International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Multi-Class Cancer Classification with SVM Using Wrapper Forward and Backward Feature Selection for Dimension Reduction

May Myat Myat Khaing[a*] , May Mar Oo[b]

[a]Faculty of Computer Science,University of Computer Studies,Yangon, Myanmar

[b]Faculty of Information and Communication Technology,University of Technology (Yatanarpon Cyber City)

Pyin Oo Lwin, Myanmar

[a]Email: mmyatm@gmail.com

[b]Email: maymaroo@gmail.com

Abstract

The use of machine learning (ML) into healthcare has shown enormous growth in recent years. The efficacy of supervised ML models is significantly influenced by the quality of the training data. Feature selection is a crucial factor that affects the performance of machine learning models, especially in complex tasks like multi-class cancer classification. This research investigates the efficacy of using forward feature selection and backward feature elimination approaches in combination with logistic regression. The features generated using these approaches are then used for cancer type classification using support vector machines (SVM).The focus of our study is to use a partially complete gene dataset obtained from the Indian Council of Medical study (ICMR) for the purpose of classifying different types of cancer using Support Vector Machines (SVM). Our approach demonstrated a remarkable success rate of 96% when using features selected via the forward selection method and 97% when using features obtained through the backward selection method in multi-class cancer classification.

Keywords: Cancer type detection; gene expression data; Minimum Redundancy Maximum Relevance; wrapper-based feature selection method; forward feature selection; backward feature elimination; ICMR; SVM; logistic regression.

Received: 4/23/2024
Accepted: 6/23/2024
Published: 7/1/2024

----

* Corresponding author.

1. Introduction

Machine learning (ML) has emerged as a powerful tool in healthcare, aiding in tasks such as disease diagnosis, prognosis, and treatment prediction. However, the quality of training data significantly influences the performance of ML models, particularly in supervised learning scenarios. Feature selection plays a crucial role in enhancing model performance by identifying relevant features and reducing computational complexity. In the context of multi-class cancer classification, identifying important features poses a significant challenge due to the complexity and heterogeneity of cancer types.

The imbalanced nature of cancer datasets, where certain types have fewer instances, poses a challenge. Biased models may excel in majority classes while performing poorly in minority classes. Techniques such as oversampling, under sampling, or class weights can address this issue, leading to a more balanced model. Feature extraction and selection, incorporating domain knowledge, dimensionality reduction, and feature engineering, contribute to improved cancer classification model performance.

Interpretability is crucial in medical diagnostics, and logistic regression provides a clear probabilistic explanation of how each variable impacts the probability of a specific cancer type. Gaining the confidence of healthcare practitioners and facilitating the integration of machine learning into clinical decision-making processes is achievable through this interpretability.

Our research focuses on evaluating two feature selection methods, namely forward feature selection and backward feature elimination, using logistic regression as part of the classification process. We aim to address the challenges associated with multi-class cancer classification by leveraging these techniques on an incomplete gene dataset obtained from the Indian Council of Medical Research (ICMR). The ultimate goal is to develop a robust methodology that can accurately classify cancer types using Support Vector Machine (SVM) with high predictive performance.

This paper is structured as follows: This paper's introduction and related works are described in sections I and II. Section III contains information about the dataset and the proposed system design. The experimental results and conclusions are addressed in sections IV and V.

2. Related Work

All Gene expression refers to the mechanism through which DNA information is transcribed into instructions for synthesizing proteins or other molecules. Transcription of DNA leads to the formation of messenger RNA (mRNA), and subsequently, proteins are generated through the process of translation. Gene expression analysis is employed to ascertain the sequence of genetic changes occurring in a tissue or a single cell under specific conditions [1]. This analytical process involves quantifying the DNA transcripts in a sample of cells or tissue to identify the expressed genes and their respective quantities. One step in evaluating gene expression involves comparing sequenced reads, representing the number of base pairs in a DNA fragment, to a known genomic or transcriptome reference.

Analyzing gene expression involves the application of computational methods to comprehend gene regulation and their roles in tissue and cell functions. Machine learning (ML) approaches are commonly employed to gain insights into how genetic variations and regulatory regions contribute to phenotypic changes, including traits, wellness, and health [2,3]. While traditional ML methods like Decision Trees and Support Vector Machines were initially prevalent in computational gene expression analysis, the past decade has witnessed a surge in the prominence of deep learning (DL) techniques. Specifically, DL-based methods have gained traction for predicting the structure and function of genomic components such as promoters, enhancers, and gene sequences [4,5].

In [6], a pioneering RFODL-MGEC model was introduced for the classification of microarray gene expressions. This innovative RFODL-MGEC model primarily incorporated the RFO-FS technique to extract an optimal subset of features. Subsequently, the BCDNN model was applied for data classification, and the parameters of the BCDNN technique were finely tuned using a CGO algorithm. Extensive experiments conducted on benchmark datasets demonstrated that the RFODL-MGEC model exhibited superior performance in subtype classifications. Consequently, the RFODL-MGEC model proved to be effective in identifying different classes within high-dimensional and small-scale microarray data.

Conditional mutual information maximization (CMIM) and adaptive genetic algorithms (AGA) were combined by Hukla and his colleagues. [7] to create a unique hybrid framework known as CMIMAGA. To find important biomarkers in gene expression data, this approach is used. AGA was the wrapper strategy that was used to pick highly discriminating genes, while CMIM was used as a filter to remove unnecessary genes.

Feature engineering is a crucial aspect of computational methods designed for gene expression analysis. It addresses the inherent challenge posed by high dimensionality and the relatively limited number of samples in gene expression data.

In this study, a stacking ensemble deep learning model was employed to categorize multiple classes, specifically five types of cancer. The LASSO regularization method was utilized for feature selection, and the analysis was conducted on the Pan-Cancer Atlas dataset[8].

The importance of accurately classifying phenotypes by selecting a concise gene subset from extensive microarray data. Traditional methods, based on varying gene expressions, often yield redundant feature sets. The proposed solution is the Minimum Redundancy-Maximum Relevance (MRMR) feature selection framework, aiming to reduce duplication while enhancing relevance. Genes selected through MRMR exhibit a broader coverage of phenotypic traits and genomic space balance. Across six gene expression datasets, extensive tests demonstrate that MRMR-selected genes consistently improve class predictions, irrespective of the classification technique used. The text also mentions additional details, including the top 60 MRMR genes for each dataset [9].In this study, introduce an innovative framework designed for the classification of cancer types using gene expression levels. This study presents a comprehensive framework designed to enhance the accuracy and efficiency of cancer type classification, with significant implications for both biomedical research and clinical practice. The framework encompasses several crucial steps, beginning with dimension reduction techniques implemented through wrapper-based methods. Both forward and backward wrapper-based approaches are utilized, leveraging a logistic regression machine learning classifier.The forward wrapper-based technique incrementally adds features to the model and evaluates their impact on classification performance. Conversely, the backward wrapper-based method systematically removes features to improve model efficiency. These techniques streamline feature selection and optimize the model for better predictive accuracy.In the classification phase following dimension reduction, support vector machines (SVMs) are employed due to their effectiveness in handling complex datasets. SVMs play a vital role in the final stage of classification, enhancing the overall robustness and effectiveness of the framework.The primary objective of this study is to enhance the accuracy and efficiency of cancer type classification by integrating dimension reduction techniques with advanced classification algorithms. This integrated approach not only enhances the reliability of predictions but also fosters advancements in biomedical research and enables more precise clinical applications in cancer diagnosis and treatment.

3. Cancer Type Detection Based On Gene Expression Data Using Support Vector Machine

The methodology we propose for our system is well-organized and comprises four critical stages, all of which are essential for the successful achievement of our objectives. Each of these stages has been carefully designed to ensure the accuracy, effectiveness, and reliability of our system. In order to comprehend the significance and contribution of every step to the overall procedure, we shall analyze them separately: data cleansing, feature scaling wrapper-based feature reduction and classification.By integrating these four fundamental stages, the system we have proposed offers a comprehensive and systematic methodology for data analysis and modelling. This well-organized process makes sure that it is reliable, accurate, and scalable, so it can be used for a wide range of tasks in many different areas.

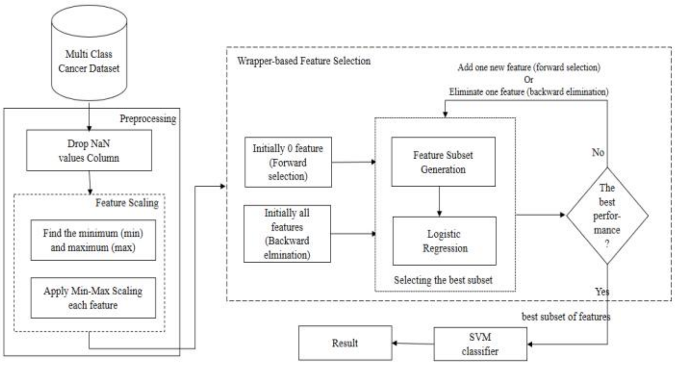

Multi Class Cancer Dataset
Preprocessing
Drop NaN values Column
Feature Scaling
Find the minimum (min) and maximum (max)
Apply Min-Max Scaling each feature
Wrapper-based Feature Selection
Add one new feature (forward selection)
Or
Eliminate one feature (backward elimination)
Initially 0 feature (Forward selection)
Initially all features (Backward elmination)
Feature Subset Generation
Logistic Regression
Selecting the best subset
No
The best perfor-mance ?
Yes
best subset of features
SVM classifier
Result

Figure 1: Proposed system design of cancer type detection using Support Vector Machine

3.1. Dataset

The ICMR dataset undertook a thorough effort to classify and study various types of concerning cancers that have arisen in recent years. These include prostate cancer (PRAD), lung cancer (LUAD), colon cancer (COAD), breast cancer (BRCA), and kidney cancer (KIRC). The main goal of this initiative was to collect and analyze extensive data to pinpoint the genetic factors linked to each specific cancer type. Through careful research and data gathering, the ICMR aimed to uncover the fundamental causes and mechanisms contributing to the onset of these cancers.The dataset compiled by the ICMR contains information from 802 individuals diagnosed with one of the mentioned cancer types. Each individual is represented by 802 associated samples, with each sample containing expression levels for 20,532 genes. These genes are crucial for understanding the unique molecular pathways and genetic signatures associated with each cancer type. Analyzing the gene expression patterns across these samples enables researchers to gain valuable insights into the molecular characteristics and potential biomarkers associated with PRAD, LUAD, COAD, BRCA, and KIRC.

This extensive dataset is a valuable asset for conducting detailed analyses, such as identifying gene mutations, detecting abnormal gene expression profiles, and exploring potential therapeutic targets specific to each cancer subtype. The comprehensive data provided by the ICMR's initiative not only enhances our comprehension of these complex diseases but also opens up avenues for advancements in precision medicine, personalized treatment approaches, and targeted therapies tailored to the genetic composition of individual patients.

3.2. Data Cleaning

The ICMR1 dataset includes cells with NaN values that need to be removed. Following that, Feature normalization, also known as min-max scaling, is a technique employed to transform numerical features into a specific range, typically bounded between 0 and 1. This process plays a pivotal role in standardizing data for analysis and modeling purposes. The methodology behind normalization involves adjusting each data point by subtracting the minimum value of the feature and then dividing it by the range, which is calculated as the difference between the maximum and minimum values of the feature. The formula for feature normalization is expressed as:

$$x_{normalized} = \frac{x - min(x)}{max(x) - min(x)} \qquad (1)$$

Here, x represents the original feature value, min(x) denotes the minimum value of the feature, and max(x) signifies the maximum value of the feature.Normalization proves to be particularly advantageous when dealing with features that exhibit varying ranges or magnitudes. It becomes essential in scenarios where algorithms necessitate all features to be on a similar scale for optimal performance. By applying normalization, we mitigate the risk of features with larger magnitudes overshadowing the learning process and skewing the model's outcomes.This technique finds widespread application across various domains, especially in machine learning algorithms such as neural networks, clustering algorithms like K-means clustering, and distance-based algorithms. The normalization of features ensures that the model's computations are not biased towards specific features due to their scale, thereby enhancing the overall accuracy and robustness of the learning process.

4. Wrapper-Based Feature Reduction

We employed a wrapper method on our dataset to improve our model's effectiveness and simplify its structure. Wrapper methods for feature selection are advanced techniques that assess subsets of features based on their impact on the performance of a machine learning model. Unlike filter methods that evaluate features independently of the chosen model, wrapper methods use the actual predictive model to evaluate feature subsets. This involves iteratively training the model, determining which features to include or exclude, and evaluating its performance.Various wrapper techniques are available, such as Forward Selection, Backward Elimination, Recursive Feature Elimination (RFE), and Bidirectional Elimination. In our project, we assessed forward selection and backward elimination methods to classify with Support Vector Machine (SVM).

4.1. Forward Feature Selection

Through the incremental addition of features, forward wrapper-based feature selection identifies the most relevant attributes for a logistic regression-based predictive model. Each feature is assessed using logistic regression during the process; the model is trained using the current feature set plus one additional feature; and the performance of the model is evaluated using a variety of metrics. We choose and integrate the feature that has the greatest impact on improving model performance into the feature set. Until a halting criterion, such as a predetermined number of features or a lack of significant improvement in model performance, is satisfied, this iterative process persists. The aim is to generate an ideal feature set that enhances the ability to determine outcomes, taking into account the interaction between features. By modifying stopping criteria, evaluation metrics, and parameters, one can customize the algorithm to fit particular problem domains and analysis objectives, thereby enhancing the accuracy and interpretability of the model.

4.2. Backward Feature Selection

By systematically removing features from the feature set, the machine learning technique known as "wrapper-based feature elimination" is utilized to select the most important characteristics for a predictive model. The process begins by training a predictive model using every feature extracted from the dataset. Any machine learning algorithm, such as logistic regression, support vector machines, decision trees, or random forests, can construct this model. We subsequently assess the model's performance by employing metrics such as accuracy, precision, recall, F1-score, AUC-ROC, and MSE.

Subsequently, the algorithm proceeds to eliminate each feature individually, beginning with the least significant feature determined by an initial criterion such as the coefficient value in logistic regression or feature importance. The model is retrained with the decreased feature set, and its performance is reevaluated using the same metrics following the removal of a feature. If the model's performance fixes or improves, we retain the feature elimination; if not, we restore the feature.

Until a stopping criterion is satisfied, such as completing a predetermined number of iterations or observing no sizable enhancement in model performance, this iterative procedure persists. In conclusion, the algorithm trains the final predictive model using the optimal subset of identified features.

The goal of feature elimination based on wrappers is to find the most important features that make a big difference in how well the model can predict the future, while also making the model easier to understand and less complicated to compute. That which determines the stopping point, the evaluation metric, and the feature removal criterion depends on the problem domain, the dataset's features, and the analysis's goals.

5. Classification by Support Vector Machine

Prior to the initiation of the classification process in machine learning, it is imperative to perform hyperparameter optimization. Hyperparameters, determined by the user while creating a machine learning model, precede the definition of parameters. Their objective is to identify the most advantageous parameters for the model. The primary objective of hyperparameter tuning is to identify the optimal settings for these hyperparameters in order to get the highest quality prediction outcomes from the model.An efficient method used for hyperparameter optimization is GridSearchCV (Grid Search Cross-Validation). This approach employs a systematic process to search for and determine the optimal combination of hyperparameters for a specified model. It does this by systematically examining a predetermined range of hyperparameter values, generating a "grid" of possible combinations. The system then evaluates each combination through cross-validation, selecting the one that exhibits the highest performance level. GridSearchCV greatly simplifies the process of tuning hyperparameters, resulting in enhanced model performance and reducing the need for human trial-and-error. Our plan uses both forward and backward feature selection methods to find the most important features. Then, we use the Support Vector Machine (SVM) classifier to put them into different groups. This integration improves the model's capacity to identify significant patterns in the data, offering a complete method for classifying different types of cancer.We generate two models to evaluate the performance of different feature selection methods combined with Support Vector Machine (SVM) classification. The first model employs forward selection, a stepwise process that starts with an empty set of features and adds one feature at a time. At each step, the feature that improves the model the most is added until no further improvement is observed. Once the optimal set of features is determined, we use these features to train an SVM classifier.The second model utilizes backward elimination, another stepwise approach that starts with the complete set of features. In each iteration, the least significant feature is removed. This process continues until further removal of features does not significantly affect the model's performance. After identifying the optimal subset of features, we train an SVM classifier using this refined feature set.By comparing the performance of these two models, we aim to understand the impact of different feature selection techniques on the classification accuracy of SVM.This combination of factors not only enhances the effectiveness of the classification process but also plays a crucial role in identifying important biomarkers and gene expressions that are linked to certain forms of cancer. Our goal is to improve the accuracy and reliability of our cancer classification model by using hyperparameter tuning and feature selection strategies in combination with the SVM classifier.

6. Experimental Results

In this section, we analyze the experimental results and conduct a thorough evaluation of our work. Utilizing a confusion matrix, we assess performance and determine the highest accuracy by adjusting the number of features (k values).

6.1. Analysis on Feature Reduction Result

The ICMR dataset, containing a high-dimensional feature set of 20,532 features, was used to evaluate our proposed framework. Feature reduction was achieved using a wrapper-based forward and backward selection methods with a Logistic Regression classifier. After feature extraction, classification was performed with an SVM for multiple outputs.

6.2. The wrapper-based forward selection method uses SVM for classification

The forward selection process was implemented in a Python Jupyter Notebook, providing an interactive and flexible environment for data analysis and visualization. During our investigation, we configured the forward selection algorithm to select the top 30 features from the dataset, setting the parameter k to 30. This means that out of the original 20,532 high-dimensional features, the algorithm identified the 30 most relevant features for our analysis. This selection process and the identified features are illustrated in Figure 2. Upon further analysis, we discovered that the model's performance remained unchanged when we used only the top 10 features instead of the full set of 30 selected features. This finding suggests that the additional 20 features did not contribute significantly to the model's accuracy or reliability. Consequently, we decided to streamline our model by utilizing only the top 10 features. This decision is also supported by the visual representation in Figure 2, which confirms that the performance was consistent whether we used the top 10 or all 30 features. This approach allowed us to simplify the model without sacrificing performance, making it more efficient and easier to interpret.

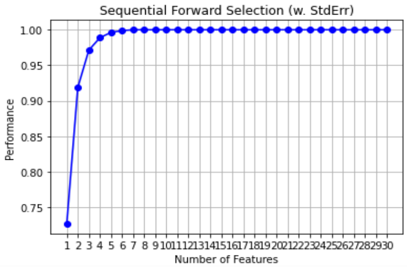
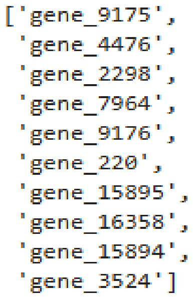


Figure 2: Performance of the model based on the number of features(k values)



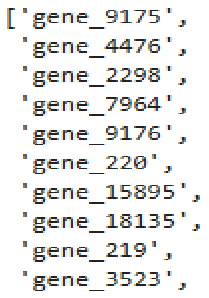
```
['gene_9175',
 'gene_4476',
 'gene_2298',
 'gene_7964',
 'gene_9176',
 'gene_220',
 'gene_15895',
 'gene_16358',
 'gene_15894',
 'gene_3524']
```

Figure 3: The best 10 features by forward selection method

Using the wrapper-based forward selection method, we systematically identified the top 10 most relevant features from our extensive dataset. This method involves evaluating subsets of features and selecting the combination that optimizes the performance of the classifier, in this case, an SVM (Support Vector Machine). By focusing on these 10 features, we were able to streamline the model while retaining the most informative variables for classification. Once the top 10 features were selected, we employed an SVM to classify the data into different cancer types. The classification process yielded highly impressive accuracy rates, demonstrating the robustness of our approach.

6.3. The wrapper-based backward elimination method uses SVM for classification

At present, we are using a wrapper-based backward elimination method to perform feature selection in our machine learning pipeline, with a Support Vector Machine (SVM) as our classifier. This method iteratively evaluates and removes features based on their impact on model performance. In a manner similar to the forward selection method, we start by identifying the top 30 features from a large set of 20,532 high-dimensional features. These 30 features are selected because they provide the best initial model performance.After narrowing down to these 30 features, we further analyze their contribution to the model’s accuracy. Interestingly, we observe that the model’s accuracy remains stable when we limit the feature set to the top 10 features out of the initial 30. This suggests that these top 10 features are highly informative and sufficient for maintaining optimal model performance. Moreover, it is noteworthy that 7 out of these top 10 features are identical to those identified through the forward selection method. This overlap indicates a consistency in the most significant features across different selection methods, reinforcing their importance. The top 10 selected features are illustrated in the following figure (4), providing a clear illustration of the most critical features for our SVM classifier.



```
['gene_9175',
 'gene_4476',
 'gene_2298',
 'gene_7964',
 'gene_9176',
 'gene_220',
 'gene_15895',
 'gene_18135',
 'gene_219',
 'gene_3523',
```

Figure 4: The top 10 features were selected by the wrapper method

By using the wrapper-based backward selection method, we systematically identified the top 10 most relevant attributes from our extensive dataset. This method involves evaluating subsets of features to find the combination that maximizes the performance of our classifier, specifically an SVM (Support Vector Machine) in this case. By focusing on these 10 features, we optimized the model by selecting the most relevant variables for classification while eliminating unnecessary ones. After selecting the top 10 features, we used the SVM algorithm to categorize the data into different cancer types. The classification process resulted in remarkably high accuracy rates, demonstrating the strength and reliability of our approach.

6.4. Evaluation of our proposed model

In our experiment, we used Support Vector Machines (SVM) to classify various cancer types in the ICMR test dataset. The training was carried out using a Python Jupyter notebook. To assess our proposed methodology, we calculated accuracy, precision, recall, and F1-score using confusion matrices. Confusion matrices are essential for classification tasks and consist of four elements: true positive (TP), false positive (FP), false negative (FN), and true negative (TN).

These matrices allow us to calculate evaluation metrics for each classifier. TP represents true positives, which are correctly classified positive observations. FN represents false negatives, which are positive observations incorrectly classified as negative. FP represents false positives, which are negative observations incorrectly classified as positive. TN represents true negatives, which are correctly classified negative observations.

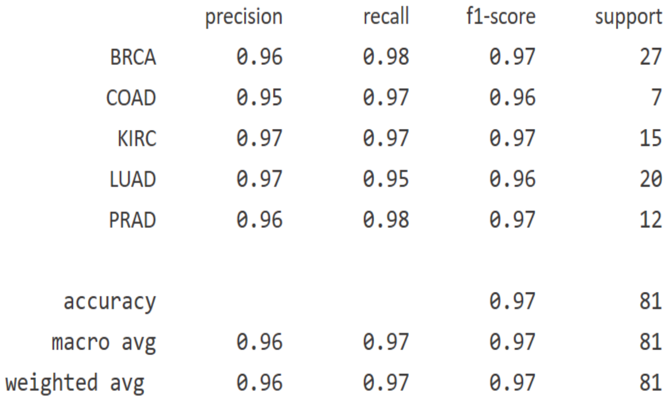
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

$$Precision = \frac{TP}{TP+FP} \qquad (4)$$

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

$$F1 - score = \frac{2*Precision*Recall}{Precision+Recall} \qquad (6)$$

The accuracy, precision, recall, and F1-score were computed using specific mathematical formulas (Equations 3, 4, 5, and 6, respectively) to evaluate the model's overall effectiveness. Accuracy measures the model's ability to correctly classify instances, precision quantifies the proportion of true positive predictions, recall assesses the proportion of actual positive instances correctly identified, and the F1-score is a combined metric that considers both precision and recall across the entire dataset.



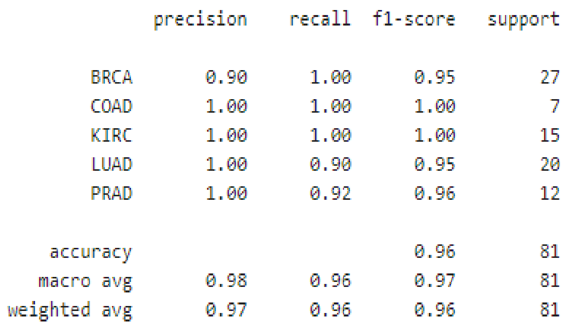
```
              precision    recall  f1-score   support
        BRCA       0.96      0.98      0.97        27
        COAD       0.95      0.97      0.96         7
        KIRC       0.97      0.97      0.97        15
        LUAD       0.97      0.95      0.96        20
        PRAD       0.96      0.98      0.97        12

    accuracy                           0.97        81
   macro avg       0.96      0.97      0.97        81
weighted avg       0.96      0.97      0.97        81
```

Figure 5: The average results of accuracy, precision, recall and f1-score generated by using forward selection for selecting features and classifying them using SVM

Specifically, the wrapper-based forward selection method classified by the SVM achieved a 98% accuracy rate for identifying BRCA, a 96% accuracy rate for COAD, a 97% accuracy rate for KIRC, a 96% accuracy rate for LUAD, and a 97% accuracy rate for PRAD, as illustrated in the following figure(5). These results underscore the effectiveness of our feature selection and classification strategy.

Based on Figure 6, this report offers a succinct summary of the multiclassification results for cancer types derived from the ICMR dataset. The evaluation indicates that BRCA, COAD, KIRC, LUAD, and PRAD demonstrate notable accuracy scores of 0.95, 1.0, 1.0, 0.98, and 0.96, correspondingly.



```
              precision    recall  f1-score   support

        BRCA       0.90      1.00      0.95        27
        COAD       1.00      1.00      1.00         7
        KIRC       1.00      1.00      1.00        15
        LUAD       1.00      0.90      0.95        20
        PRAD       1.00      0.92      0.96        12

    accuracy                           0.96        81
   macro avg       0.98      0.96      0.97        81
weighted avg       0.97      0.96      0.96        81
```

Figure 5: The average results of accuracy, precision, recall and f1-score generated by using backward selection for selecting features and classifying them using SVM

The proposed system utilized Support Vector Machines (SVM) to classify various cancer types within the ICMR test dataset. Training occurred through a Python Jupyter notebook, with evaluation involving the assessment of accuracy, precision, recall, and F1-score via confusion matrices. These matrices enable the computation of evaluation metrics such as true positives, false positives, false negatives, and true negatives. Mathematical formulas were applied to determine accuracy, precision, recall, and F1-score, providing insights into the model's classification performance. The wrapper-based forward selection approach, combined with SVM, exhibited remarkable accuracy in identifying diverse cancer types, underscoring the efficacy of the feature selection and classification methodology.

7. Conclusions And Future Works

In the realm of deciphering complex patterns within gene expression data, advanced machine learning and artificial intelligence techniques have become indispensable. Researchers are actively exploring sophisticated algorithms capable of analyzing vast genomic datasets to extract meaningful information for precise classification of various cancer types. However, the high feature dimensions inherent in gene datasets pose challenges in selecting significant features tailored to specific cancer types.

In our study, we proposed a model employing wrapper-based forward and backward elimination techniques to reduce the feature set from 20,532 to a mere 10. This reduction streamlines the process, making it more efficient and effective in identifying unknown gene data values and detecting cancer types. Notably, our proposed model achieves impressive average accuracies of 97% with forward selection and 96% with backward elimination.

While these initial results are promising, further testing with larger datasets is necessary to validate the robustness and scalability of our approach.

References

[1] Anna, A.; Monika, G. Splicing Mutations in Human Genetic Disorders: Examples, Detection, and Confirmation. J. Appl. Genet. 2018, 59, 253–268

[2] Lunshof, J.E.; Bobe, J.; Aach, J.; Angrist, M.; Thakuria, J.V.; Vorhaus, D.B.; Hoehe, M.R.; Church, G.M. Personal Genomes in Progress: From the Human Genome Project to the Personal Genome Project. Dialogues Clin. Neurosci. 2010, 12, 47–60.

[3] Khan, M.F.; Ghazal, T.M.; Said, R.A.; Fatima, A.; Abbas, S.; Khan, M.A.; Issa, G.F.; Ahmad, M.; Khan, M.A. An IoMT-Enabled Smart Healthcare Model to Monitor Elderly People Using Machine Learning Technique. Comput. Intell. Neurosci. 2021, 2021, 2487759.

[4] Bhonde, S.B.; Prasad, J.R. Deep Learning Techniques in Cancer Prediction Using Genomic Profiles. In Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2–4 April 2021; pp. 1–9

[5] Celesti, F.; Celesti, A.; Wan, J.; Villari, M. Why Deep Learning Is Changing the Way to Approach NGS Data Processing: A Review. IEEE Rev. Biomed. Eng. 2018, 11, 68–76.

[6] Vaiyapuri, T.; Liyakathunisa; Alaskar, H.; Aljohani, E.; Shridevi, S.; Hussain, A. Red Fox Optimizer with Data-Science-Enabled Microarray Gene Expression Classification Model. Appl. Sci. 2022, 12, 4172. https://doi.org/10.3390/ app1209417P. D. Turney, "*Similarity of Semantic Relations*," 2006.

[7] hukla, A.K.; Singh, P.; Vardhan, M. A two-stage gene selection method for biomarker discovery from microarray data for cancer classification. Chemom. Intell. Lab. Syst. 2018, 183, 47–58

[8] Mohammed, M., Mwambi, H., Mboya, I.B. *et al.* A stacking ensemble deep learning approach to cancer type classification based on TCGA data. *Sci Rep* **11**, 15626 (2021). https://doi.org/10.1038/s41598-021-95128-x.

[9] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol. 2005 Apr;3(2):185-205. doi: 10.1142/s0219720005001004. PMID: 15852500.

[10] Z. Zhao, R. Anand and M. Wang, "Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform," 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Washington, DC, USA, 2019, pp. 442-452, doi: 10.1109/DSAA.2019.00059.