Educational Administration: Theory and Practice
2024, 30(11), 1387-1394
ISSN: 2148-2403
https://kuey.net/

Research Article

Enhanced Predictive Modeling For Lung Cancer Using Advanced Relief Feature Selection (ARFS)

Kanimozhi V A[1*], Dr. V. Krishnapriya[2]

[1*]Research Scholar,Department of Computer Science,Sri Ramakrishna College of Arts &Science,Coimbatore, Tamilnadu, India. Email: vakanimozhipsg@gmail.com
[2]Associate Professor & Head,Department of Computer Science with Cognitive System,Sri Ramakrishna College of Arts & Science,Coimbatore, Tamilnadu, India.Email: kp@srcas.ac.in

***Corresponding Author:** Kanimozhi V A
*Email: vakanimozhipsg@gmail.com

Citation:Kanimozhi V A et al. (2024).Enhanced Predictive Modeling For Lung Cancer Using Advanced Relief-FRFE Feature Selection (ARFS)*Educational Administration: Theory and Practice*, 30(11) 1387-1394
Doi: 10.53555/kuey.v30i11.9465

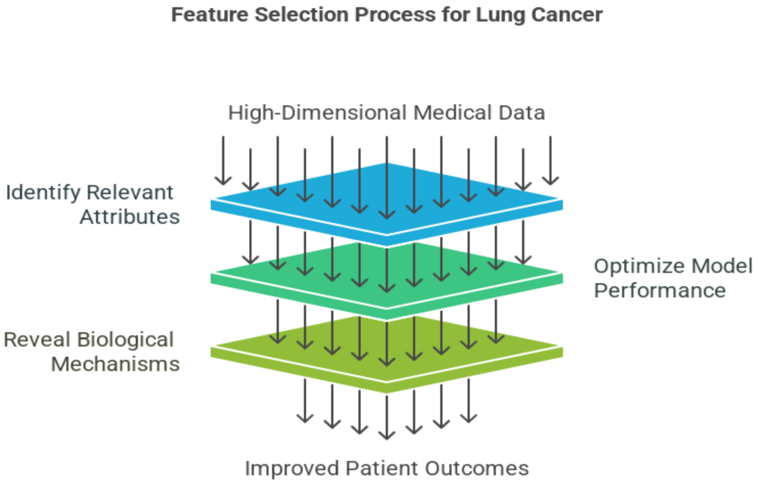
ARTICLE INFO **ABSTRACT**

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Early detection significantly improves survival rates.This paper proposes an Advanced ReliefF-RFE Feature Selection (ARFS) methodology tailored for lung cancer research, aiming to optimize feature selection for predictive modeling. Leveraging the Improved ReliefF algorithm and Recursive Feature Elimination (RFE), the method iteratively evaluates and refines feature subsets, enhancing predictive model performance. ReliefF initially assesses feature importance based on their discriminatory power, while RFE iteratively eliminates the least significant features. The proposed methodology offers a structured approach to feature selection, contributing to the effectiveness, efficiency, and interpretability of predictive models in lung cancer research. Experimentation validates the efficacy of ARFS in optimizing feature subsets, thus enhancing predictive accuracy in lung cancer prediction tasks.

Keywords: Recursive Feature Elimination, Feature Selection, ReliefF-RFE, Lung Cancer Prediction, Machine Learning;

1. Introduction

Lung cancer poses a significant global health challenge, primarily due to its high mortality rate, which is often a result of late diagnosis and complex causes. Factors contributing to lung cancer include smoking, exposure to carcinogens in certain occupations, air pollution, and genetic predispositions. Effective treatment and better patient outcomes rely heavily on early detection and accurate diagnosis. However, the disease's heterogeneous nature and its progression complicate the diagnostic process. This complexity underscores the need for advanced methodologies to pinpoint critical predictive features from extensive and diverse datasets. Data mining and machine learning techniques offer considerable promise in addressing these challenges. By leveraging these technologies, it is possible to enhance diagnostic accuracy and improve the identification of key factors influencing lung cancer, ultimately leading to more effective treatment strategies and better patient outcomes. Feature selection is a critical component of data mining, focused on identifying the most relevant attributes in a dataset that significantly influences the prediction or classification of a target variable. By optimizing machine learning model performance, effective feature selection reduces dimensionality, prevents overfitting, and enhances generalization. For lung cancer, feature selection is especially vital due to the high-dimensional nature of medical data, which includes genomic, proteomic, and imaging information. The figure below illustrates the feature selection process for lung cancer prediction.


Feature Selection Process for Lung Cancer
High-Dimensional Medical Data
Identify Relevant Attributes
Optimize Model Performance
Reveal Biological Mechanisms
Improved Patient Outcomes

This process not only helps to isolate the most impactful features but also reveals underlying biological mechanisms and risk factors associated with lung cancer. By improving the efficiency and insightfulness of machine learning models, feature selection contributes to more accurate diagnostics and more effective treatment strategies, ultimately leading to better patient outcomes.

2. Literature Review

2.1Christo VE et.al proposed Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest. The proposed framework for a Clinical Decision Support System (CDSS) employs a co-operative co-evolution approach to improve system performance by addressing Feature Selection (FS) and Instance Selection (IS) as independent sub-problems. By treating FS and IS separately, the system effectively removes less relevant features and instances, thereby enhancing overall performance. This method utilizes a wrapper approach combined with a random forest classifier for both FS and IS. The proposed framework achieved accuracy rates of 97.1%, 82.3%, 81.01%, 93.4%, 96.8%, 91.4%, and 72.2% for the WDBC, Hepatitis, PID, CHD, SHD, Vertebral Column, and HCC datasets, respectively. These results demonstrate that the CDSS developed using the co-operative co-evolution approach can effectively support physicians in decision-making processes.
2.2P. Liu et.al proposed Prediction of Second Primary Lung Cancer Patient’s Survivability Based on Improved Eigenvector Centrality-Based Feature Selection. Improved eigenvector centrality-based feature selection (IECFS) is then used to preprocess the data set. The ranking criteria as well as intraclass and interclass dispersions are used in the IECFS approach. The approach attains optimal performance by varying the parameter's value and the quantity of characteristics chosen.There are five folds in the experiment. In comparison to the original classification accuracy (89.16%) and other cutting-edge feature selection approaches, our method produces a prediction accuracy of 90.998% for the five-year survivability. The suggested strategy produces a forecast accuracy of 83.16% for the three-year survival, marginally surpassing all of the examined methods. The approach works well and can be applied broadly.
2.3 B. N. Narayanan et.al proposed Performance Analysis of Feature Selection Techniques for Support Vector Machine and its Application for Lung Nodule Detection. A feature-based classifier plus a candidate detector make up a typical CAD system. In this work, they examine and investigate the Support Vector Machine's (SVM) performance using a huge feature set. They examine how well SVM performs in relation to the quantity of features. In comparison to classical classifiers, our findings show that SVM is less prone to over-training, more computationally rapid, and resilient with a large feature set. Results for the Lung Nodule Analysis 2016 dataset are made available to the public. This 10-fold validation findings show that the SVM-based classification approach performs 14.8% better than the Fisher linear discriminant classifier.
2.4M. Sobhan et.al proposed Deep Learning to Discover Genomic Signatures for Racial Disparity in Lung Cancer. Identifying genomic signatures that differentiate lung cancers between these groups can help understand this disparity. We analyzed gene expression profiles from whole blood samples of AAM and EAM patients using a deep learning-based unsupervised feature selection approach, the concrete autoencoder (CAE). Given the highly imbalanced nature of the dataset (AAM: 15, EAM: 153). In contrast, the CAE isolated 34 key genes capable of making this distinction. Using these genes, a Random Forest classifier achieved more accuracy, with only one false negative. These key genes can serve as biomarkers to comprehend the differences in lung cancer development between AAM and EAM, highlighting CAE's effectiveness in extracting relevant features from imbalanced datasets.

2.5W. Zhang et.al proposed Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer. In order to choose the informative genes by unsupervised feature selection, the suggested approach combined linear discriminant analysis, adaptive structure preservation, and l2, 1-norm sparse regression into a joint learning framework. In the suggested solution, an efficient algorithm was created to address the optimization problem. In order to lower the computational cost, we also carried out module-based gene filtering prior to feature selection. Using an NSCLC gene expression dataset from The Cancer Genome Atlas (TCGA), we assessed the suggested methodology. The experimental findings demonstrate that a limited set of gene signatures were found by the suggested strategy for precise NSCLC subtype determination. A summary of the major biological processes was also done in order to do enrichment analysis on the detected gene signatures.

3. Research Methodology

The research methodology combines the Improved ReliefF algorithm with the RFE algorithm to achieve an optimal subset of features for lung cancer prediction. Initially, Improved ReliefF is employed to evaluate the importance of each feature based on its ability to distinguish between instances of different classes. The weights of the features are updated iteratively, with features that contribute more to the class discrimination process receiving higher weights.Subsequently, the RFE algorithm is applied to further refine the feature selection process. RFE iteratively removes the least significant features based on their importance scores, as determined by Improved ReliefF. This iterative elimination process continues until the optimal subset of features is identified, maximizing the predictive performance of the machine learning model.The proposed feature selection method combines the Improved ReliefF algorithm with the Recursive Feature Elimination (RFE) algorithm to enhance the selection of relevant features for lung cancer prediction.
The ARFS method integrates the ReliefF and Recursive Feature Elimination (RFE) techniques for optimal feature selection in lung cancer prediction.

Input the Dataset

First, we need to have a dataset X with features $F = \{f_1, f_2, \dots, f_n\}$ and corresponding labels y. Ensure the dataset is cleaned and ready for feature selection.

ReliefF Feature Evaluation

The ReliefF algorithm evaluates the importance of each feature through the following steps:

Initialize Feature Weights:$W(f_i) = 0\, for\ all\ features\ f_i \epsilon\ F$
Identify Nearest Hits and Misses: For each instance x_i in the dataset:
Find the nearest hit H_i (nearest instance of the same class).
Find the nearest miss M_i (nearest instance of a different class).

Update Feature Weights

The weight update for each feature f_k based on the nearest hit and miss is:

$$W(f_k) = W(f_k) - \frac{1}{m}\sum_{i=1}^{m}(\delta(f_k, x_i, H_i) \sum m - \delta(f_k, x_i, M_i))$$

Where $\delta(f_k, x_i, y)$ is a function that measures the difference between the feature$f_k$in instance x_i and its nearest hit or miss y.

Normalize Feature Weights

Normalize the feature weights to ensure they are within a specific range or sum up to one.
Select Top N Features: Select the top N features with the highest weights for further analysis:

$$F_{ReliefF} = \{f_1, f_2, \dots, f_n\}$$

Where $W(f_i)$ are among the highest.

RFE with Filtered Features

Using the filtered features $F_{ReliefF}$ obtained from ReliefF as input for RFE, the process proceeds as follows:
Train a Model: Select a suitable model (e.g., support vector machine, logistic regression, or a tree-based model) and train it using the filtered feature set. Let the model be represented by M.
Rank Features Based on Importance: Calculate the importance score for each feature f_k in $F_{ReliefF}$ based on the model M:

$$F_{RFE} = F_{ReliefF}\{f_{least\ important}\}$$

Remove Least Important Features: Identify the least important feature(s) based on the importance scores and remove them:

$$F_{RFE} = F_{ReliefF} \setminus \{f_{least\ important}\}$$

Iterative Elimination: Retrain the model with the updated feature set and repeat the ranking and elimination process iteratively:

$$F_{RFE}^{(t+1)} = F_{RFE}^{(t)} \setminus \{f_{least\ important}\}$$

Continue this process until a predefined stopping criterion is met, such as achieving a specific number of features or a performance plateau.

Model Training and Evaluation

After obtaining the final set of features F final through the integrated ReliefF and RFE process, proceed with the following steps:

Train the final Model:

Train the final predictive model using the selected features F_{final}:

$$\mathcal{M}_{final} = Train(X_{Ffinal}, y)$$

Evaluate Model Performance:

Evaluate the model performance using appropriate metrics such as accuracy, precision, recall, and F1 score on a validation set:

$$Performance\ (M_{final}) = metrics\ (X_{val}, y_{val})$$

This ensures that the selected features positively contribute to the models predictive capability.

Algorithm: Advanced ReliefF Feature Selection (ARFS)

```
Input:
$X$: Cleaned dataset
$y$: Corresponding labels
$N$: Number of top features to select from ReliefF
$K$: Number of features to select after RFE
$M$: Chosen model for RFE (e.g., SVM, logistic regression)
Output:
$F_{final}$: Final selected features
Initialize feature weights $W(f_i) = 0$ for all features$f_i \in X$.
For i =1 to $x_i$ do
    For j=1 to $x_i$ do
        Find the nearest hit $H_i$ (same class).
        Find the nearest miss $M_i$ (different class).
    End
    For all hits and misses do
        # Update Feature Weights:
        For each instance $x_i$ do
            For each feature $f_k$ do
                $W(f_k) = W(f_k) - \frac{1}{m}\sum_{i=1}^{m}(\delta(f_k, x_i, H_i)\sum m - \delta(f_k, x_i, M_i))$
            End
            Normalize the feature weights $W(f_i)$.
        End
    Select the top N features with the highest weights:
        $F_{ReliefF} = \{f_1, f_2, f_3 \ldots\ldots f_N,\}$ where $W(f_i)$ are among the highest
End
Use the filtered features $F_{ReliefF}$ from ReliefF as the input for RFE.
Train the model $\mathcal{M}$ using the filtered feature set $F_{ReliefF}$.
Calculate importance scores for each feature$f_k$: $Importance(f_k) = score_k$
Remove the least important feature(s) based on the importance scores:
    $F_{RFE} = F_{ReliefF} \setminus \{f_{least\ important}\}$
Retrain the model with the updated feature set and repeat the ranking and elimination process until K features remain or a performance plateau is reached.
The final feature set $F_{final}$ is the set of K features remaining after the RFE process.
Train the final model using $F_{final}$:
    $\mathcal{M}_{final} = Train(X_{Ffinal}, y)$
Return$F_{final}$ as a selected fetures.
```

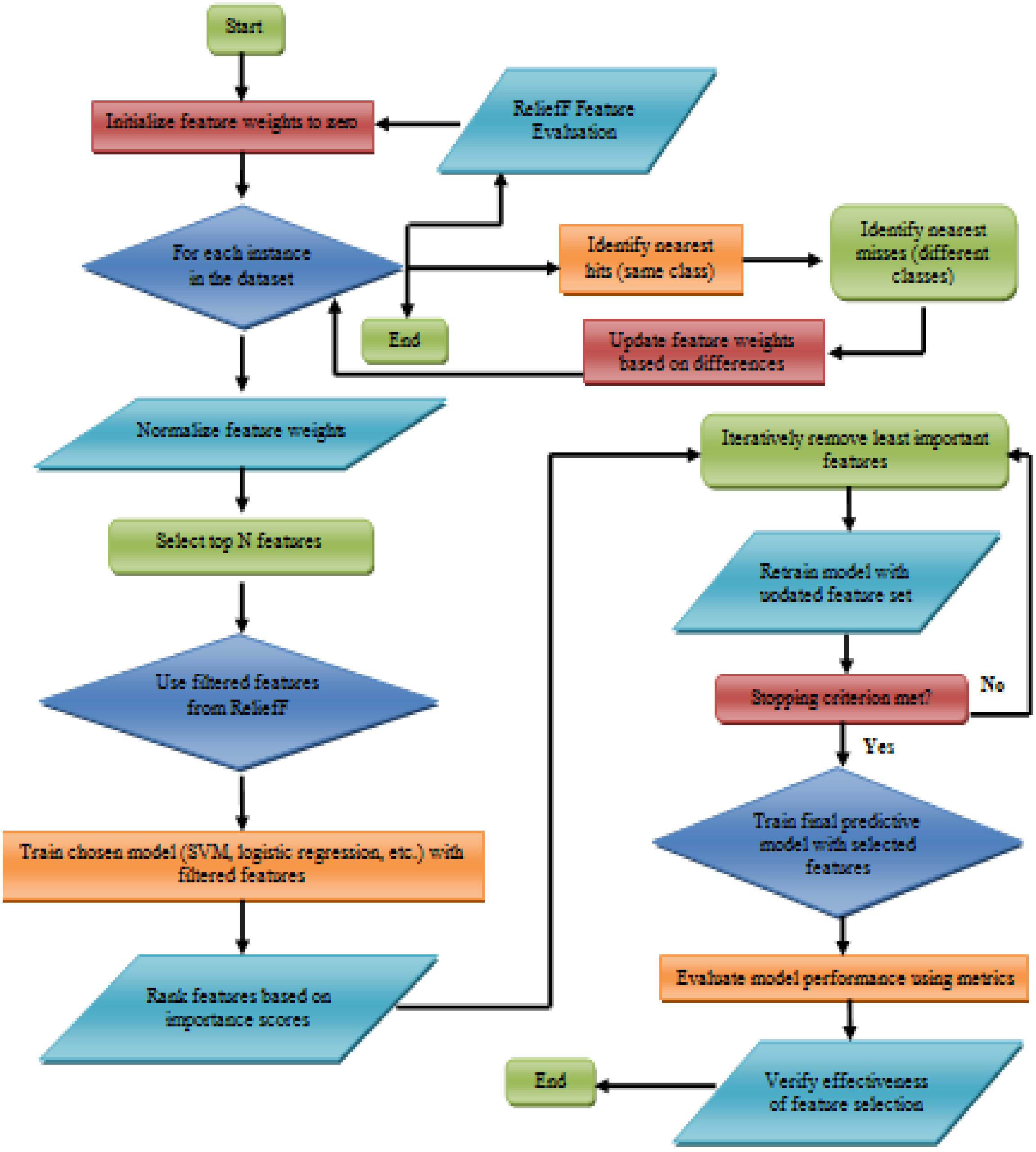

Start
Initialize feature weights to zero
ReliefF Feature Evaluation
For each instance in the dataset
Identify nearest hits (same class)
Identify nearest misses (different classes)
End
Update feature weights based on differences
Normalize feature weights
Select top N features
Use filtered features from ReliefF
Train chosen model (SVM, logistic regression, etc.) with filtered features
Rank features based on importance scores
Iteratively remove least important features
Retrain model with updated feature set
Stopping criterion met?
No
Yes
Train final predictive model with selected features
Evaluate model performance using metrics
Verify effectiveness of feature selection
End

Figure 1.*Advanced ReliefF Feature Selection (ARFS)* Flow Chart

This Figure 1 flow chart outlines the process of using the Advanced ReliefF-RFE Feature Selection (ARFS***)*** algorithm for feature selection, including steps for evaluating, normalizing, and selecting the most important features for model training and performance evaluation.

4. Experimental Results

4.1 Precision

Precision is a measure of how well a model can predict a value based on a given input. The precision of a model is the ratio of true positive predictions to all positive predictions.

$$\boldsymbol{Precision} = \frac{\boldsymbol{true\ positive}}{(\boldsymbol{true\ positive}\ +\ \boldsymbol{false\ positive})}$$

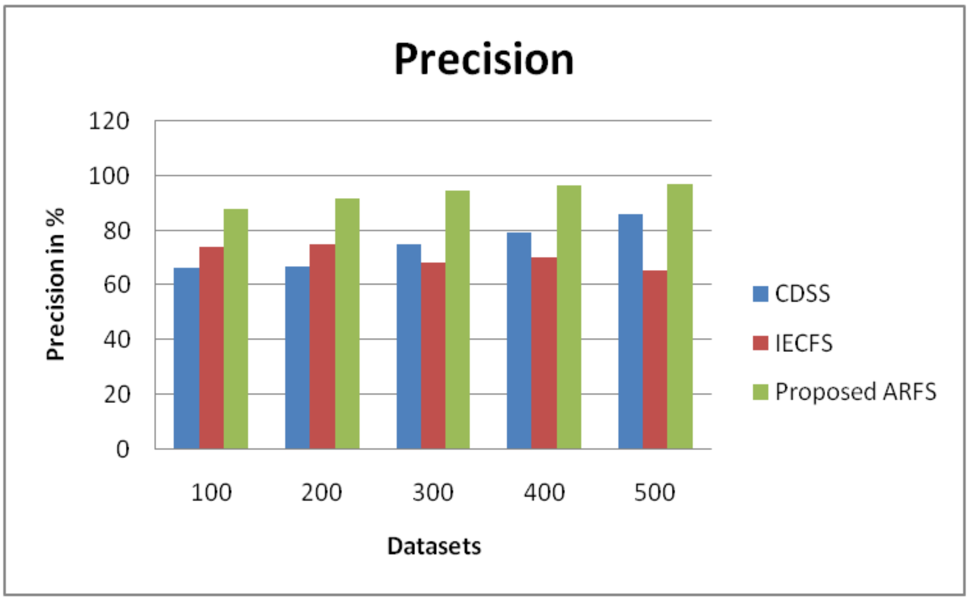


Figure 2.Comparison chart of Precision

The Figure 2 Shows the comparison chart of Precision demonstrates the existing IECFS, CDSS and Proposed ARFS. X axis denote the Datasets and y axis denotes the Precision ratio. The proposed values are better than the existing algorithm. The existing algorithm values start from 66.45 to 86.86, 65.33 to 74.12 and Proposed ARFS values starts from 87.76 to 97.12. The proposed method provides the great results.

4.2 Recall

Recall is a measure of a model's ability to correctly identify positive examples from the test set:

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Positives}}{(\boldsymbol{True\ Positives} + \boldsymbol{False\ Negatives})}$$

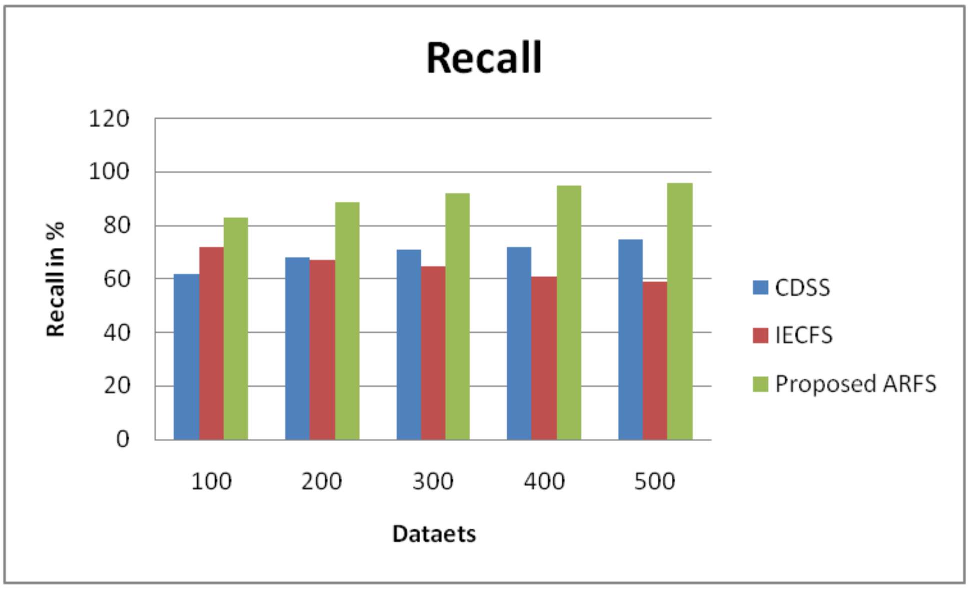


Figure 3.Comparison chart of Recall

The Figure 3 Shows the comparison chart of Recall demonstrates the existing IECFS, CDSS and Proposed ARFS. X axis denote the Datasets and y axis denotes the Recall ratio. The proposed values are better than the existing algorithm. The existing algorithm values start from 62 to 75, 59 to 72 and Proposed ARFS values starts from 83 to 96. The proposed method provides the great results.

4.3 F-Measure

F-measure is a test's accuracy that combines precision and recall. It is calculated by taking the harmonic mean of precision and recall.

$$\boldsymbol{F - Measure} = \frac{(2 * \boldsymbol{Precision} * \boldsymbol{Recall})}{(\boldsymbol{Precision} + \boldsymbol{Recall})}$$

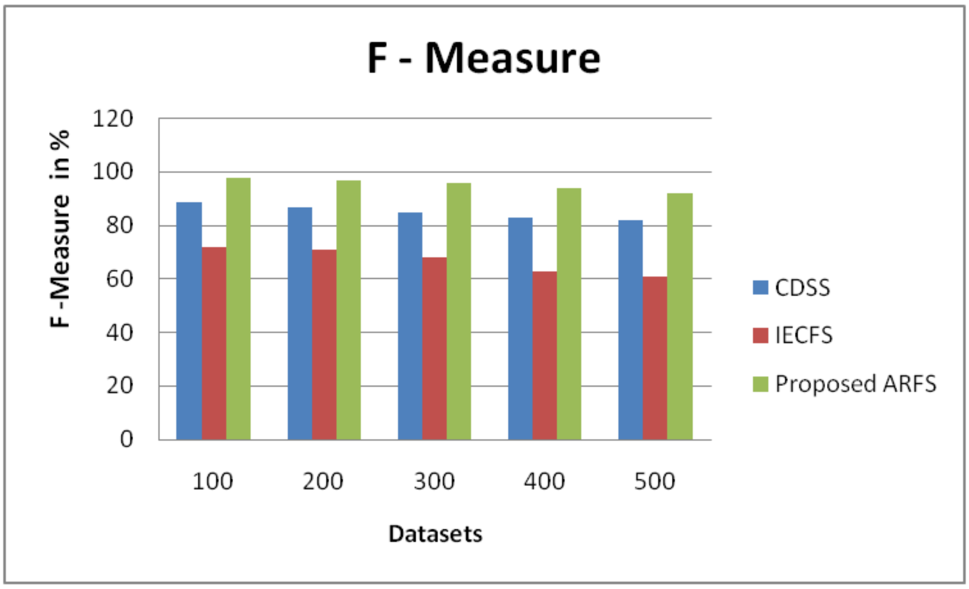


Figure 4.Comparison chart of F -Measure

The Figure 4 Shows the comparison chart of F -Measure demonstrates the existing IECFS, CDSS and Proposed ARFS. X axis denote the Datasets and y axis denotes the F -Measure ratio. The proposed values are better than the existing algorithm. The existing algorithm values start from 82 to 89, 61 to 72 and Proposed ARFS values starts from 92to 98. The proposed method provides the great results.

4.4 Accuracy

Accuracy is the degree of closeness between a measurement and its true value. The formula for accuracy is:

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{true\ value} - \boldsymbol{measured\ value})}{\boldsymbol{true\ value}} * \mathbf{100}$$

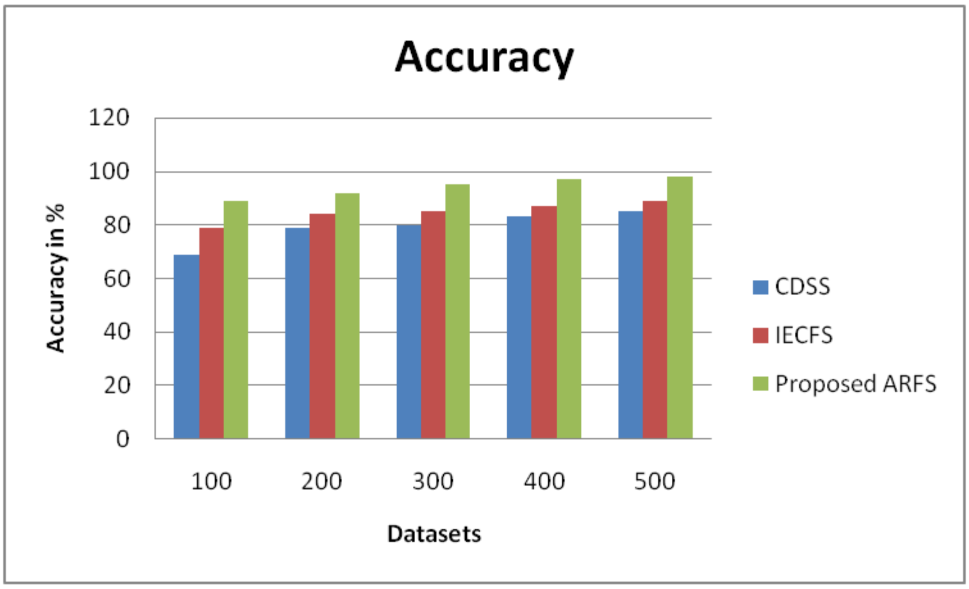


Figure 5.Comparison chart of Accuracy

The Figure 5 Shows the comparison chart of Accuracy demonstrates the existing IECFS, CDSS and Proposed ARFS. X axis denote the Dataset and y axis denotes the Efficiency Measure ratio. The proposed values are better than the existing algorithm. The existing algorithm values start from 69 to 85, 79 to 89 and Proposed ARFS values starts from 89 to 98. The proposed method provides the great results.

5. Conclusion

In this paper, the Advanced ReliefF-RFE Feature Selection (ARFS) methodology presents a robust approach to optimizing feature selection for lung cancer prediction. By leveraging the Improved ReliefF algorithm and Recursive Feature Elimination (RFE), ARFS effectively identifies and refines feature subsets, enhancing predictive model performance. The structured approach offered by ARFS contributes to the effectiveness, efficiency, and interpretability of predictive models in lung cancer research. Experimental validation demonstrates the efficacy of ARFS in optimizing feature subsets, thereby improving predictive accuracy in lung cancer prediction tasks. Overall, ARFS holds promise for advancing research in lung cancer prediction and potentially other biomedical applications requiring feature selection for predictive modeling.

References

1. Christo VE, Nehemiah HK, Brighty J, Kannan A. Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest. IETE Journal of Research. 2022 Jul 4;68(4):2508-21.
2. B. N. Narayanan, R. C. Hardie and T. M. Kebede, "Performance Analysis of Feature Selection Techniques for Support Vector Machine and its Application for Lung Nodule Detection," NAECON 2018 - IEEE National Aerospace and Electronics Conference, Dayton, OH, USA, 2018, pp. 262-266, doi: 10.1109/NAECON.2018.8556669.
3. P. Liu, K. Jin, Y. Jiao, M. He and S. Fei, "Prediction of Second Primary Lung Cancer Patient's Survivability Based on Improved Eigenvector Centrality-Based Feature Selection," in IEEE Access, vol. 9, pp. 55663-55672, 2021, doi: 10.1109/ACCESS.2021.3063944.
4. M. Sobhan, A. A. Mamun, R. B. Tanvir, M. J. Alfonso, P. Valle and A. M. Mondal, "Deep Learning to Discover Genomic Signatures for Racial Disparity in Lung Cancer," 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), 2020, pp. 2990-2992, doi: 10.1109/BIBM49941.2020.9313426.
5. X. Ye, W. Zhang and T. Sakurai, "Adaptive Unsupervised Feature Learning for Gene Signature Identification in Non-Small-Cell Lung Cancer," in IEEE Access, vol. 8, pp. 154354-154362, 2020, doi: 10.1109/ACCESS.2020.3018480.
6. D. Balasubramanian, P. Srinivasan and R. Gurupatham, "Automatic Classification of Focal Lesions in Ultrasound Liver Images using Principal Component Analysis and Neural Networks," 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France, 2007, pp. 2134-2137, doi: 10.1109/IEMBS.2007.4352744.
7. P. Chhabra and R. Madaan, "Data mining concepts in healthcare with discussion on prediction of diseases," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022, pp. 71-77, doi: 10.1109/COM-IT-CON54601.2022.9850851.
8. M. Y. SyarifahAdilah, R. Abdullah and I. Venkat, "ABC algorithm as feature selection for biomarker discovery in mass spectrometry analysis," 2012 4th Conference on Data Mining and Optimization (DMO), Langkawi, Malaysia, 2012, pp. 67-72, doi: 10.1109/DMO.2012.6329800.
9. P. Klemm et al., "3D Regression Heat Map Analysis of Population Study Data," in IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, pp. 81-90, 31 Jan. 2016, doi: 10.1109/TVCG.2015.2468291.
10. M. Sharma and R. Parveen, "A Comparative Study of Data Mining, Digital Image Processing and Genetical Approach for Early Detection of Liver Cancer," 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, 2020, pp. 687-692, doi: 10.1109/GUCON48875.2020.9231193.
11. F. Pinheiro, M. -H. Kuo, A. Thomo and J. Barnett, "Extracting association rules from liver cancer data using the FP-growth algorithm," 2013 IEEE 3rd International Conference on Computational Advances in Bio and medical Sciences (ICCABS), New Orleans, LA, USA, 2013, pp. 1-1, doi: 10.1109/ICCABS.2013.6629208.
12. P. R. Anisha, C. K. K. Reddy and L. V. N. Prasad, "A pragmatic approach for detecting liver cancer using image processing and data mining techniques," 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, India, 2015, pp. 352-357, doi: 10.1109/SPACES.2015.7058282.
13. N. Ramkumar, S. Prakash, S. A. Kumar and K. Sangeetha, "Prediction of liver cancer using Conditional probability Bayes theorem," 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2017, pp. 1-5, doi: 10.1109/ICCCI.2017.8117752.
14. J. Yang, Y. Wen, G. Zhao and J. Duan, "Research on Association Rules of Breast Cancer and TCM : Syndrome Based on Data Mining," 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 2019, pp. 2788-2792, doi: 10.1109/SSCI44817.2019.9003024.
15. J. K. Chiang and R. Chi, "Comparison of Decision Tree J48 and CART in Liver Cancer Symptom with CARNEGIE-MELLON UNIVERSITY Data," 2022 IEEE 4th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), Tainan, Taiwan, 2022, pp. 28-31, doi: 10.1109/ECBIOS54627.2022.9945039.