DEVELOPING A MULTIDIMENSIONAL FUZZY DEEP LEARNING FOR CANCER CLASSIFICATION USING GENE EXPRESSION DATA

Mahmood Khalsan[1,3] , Mu Mu[1] (Member, IEEE), Eman Salih Al-shamery,
Suraj Ajit[1] , Lee Machado[2], and Michael Opoku Agyeman[1], (Senior
Member, IEEE)

[1]Advanced Technology Research Group, Faculty of Arts, Science and Technology,
The University of Northampton, UK.
[2]Centre for Physical Activity and Life Science, Faculty of Arts, Science, and
Technology, The University of Northampton, UK
[3]Computer Science Department, University of Babylon, College of Information

ABSTRACT

In the realm of cancer research, the identification of biomarker genes plays a pivotal role in accurate classification and diagnosis. This study delves into the intersection of machine learning and gene selection to enhance the precision of biomarker identification for cancer classification. Leveraging advanced computational techniques. In the quest for improved cancer classification, studies face challenges due to high-dimensional gene expression data and limited gene relevance. To address these challenges, we developed a novel multidimensional fuzzy deep learning (MFDL) to select subset of significant genes and using those genes to train the model for better accuracy. MFDL is exploring the integration of fuzzy concepts within filter and wrapper methods to select significant genes and applying a fuzzy classifier to improve cancer classification accuracy. Through rigorous experimentation and validation, six gene expression data used, the findings demonstrated the efficacy of our methodology on diverse cancer datasets. The results underscore the importance of integrative computational methods in deciphering the intricate genomic landscape of cancer and spotlight the potential for improved diagnostic accuracy. The developed model showcased outstanding performance across the six employed datasets, demonstrating an average accuracy of 98%, precision of 98.3%, recall of 97.6%, and an f1-score of 97.8%.

KEYWORDS

Deep learning, Gene selection, Cancer classification , Gene expression

1. INTRODUCTION

Cancer, a complex and heterogeneous group of diseases, poses a significant global health challenge [1][15]. With advancements in molecular biology and genomics, the use of gene expression data for cancer classification has gained substantial attention. However, the dimensionality and noise present in high-throughput genomic data can hinder the effectiveness of classification models. To address these challenges, the strategic selection of relevant genes has emerged as a critical preprocessing step in enhancing the accuracy, interpretability, and

David C. Wyld et al. (Eds): SEAPP, Fuzzy, SIGEM, CSEA, MLTEC -2023
pp. 37-49, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.132304

generalizability of cancer classification models. The advent of microarray and next-generation sequencing technologies (RNA-seq) has enabled the simultaneous measurement of expression levels of thousands of genes in a single experiment [2]. While this wealth of data holds immense potential, it also presents computational and statistical challenges. The curse of dimensionality, where the number of features (genes) far exceeds the number of samples, can lead to overfitting and suboptimal model performance [3]. Moreover, the presence of noise and irrelevant genes can further degrade the predictive power of classification models [4][5].

Gene selection techniques offer a solution by identifying a subset of genes that are most informative for distinguishing between different cancer types or subtypes. These techniques encompass a spectrum of methodologies, ranging from filter methods that rank genes based on statistical measures of their relevance, to wrapper methods that evaluate subsets of genes using machine learning algorithms incorporate gene selection into the model training process. The primary goal of this paper is to explore the efficacy the developed model for enhancing cancer classification. By systematically reducing the dimensionality of the gene expression data, the algorithm aims to improve the performance of classification models while maintaining or even enhancing their interpretability.

The paper presents valuable contributions, as outlined below:

- Development of a state-of-the-art Multidimensional Fuzzy Deep Learning Model capable of processing cancer gene expression datasets. This model incorporates a gene selection method to identify an optimal subset of genes, followed by the application of a classifier for cancer classification.

- Selection of a limited number of significant genes to reduce dataset dimensionality. This reduction leads to shorter training times, mitigates overfitting, simplifies the classifier, and enhances classification accuracy.

- Introduction of a novel deep learning architecture designed to minimize information loss and reduce processing time.

The rest of the article structured as follows: Section 2 reviews previous studies on cancer classification using gene expression data and machine learning integration. Section 3 details the development steps of our proposed model, while Section 4 covers the experimental setup, including programming language, hardware, evaluation metrics, and data splitting. Section 5 presents experiment outcomes, including datasets, results, and discussions. In Section 6, we compare our model to prior work using the same datasets for fairness. Finally, the concluding section summarizes the paper's main findings.

2. RELATED WORK

Using microarray cancer data extracted from the Gene Expression Omnibus (GEO) dataset with accession number GSE43580, Support Vector Machine (SVM) and Random Forest (RF) methodologies were employed for the categorization of distinct lung cancer subtypes, namely adenocarcinomas (AC) and squamous cell carcinomas (SCC) [6]. This analysis also incorporated Monte-Carlo simulation along with incremental feature selection (MCSF) to ascertain the pivotal genes influencing the classification of cancer types. Upon integrating the results of MCSF and SVM, a set of 43 genes, the performance metrics stood at 86%, 80%, 98%, and 88% for accuracy, precision, recall, and F1-score, respectively. When the identical 43 genes were harnessed through the lens of MCSF and RF, the corresponding outcomes improved to 88%, 82%, 97%, and 89% for accuracy, precision, recall, and F1-score.

RelifF, a feature selection technique, was harnessed to discern the top 19 genes from an acquired lung cancer dataset sourced from (DI GSE10072). Subsequently, this selected gene set was employed as the training data for the Naive Bayes (NB) classifier [7]. The dataset encompasses 58 samples of Adenocarcinoma and 49 samples of normal lung tissue. Implementing the NB classifier yielded an impressive accuracy of 95%. The outcomes highlight the positive influence of RelifF in enhancing accuracy while concurrently curtailing the gene count. Nevertheless, it's crucial to note that the evaluation of the proposed model is based on a diminutive dataset, potentially limiting the achieved accuracy when extrapolated to more expansive datasets. Moreover, the study exclusively concentrates on microarray data, neglecting alternative data modalities like RNA-seq. The research's applicability in a broader context, particularly in multi-class datasets where only binary classification was explored, remains unexplored. Additionally, the absence of comprehensive assessment metrics such as precision, recall, and F1-score are a notable limitation in this study.

Deng et al. [8], the k-Nearest Neighbors (KNN) technique was employed in conjunction with the principles of maximum relevance and minimum redundancy (mRMR). Their focus was on classifying thyroid carcinoma, where KNN served as the classification method. To pinpoint informative genes suitable for training KNN, the study employed mRMR as a gene selection approach. The microarray gene expression dataset (GEO GSE33630) used in this research encompassed a total of 105 samples. The results showcased an accuracy of 85.7% when utilizing the top 10 genes chosen through the mRMR methodology. This underscores the study's ability to streamline the gene pool while acknowledging that the achieved accuracy did not reach the level demanded for accurate cancer sensitivity assessment. However, it's noteworthy that the proposed model lacked validation with additional gene expression datasets, a step that would ensure its efficacy in accurately categorizing various types of cancer.

Hila et al. [19] introduced a novel feature subset selection method using an adaptive neuro-fuzzy inference system for gene expression classification. The study assessed this approach on four microarray gene expression datasets related to different types of cancer (Leukemia, Prostate, DLBC Stanford, and Colon Cancer). Comparative evaluations against existing classifiers demonstrated classification accuracies of 89.47%, 83.33%, 80.65%, and 73.33% for Colon Cancer, Leukemia, Prostate Cancer, and DLBC Stanford datasets, respectively. However, these results fell short of achieving high performance compared to previous studies in the field. The study faced limitations due to small dataset sizes (all contained less than 100 samples, except for Prostate with 102 samples) and exclusive use of binary class datasets. It's worth noting that the proposed model's efficiency may vary when applied to multi-class datasets.

For a more comprehensive understanding of the topic, it's worth noting that our earlier research [9] extensively explores a multitude of studies in this field. Drawing from these investigations and others discussed in our prior survey [9], it becomes evident that there is a significant demand for the development of an end-to-end process capable of efficiently selecting a minimal number of optimal genes and achieving accurate cancer classification.

3. THE PROPOSED MODEL

The proposed model involves two main stages, firstly developing gene selection method that aims to select a subset of informative genes that highly correlated with the target (class). Therefore, developing classifier method that be able to accurately classify cancer expression datasets. These two stages are combined in a single model, namely multidimensional fuzzy deep learning (MFDL). The topology architectural of the developed model is visually presented in Figure 1. This model is constructed with a total of 22 hidden layers , complementing the one input layer and one output layer.

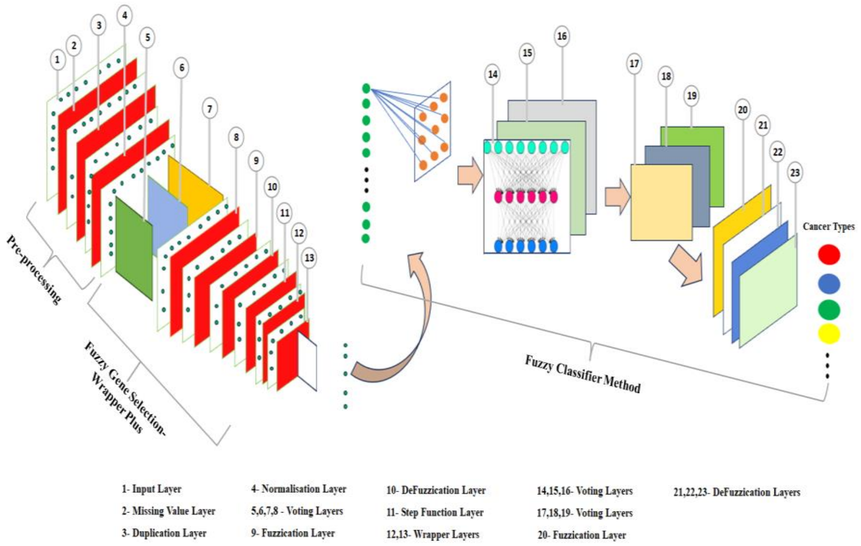

Figure 1: The developed Topology of the Proposed Model.

3.1. Fuzzy Gene Selection-Wrapper Plus (FGSWP)

FGSWP was developed by integrating filter and wrapper methods and using fuzzification and defuzzification methods. FGSWP aims to select a subset of genes that positively impact cancer classification by reducing the classifier complexity and improving the accuracy. Furthermore, it reduces the overfitting and time consuming for training stage. FGSWP using three feature selection method (Mutual information, F-classif, and Chi-squared and getting the score and rank for each gene for these three gene selection methods. Then, selecting three different lists based on Step Function (SF) by using the formula . This process called voting step.

$$SF = max(FSS) * 0.3 \qquad (1)$$

SF is the step function and GS is the gene score for each gene for each gene selection method.

The outcome is three lists of top genes are selected as first step for further investigation to select informative genes. Next, to select best single score for each in the three list, fuzzification and defuzzification were used. Fuzzification aims to make the gene score between [0,1] in the three selected lists of genes by using three member functions. While the defuzzification attempts to find the best score for each gene in the three lists of genes. Mathematically the member function method calculated as follows.

$$MF = \frac{\mathrm{Wi-a}}{b-a} \qquad (2)$$

The outcome of the previous step is the score for each gene in the three lists is between (0,1). To obtain single best score for each gene in three list fuzzification calculated as follows.

$$\mathrm{ASG} = \frac{\mathrm{MFi + MFi + MFi}}{N} \qquad (3)$$

Where ASG is the average score for gene in the three lists, while MF is the member function for each gene selection method.

Through the two processes, it becomes evident that the utilization of fuzzification and defuzzification has been instrumental in attaining the objective of obtaining an optimal singular score for each gene. In contrast, filter feature selection methodologies yield disparate scores for the identical gene. Consequently, the adoption of a SF takes precedence in determining the genes of significance to serve as valuable markers for accurate cancer classification, as elucidated in the equation provided below.

$$\mathbf{SF} = \mathbf{max(FSS)} * \mathbf{0.5} \tag{4}$$

In this context, the step Function (SF) exhibits an adaptable characteristic, wherein it captures the highest score among the obtained scores and subsequently multiplies it by 0.5 an outcome stemming from the prior procedural steps. This computation offers a trio of pivotal advantages. Firstly, it serves to avert the scenario of selecting null genes in all instances, an occurrence that could transpire when employing a constant SF. As an illustration, if the SF is set at 0.5 and all gene scores hover around 0.49, the application of a constant SF could result in the exclusion of all genes. Secondly, this approach sidesteps the inadvertent omission of genes that bear identical scores, a scenario that could arise when comparing the selection of the top 10 genes. For instance, if the SF applied to the top 10 genes yields identical scores for the tenth and the eleventh gene, the latter might be overlooked unless the SF strategy is introduced. The proposed methodology thus ensures robust gene selection outcomes by adapting to the dynamic range of scores, avoiding the risk of null selections or gene dismissals due to tie scores, and fostering a more comprehensive gene marker selection process. The outcome of this methods will be used to further investigation by employing backword elimination method for further reducing the number of selected genes without sacrificing the accuracy and other evaluation metrics.

3.2. Fuzzy Classifier

The primary objective of developing the Fuzzy Classifier (FC) is to enhance cancer classification accuracy and improve algorithm generalization across diverse datasets. FC employs three classifier techniques (LR, SVM, and MLP) for each dataset, generating class label prediction probabilities. It selects the class label with the highest average max probability as the predicted label through a "soft" process. Additionally, FC incorporates a "majority" process, where the most frequently predicted label among classifiers is chosen. The FC method combines both processes. For example, if the soft process predicts class A and the majority process predicts class B, FC calculates a member function that considers both methods. This function adds 0.6 to the majority-selected class's average max and divides it by two. If the calculated output exceeds the original max average, it becomes the predicted class; otherwise, the soft-predicted class remains as the prediction.

4. EXPERIMENTAL SETUP

The MFDL model was built using Python software, leveraging the power of an Intel Core i7-8565U processor and 32 GB of RAM. To assess the model's effectiveness comprehensively, we conducted evaluations across various cancer types. In this rigorous analysis, we employed thirteen distinct datasets, comprising nine microarray and four RNA-seq datasets. This diverse set of data enabled us to thoroughly investigate the performance and versatility of the MFDL model. Employing a cross-validation approach, we meticulously split the datasets into training and testing subsets, ensuring a robust assessment of model generalization and the attainment of reliable results.

4.1. Cross-Validation (Cv)

Cross-validation (CV) is a crucial statistical technique employed in the field of machine learning (ML) to tackle the persistent issue of overfitting across diverse classifier paradigms [17]. This technique, through the utilization of k-fold cross-validation, enhances the model's capacity to learn and generalize effectively [18]. Instead of relying solely on a single training dataset, k-fold cross-validation partitions the data into k subsets or folds, training the model on each fold in turn. This practice fosters robustness in the model's ability to make accurate predictions on unseen data. The benefits of employing k-fold cross-validation are twofold. Firstly, it encourages the model to generalize more effectively, thereby reducing the likelihood of overfitting to the training data. Secondly, it provides a more comprehensive and reliable assessment of the algorithm's predictive performance. As depicted in Figure 4, the dataset is divided into k equal-sized folds, often with a common choice being k=5. Each fold is subsequently used for training the model while the remainder are used for validation. This process is repeated k times, ensuring that every data point has an opportunity to be part of the validation set. The model's performance is then evaluated using metrics such as accuracy, precision, recall, or F1-score across these k iterations. This rigorous evaluation procedure offers valuable insights into the model's robustness and its predictive capabilities, making it an indispensable tool for ML practitioners.

4.2. Evaluation Mitrics

To evaluate the effectiveness of our proposed model, we employed four key evaluation metrics. While achieving a high level of accuracy is undoubtedly important, it does not always provide a comprehensive measure of a model's quality. For instance, consider a dataset where 90% of the samples are normal and only 10% are cancerous. In such a scenario, even if the model achieves an accuracy of 90%, it may fail to correctly predict any of the cancer samples. Conversely, the model could classify all samples as normal and still achieve a 90% accuracy rate. This highlights the necessity of utilizing additional evaluation metrics to gauge the model's performance comprehensively.

4.2.1. Accuracy (Ac)

Ac is a metric that measures the proportion of correctly predicted instances (or samples) out of the total instances in a dataset. It is a common evaluation metric used for classification problems, where the goal is to assign a label or class to each input data point. Mathematically calculated as follows [20].

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN) \qquad (5)$$

In this context, TP represents "true positives," TN stands for "true negatives," FP signifies "false positives," and FN denotes "false negatives.

A TP signifies a prediction made by the model that is both accurate and correctly identified as a positive class. TN corresponds to a correct prediction where the model accurately recognizes a case as a negative class. For instance, non-cancerous instances are appropriately classified as such by the model, representing TN. FP refers to an erroneous prediction of the positive class by the model. On the other hand, FN indicates that the model incorrectly identifies a case as belonging to the negative class when it actually belongs to the positive class.

4.2.2. Precision (Pre)

Pre is a metric that quantifies the accuracy of positive predictions made by a model. It measures the proportion of correctly predicted positive instances (True Positives, TP) out of all instances that the model predicted as positive (True Positives + False Positives, TP + FP). It calculated as follows [20].

$$\text{Precision} = TP/\,(TP+FP) \qquad (6)$$

4.2.3. Recall (Rec)

Rec is a metric that measures the ability of a model to correctly identify all relevant instances of a particular class, also known as the "True Positive Rate" or "Sensitivity." It quantifies the proportion of correctly predicted positive instances (True Positives, TP) out of all instances that truly belong to the positive class (True Positives + False Negatives, TP + FN). Mathematically calculated as follows [20].

$$\text{Recall} = TP/(TP+FN) \qquad (7)$$

4.2.4. F1-Score

F1-score is a metric that combines both precision and recall providing a balanced measure of a model's performance. It is particularly useful when dealing with imbalanced datasets or when there is a need to strike a balance between minimizing false positives and false negatives. The formula for calculating the F1-score is:

$$\text{F1-score} = 2^{*}\,(\text{Precision} \,{*}\text{Recall})/(\text{Precision} + \text{Recall}) \qquad (8)$$

5. EXPERIMENTS

5.1. Employed Datasets

Six cancer expression datasets used for training and testing the proposed model. Table 1 presents full details of used datasets, including the Dataset ID, the measurement tools , the number of samples and number of genes for each dataset as well as the number of classes. These datasets were obtained from the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas Program (TCGA). The utilized datasets included both binary and multi-class data to ensure the adaptability of the proposed model to both scenarios.

Table 1: Full details of employed datasets

Datasets	Measurement	N-Samples	N-Gene	N-Class	Reference
GSE53757	Microarray	144	23516	2	[16]
GSE33630	Microarray	105	23518	3	[8]
GSE45827	Microarray	155	29873	6	[9]
TCGA	RNA-seq	2086	971	5	[10]
GSE10072	Microarray	107	13298	2	[9]
GSE43580	Microarray	150	54675	2	[5]

5.2. Experiment Results

Table 2 describes the findings of the proposed model using 5 k-fold cross validation for six gene expression datasets. Cross validation used to reduce the overfitting issue and increase the generalization of classifier algorithm. The results highlight the significant success of the proposed model in both gene reduction and accuracy. Across the six datasets utilized, the model consistently achieved outstanding performance, with average results of 98% accuracy, 98.3% precision, 97.6% recall, and 97.8% F1-score. Moreover, the number of selected genes in the six datasets were ranging between (69 and 5). The minimal count of selected genes not only optimizes training time but also effectively addresses overfitting concerns, resulting in an enhanced classifier performance. Collectively, the accomplishments of the proposed model align seamlessly with its intended objectives, attesting to its successful development. The findings unequivocally establish the superiority of the proposed model over existing classifier approaches, including Support Vector Machine (SVM) and Multilayer Perceptron (MLP). This outperformance was achieved even when trained these classifier approaches on the same number of selected genes using our FGSWP method. On the other hand, the noteworthy results obtained by SVM and MLP can be attributed to the selection of crucial genes by a component of our method. This underscores the effectiveness of the gene selection methodology we introduced as part of our model.

Table 2: The performance of employing MFDL method against SVM and MLP.

Dataset	N-genes	Classifier	Accuracy %	Precision %	Recall %	F1-score %
GSE53757	69	SVM	97	96	98.5	97
		MLP	97	97	97	97
		Our Model	100	100	100	100
GSE33630	17	SVM	93.3	90	90.5	89.5
		MLP	94	96	92.8	93
		Our Model	100	100	100	100
GSE45827	30	SVM	98.7	99	98.8	99
		MLP	99.3	99.4	99.4	99.4
		Our Model	100	100	100	100
TCGA	18	SVM	93.4	90	89.4	89.5
		MLP	94.6	91	91	90.7
		Our Model	96	95	94	94
GSE10072	5	SVM	97	98	96	97
		MLP	96	94.8	98	96
		Our Model	100	100	100	100
GSE43580	8	SVM	85.3	98	71	82
		MLP	86.6	90.4	83.5	86
		Our Model	93	95	92	93

To clarify the results obtained in this experiment, bar charts have been employed to illustrate the comparison between the proposed model and traditional classifier approaches (SVM, MLP) across four evaluation metrics. Figure 2 presents the average accuracy scores across 5-fold cross-validation for each dataset, showcasing the performance of MFDL in comparison to SVM and MLP. In Figure 3, we observe the precision scores averaged over 5-fold cross-validation for each dataset, highlighting the differences between the proposed model and the two other classifier algorithms. Figures 4 and 5 provide visualizations of the average recall and F1-score, respectively, also computed over 5-fold cross-validation. In summary, these figures collectively demonstrate the superior performance of the proposed model when compared to traditional classifier methods across all evaluated metrics. Across the six datasets, SVM achieved average scores of 93.8% for accuracy, 95% for precision, 90.3% for recall, and 92% for F1-score. Meanwhile, MLP achieved average scores of 94.3% for accuracy, 94.5% for precision, 93.3% for recall, and 93.5% for F1-score. In contrast, the proposed model consistently outperformed both SVM and MLP, achieving average scores of 98% for accuracy, 98.3% for precision, 97.6% for recall, and 97.8% for F1-score.

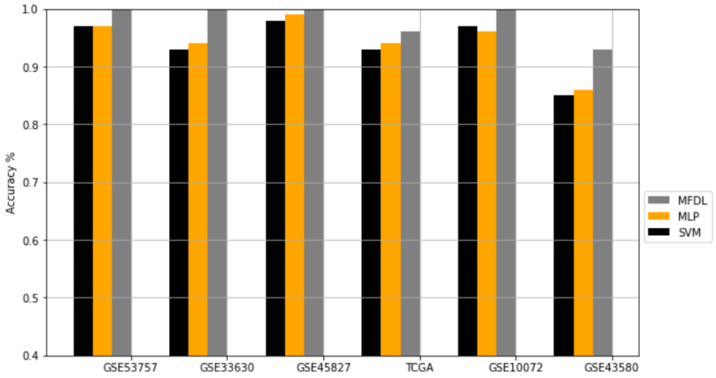

Figure 2: Comparing accuracy of the proposed mode against SVM and MLP for each dataset.

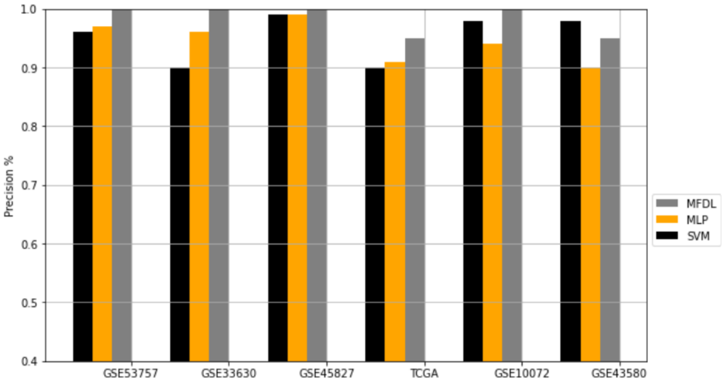

Figure 3: Comparing precision score of the proposed model against SVM and MLP for each dataset.

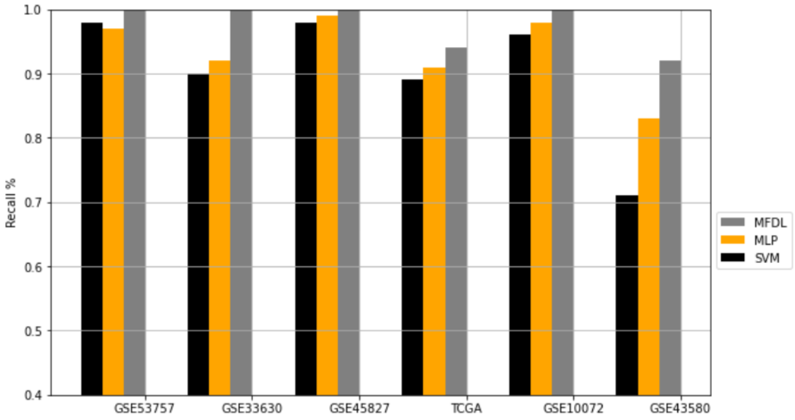

Figure 4: Comparing the recall scores of the proposed model against SVM and MLP for each dataset.

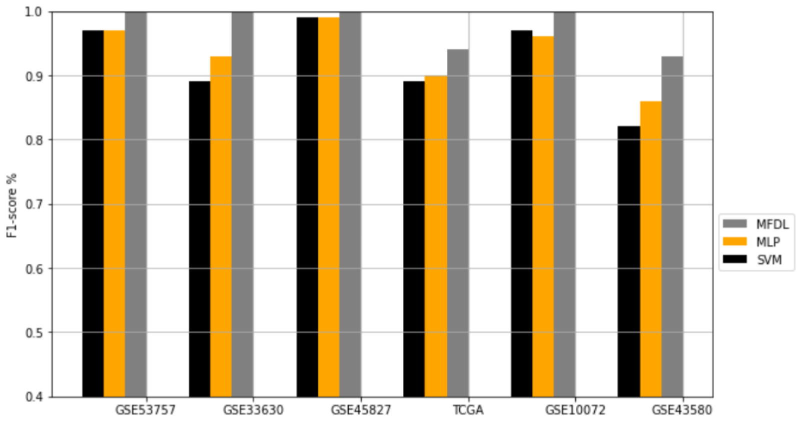

Figure 5: Comparing the f1-score of the proposed model against SVM and MLP for each dataset.

6. Comparing MFDL to Published Work

The introduced model showcased a notable advancement over preceding studies employing identical datasets. Our proposed algorithm exhibited superiority over existing research in terms of both the number of selected genes and the accuracy of cancer classification across all datasets employed in our experiment. This compelling performance gap is evident from the results tabulated in Table 3.

Table 3: Comparing MFDL to prior studies.

Datasets	N-gene	Approach	Ac%	Pre%	Rec%	F1%	Reference
GSE43580	43	MCSF +RF	88	82	97	89	[11]
	8	MFDL	93	95	92	93	Our Model
GSE33630	No	PCA + RF	92	92	89	No	[12]
	17	MFDL	100	100	100	100	Our Model
GSE10072	19	ReliefF+NB	95	No	No	No	[7]
	5	MFDL	100	100	100	100	Our Model
GSE45827	38	Rough set +SVM	96.86	96.9	97.34	97.8	[13]
	30	MFDL	100	100	100	100	Our Model
TCGA	971	BPSO-DT+CNN	96	94.96	95	95	[14]
	18	MFDL	96	95	94	94	Our Model

7. Conclusion

The article introduces developing new end-to-end process for the selection of informative genes and the precise classification of various cancer types using gene expression data. This approach was rigorously evaluated using six distinct gene expression datasets through 5-fold cross-validation. The experimental results reveal the exceptional performance of the proposed model, achieving outstanding outcomes in terms of gene subset size reduction and cancer classification accuracy across all datasets examined. To further validate the efficacy of the proposed algorithm, it underwent a comprehensive comparative analysis against previously published methods that utilized the same datasets. These comparisons underscore the remarkable efficacy and performance of the proposed model. Despite these significant achievements, it is crucial to recognize the inherent limitations of the study. The reliance on a limited quantity of datasets emphasizes the necessity for future research to encompass a broader array of microarray datasets.

This expansion would provide valuable insights into the algorithm's effectiveness within a more extensive context. Additionally, the exclusive focus on microarray datasets in this study prompts future investigations to shift their attention toward RNA-seq datasets. Such investigations would offer a more comprehensive understanding of the algorithm's applicability and its untapped potential in the realm of RNA-seq data analysis.

REFERENCES

[1] S. Shandilya and C. Chandankhede, "Survey on recent cancer classification systems for cancer diagnosis, "International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai,2017, pp. 2590-2594, IEEE.

[2] A. Wolff, M. Bayerlová, J. Gaedcke, D. Kube, and T. Beiÿbarth, ``A comparative study of RNA-seq and microarray data analysis on the two examples of rectal-cancer patients and burkitt lymphoma cells," PLoS ONE, vol. 13, no. 5, May 2018, Art. no. e0197162.

[3] Barbour, D. (2019). Precision medicine and the cursed dimensions. npj Digital Medicine volume, 4(2).

[4] Vanjimalar, S., Ramyachitra, D., and Manikandan, P. (2018). A review on feature selection techniques for gene expression data. In 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pages 1–4.

[5] Alhenawi, E., Al-Sayyed, R., Hudaib, A., and Mirjalili, S. (2022). Feature selection methods on gene expression microarray data for cancer classification: A systematic review. Computers in Biology and Medicine, 140:105051.

[6] F. Yuan, L. Lu, and Q. Zou, ``Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," Biochimica et Biophysica Acta (BBA) Mol. Basis Disease, vol. 1866, no. 8, Aug. 2020,Art. no. 165822

[7] A. L. Pineda, H. A. Ogoe, J. B. Balasubramanian, C. R. Escareño, S. Visweswaran, J. G. Herman, and V. Gopalakrishnan, ``On predicting lung cancer subtypes using `omic' data from tumor and tumor adjacent histologically-normal tissue," BMC Cancer, vol. 16, no. 1, p. 184, Dec. 2016.

[8] Xu Y, Deng Y, Ji Z, Liu H, Liu Y, Peng H, et al. (2014) Identification of Thyroid Carcinoma Related Genes with mRMR and Shortest Path Approaches. PLoS ONE 9(4).

[9] M. Khalsan et al., "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction," in IEEE Access, vol. 10, pp. 27522-27534, 2022, doi: 10.1109/ACCESS.2022.3146312.

[10] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M. and Cancer Genome Atlas Research Network, 2013. The cancer genome atlas pancancer analysis project. Nature genetics, 45(10), p.1113.

[11] Yuan, F., Lu, L., and Zou, Q. (2020). Analysis of gene expression profiles of lung cancer subtype with machine learning algorithms. Biochimica et Biophysica Acta (BBA)- Molecular Basis of Disease, 1866(8):165822.

[12] Cava, C., Salvatore, C., and Castiglioni, I. (2023). Pan-cancer classification of gene expression data based on artificial neural network model. Applied Sciences, 13(13):2076–3417.

[13] Sujata Patil1, Kavitha Rani Balmuri2, J. F. . P.-s. B. S. K. and Nedoma4, J. (2022). Identification of triple negative breast cancer genes using rough set based feature selection algorithm ensemble classifier. Human-centric Computing and Information Sciences, 12(54).

[14] Khalifa, N. E. M., Taha, M. H. N., Ezzat Ali, D., Slowik, A., and Hassanien, A. E. (2020). Artificial intelligence technique for gene expression by tumor rna-seq data: A novel optimized deep learning approach. IEEE Access, 8:22874–22883.

[15] Celik, A. E., Rasheed, J., and Yahyaoui, A. (2022). Machine learning approaches for lung cancer prediction. In 2022 12th International Conference on Advanced Computer Information Technologies (ACIT), pages 540–543.

[16] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A.NCBI GEO: archive for functional genomics data sets update.Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5.

[17] Alippi, C. and Roveri, M. (2010). Virtual k-fold cross validation: An effective method for accuracy assessment. In The 2010 International Joint Conference on Neural Networks (IJCNN), pages 1–6.

[18] Nie, Y., De Santis, L., Carratù, M., O'Nils, M., Sommella, P., and Lundgren, J. (2020). Deep melanoma classification with k-fold cross-validation for process optimization. In 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pages 1–6.

[19] Hilal, A. M., Malibari, A. A., Obayya, M., Alzahrani, J. S., Alamgeer, M., Mohamed, A., Motwakel, A., Yaseen, I., Hamza, M. A., Zamani, A. S., et al. (2022). Feature subset selection with optimal adaptive neuro-fuzzy systems for bioinformatics gene expression classification. Computational Intelligence and Neuroscience, 2022.

[20] M. Khalsan, M. Mu, E. S. Al-Shamery, L. Machado, M. O. Agyeman and S. Ajit, "Intersection Three Feature Selection and Machine Learning Approaches for Cancer Classification," 2023 International Conference on System Science and Engineering (ICSSE), Ho Chi Minh, Vietnam, 2023, pp. 427-433, doi: 10.1109/ICSSE58758.2023.10227163.

AUTHORS

MAHMOOD KHALSAN is a PhD student in computer science at Northampton University, Mahmood is passionately exploring the intersection of cutting-edge machine learning techniques and the intricate world of gene expression to revolutionize cancer prediction. His research journey is dedicated to pushing the boundaries of what's possible in early-stage cancer detection. In pursuit of his doctoral thesis, Mahmood endeavours to forge vital connections between the realms of machine learning and biology. His overarching goal is to uncover the hidden gems among the genetic landscape, pinpointing the most promising candidate genes associated with cancerous samples. By doing so, he aspires to usher in a new era of precision and efficacy in cancer diagnosis. Mahmood brings a wealth of academic achievement to his current pursuit, holding both a Bachelor of Science (BSc) and a Master of Science (MSc) from the prestigious University of Northampton. His relentless commitment to unravelling the complexities of cancer prediction showcases his dedication to advancing science and making a meaningful impact on healthcare.

Mu Mu (Member, IEEE) is currently an Associate Professor at the University of Northampton, U.K. He has authored more than 50 peer-reviewed papers published in international conferences and journals. His research interests include human factors in multimedia distribution, intelligent networks, and immersive media. He has been the chair and a program committee member of renowned international conferences. He has been a principal investigator and a lead researcher of various research programs funded by European Commission, U.K. Research Councils, and other funding bodies. He is an Associate Editor of the Multimedia Systems (Springer) journal.

Eman Salih Al-Shamery received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Babylon, Iraq, in 1998, 2001, and 2013, respectively. After completing her M.Sc., she worked as an Assistant Lecturer at the Department of Computer Science, University of Babylon. She is currently a Professor with the Software Department, University of Babylon. Her current research interests include artificial intelligence, bioinformatics, machine learning, neural networks, deep learning, and data mining.

Suraj Ajit received the B.E. degree (Hons.) in computer science from Bangalore University, India, and the Ph.D. degree in computer science from the University of Aberdeen. He has five years of industry experience that includes working in BAE Systems as a Research Scientist for three years. He worked as a Research Assistant for four years in a prestigious advanced knowledge technologies project at the University of Aberdeen. He joined the University of Northampton, as a Software Engineering Lecturer, in 2011. He is currently an Associate Professor and the Deputy Subject Leader overseeing the postgraduate courses in computing. His main research interests include software engineering, pedagogy (assessments and marking), constraints, and knowledge management. He is a fellow of the Higher Education Academy.

Lee R. Machado received the bachelor's degree from the University of Warwick, and the Ph.D. degree in cancer studies from the University of Birmingham, U. K. He is currently a professor of molecular medicine with the Division of Life Sciences, a Faculty Research and Enterprise Lead, the Co-Leader of the Molecular Biosciences Research Group, Physical Activity and Life Sciences Centre, University of Northampton, and an Honorary Research Fellow with the Department of Genetics and Genome Biology, University of Leicester. He did postdoctoral research at the Institute for Cancer Studies, Birmingham, the MRC Toxicology Unit, Leicester, and the Department of Genetics, Leicester, before working as a Senior Scientist with Cancer Vaccine Company, and Scancell, Nottingham. He joined the University of Northampton, as a Lecturer, in 2013. He was an Interim Head of Sport, Exercise and Life Sciences, from 2017 to 2018. He has three years of University Board level experience. His research interests include employing cellular and molecular genetic strategies to address how the host immune system responds to pathogens and cancer. The aim of this work is to increase our understanding of human health and disease and develop rational therapeutic approaches to harness the exquisite specificity and sensitivity of the immune system.

Michael Opoku Agyeman (Senior Member, IEEE) received the Ph.D. degree from Glasgow Caledonian University, U.K. He is currently a Professor and a Program Leader of computer systems engineering at the University of Northampton (UoN), U.K. He represents the Research Community of UoN at the University Senate. He is the Postgraduate (PGR) Lead of the Faculty of Arts Science and Technology and Co-Chairs the University's PGR Supervisory Forum. He has over ten years' experience in embedded systems engineering. Previously, he was a Research Fellow with Intel Embedded System Research Group, The Chinese University of Hong Kong (CUHK). He is the author of five books, two book chapters, and over 80 publications in major journals and conference proceedings. His main research interests include 3 main strands and disciplines: embedded systems and high-performance computing, such as VLSI SoC design, computer architecture, reconfigurable computing, wired and wireless NoCs, smart rehabilitation solutions, embedded systems and the Internet of Things (IoT); business administration, such as neuromarketing, advertising, and market research; and pedagogy. He is a fellow of the Higher Education Academy (UK). He is a Technical Committee Member of several conferences, such as IEEE ICCSN, IEEE ICBDA, and IEEE ICCT. His work on wireless NoC has attracted two best paper awards in IEEE/IFIP EUC 2016 and Euromicro DSD 2016, respectively. He was a recipient of the 2018 International Changemaker of the year award in the first U.K. Ashoka U Changemaker Campus. He serves as a Reviewer of several conferences and journals, including IEEE Access. He has been a Guest Editor of the EAI Endorsed Transactions on Industrial Networks and Intelligent Systems. He is a Chartered Engineer (C.Eng.) of the IET and a Chartered Manager (C.Mgr.) of CMI.