	12-16	**15**	809-919	**795**	91.93	**94.61**
Diabetes	17	14-15	**15**	44-599	**425**	98.26	**98.84**
Covid 19	61	33-49	**43**	406-563	**392**	87.5	**90**

4 Conclusion

Features are most important in dataset construction. Datasets with relevant and non-redundant features are used to predict models with greater accuracy. Most of the existing methods use MR image data and gene expression as a feature. The Enhanced Filtrate Feature Selection algorithm selects the optimal best subset for dataset construction based on patient medical records. In the proposed algorithm, top-ranked features are selected and the rate of occurrences is calculated based on the threshold and Pareto principle. This paper concludes EFFS algorithm selects the best features such that 39 from 46 for brain tumor, 29 from 31 for breast cancer, 15 from 16 for lung cancer, 15 from 17 for diabetes, and 43 from 61 for covid 19 dataset. Highly informative features are used to build a model with greater accuracy and less complexity. The experimental result shows greater performance with the random forest classifier model. The EEFS algorithm also gives minimal time for selecting features such that it takes 456, 442, 795, 425, and 392 milliseconds for brain-tumor, breast cancer, lung cancer diabetes, and covid 19 datasets respectively. Compared to filter-based feature selection with an RF classifier model our EFFS algorithm with an optimal subset of features provide an accuracy of 99.31% for a brain tumor, 99.47% for breast cancer, 94.61% for lung cancer,98.84% for diabetes, and 90% for covid 19 dataset. An exhaustive evaluation method is used to measure the performance of EFFS approach. In the future, the result of EFFS is used as an input to the wrapper-based feature selection algorithm to obtain an optimal dataset for the classification model.

References

1) Usha P. An evaluation of feature selection methods performance for dataset construction. *Futuristic Communication and Network Technologies*;p. 115–143. Available from: https://doi.org/10.1007/978-981-19-8338-2_9.
2) Usha P. Feature Selection Techniques in Learning Algorithms to Predict Truthful Data. *Indian Journal of Science and Technology 2023*;16(10):744–755. Available from: https://doi.org/10.17485/IJST/v16i10.2102.
3) Gong H, Li Y, Zhang J, Zhang B, Wang X. A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information. *Eng Appl Artif Intell*. 2024;131:107865–107865. Available from: http://dx.doi.org/10.1016/j.engappai.2024.107865.
4) Khanna M, Singh LK, Shrivastava K, Singh R. An enhanced and efficient approach for feature selection for chronic human disease prediction: A breast cancer study. *Heliyon [Internet]*. 2024;10(5). Available from: http://dx.doi.org/10.1016/j.heliyon.2024.e26799.
5) Sekhar R, Sujatha P, B. Feature extraction and Independent Subset Generation using Genetic Algorithm for Improved Classification. *Int J Intell Syst Appl Eng*;11(2):503–515. Available from: https://ijisae.org/index.php/IJISAE/article/view/2660.
6) Rad A, Salomonsson M, Cenanovic K, Balague M, Raudberget H, Stolt D, et al. Correlation-based feature extraction from computer-aided design, case study on curtain airbags design. . *Computers in Industry*. 2022;138:103634–103634. Available from: http://dx.doi.org/10.1016/j.compind.2022.103634.
7) Varzaneh ZA, Orooji A, Erfannia L, Shanbehzadeh M. A new COVID-19 intubation prediction strategy using an intelligent feature selection and K-NN method. . *Informatics in Medicine Unlocked*;p. 100825–100825. Available from: http://dx.doi.org/10.1016/j.imu.2021.100825.
8) Macaulay BO, Aribisala BS, Akande SA, Akinnuwesi BA, Olabanjo OA. Breast cancer risk prediction in African women using Random Forest Classifier. . *Cancer Treatment and Research Communications*. 2021;28:100396–100396. Available from: http://dx.doi.org/10.1016/j.ctarc.2021.100396.
9) Newaz A, Ahmed N, Haq S, F. Survival prediction of heart failure patients using machine learning techniques. . *Informatics in Medicine Unlocked*;p. 100772–100772. Available from: http://dx.doi.org/10.1016/j.imu.2021.100772.
10) Hasan AM, Shin M, Das J, U. Yakin Srizon A. Identifying prognostic features for predicting heart failure by using machine learning algorithm. In: 2021 11th International Conference on Biomedical Engineering and Technology. ACM. 2021. Available from: https://doi.org/10.1145/3460238.3460245.
11) Bhat P, Dutta K. A multi-tiered feature selection model for android malware detection based on Feature discrimination and Information Gain. *J King Saud Univ*. 2022;34(10):9464–77. Available from: http://dx.doi.org/10.1016/j.jksuci.2021.11.004.
12) Jung J, Park J, Je C. Feature engineering and evaluation for android malware detection scheme. *Journal of Internet Technology*;22(2):423–440. Available from: https://jit.ndhu.edu.tw/article/view/2500.
13) Sun L, Wang T, Ding W, Xu J, Lin Y. Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification. 2021;578:887–912. Available from: http://dx.doi.org/10.1016/j.ins.2021.08.032.
14) Gárate-Escamila AK, Hassani HE, Andrès A, E. Classification models for heart disease prediction using feature selection and PCA. 2020;19:100330–100330. Available from: http://dx.doi.org/10.1016/j.imu.2020.100330.
15) Farid A, A. A composite hybrid feature selection learning-based optimization of genetic algorithm for breast cancer detection. *Proceedings of The 2nd International Conference on Advanced Research in Applied Science and Engineering GLOBALKS; 2020*. Available from: https://doi.org/10.33422/2nd.rase.2020.03.99.
16) Niyas M, Thiyagarajan. Feature selection using efficient fusion of Fisher Score and greedy searching for Alzheimer's classification. 2022;34:4993–5006. Available from: http://dx.doi.org/10.1016/j.jksuci.2020.12.009.
17) Srivenkatesh M. Prediction of Breast Cancer Disease Using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*. 2020;9. Available from: http://dx.doi.org/10.35940/ijitee.D1866.029420.
18) Qi X, Song J, Qi H, Shi Y. Maximum Information Coefficient Feature Selection Method for Interval-Valued Data. *IEEE Access*. 2024;12:53752–53766. Available from: https://dx.doi.org/10.1109/access.2024.3387978.
19) Vommi AM, Battula TK. A hybrid filter-wrapper feature selection using Fuzzy KNN based on Bonferroni mean for medical datasets classification: A COVID-19 case study. p. 119612–119612. Available from: http://dx.doi.org/10.1016/j.eswa.2023.119612.
20) Mandal AK, Nadim MD, Saha H, Sultana T, Hossain MD, Huh EN. Feature subset selection for high-dimensional, low sampling size data classification using ensemble feature selection with a wrapper-based search. *F*. 2024;12:62341–57. Available from: http://dx.doi.org/10.1109/access.2024.3390684.
21) Parmar M, Sonker A, Sejwar V. A comparative analysis for filter-based feature selection techniques with tree-based classification. *International Journal on Recent and Innovation Trends in Computing and Communication*. 2023;11:360–369. Available from: https://doi.org/10.17762/ijritcc.v11i10s.7643.