Computer Assisted Methods in Engineering and Science
31(2): 261–272, 2024, doi: 10.24423/cames.2024.555

This article belongs to the *Special Issue on Scientific Computing and Learning Analytics for Smart Healthcare Systems* edited by Dr. C. Chakraborty, Dr. S. Barbosa and Dr. L. Garg

Measuring Comparative Statistical Effectiveness of Cancer Subtype Categorization Using Gene Expression Data

Avila Clemenshia P.*, Deepa C.

Department of Computer Science, Sri Ramakrishna College of Arts & Science, India; e-mail: deepapkd@gmail.com

* *Corresponding Author e-mail: avilajsph@gmail.com*

This work focused on the analysis of various gene expression-based cancer subtype classification approaches. Correctly classifying cancer subtypes is critical for understanding cancer pathophysiology and effectively treating cancer patients by using gene expression data to categorize cancer subtypes. When dealing with limited samples and high-dimensional biological data, most classifiers may suffer from overfitting and lower precision. The goal of this research is to develop a machine learning (ML) system capable of classifying human cancer subtypes based on gene expression data in cancer cells. These issues can be solved using ML algorithms such as Transductive Support Vector Machines (TSVM), Boosting Cascade Deep Forest (BCD Forest), Enhanced Neural Network Classifier (ENNC), Deep Flexible Neural Forest (DFN Forest), Convolutional Neural Network (CNN), and Cascade Flexible Neural Forest (CFN Forest). In inferring the benefits and drawbacks of these strategies, such as DFN Forest and CFN Forest, the findings are 95%.

Keywords: cancer subtypes, gene expression data, machine learning, Deep Flexible Neural Forest (DFN Forest) strategy.

1. Introduction

According to the WHO, cancer is now the world's second leading cause of death, accounting for one out of every six deaths. Moreover, malignancies are exceedingly diverse, as the results might vary substantially in outcomes for patients with comparable diagnoses who receive the same treatment regimens. In addition, instead of being a single ailment, cancer frequently encompasses many subtypes in multiple molecular pathogeneses and clinical aspects [1]. Therefore, it is vital to recognize cancer subtypes to ease cancer identification and therapy [2].

Thousands of genes are generally expressed in thousands of samples. For example, the Cancer Genome Atlas (TCGA) pilot experiment sequenced over

10 000 patient samples from 33 cancer types [3]. Using computational approaches, researchers can better predict cancer subtypes using diverse genome-wide data. However, defining cancer subtypes based on gene expression data is challenging and complex. Thus, an efficient model for better classification is required.

Using cancer multi-omics data, deep neural network-based subtype classification and prognostic models have recently been introduced. In such investigations, an autoencoder is a type of unsupervised deep neural network designed to identify patterns in enormous amounts of data. Autoencoders and multi-omics data were used to make prognostic assessments for neuroblastoma and liver cancer, for example in [4]. A prognostic model was developed to estimate patient survival times using the bottleneck layer values as variables. In terms of prediction error, the proposed model outperformed shallow models based on their benchmarks. Furthermore, autoencoders were utilized to categorize breast cancer subtypes based on miRNA, mRNA, and DNA methylation levels [5]. Unsupervised learning with autoencoders can successfully capture features through dimensional reduction. These models, on the other hand, may be unsuitable for classification because they do not associate with the response variable during feature selection.

Gene expression data is distinguished by high feature dimensionality, missing values, small sample volume, noise, and redundant data [6]. To improve precision, gene selection and cancer categorization have been combined, with relevant genes solely used to process the categorization. A variety of feature selection and categorization approaches have been employed, each with its benefits. Feature selection aims to reduce overfitting, improve strategy effectiveness, and produce faster and cost-effective strategies. Various feature selection techniques were used in [7] employing differential expression analysis and the minimum-Redundancy Maximum-Relevance (mRMR) approach for selecting microarray and RNA-seq data characteristics. Using RNA-seq data, García-Díaz *et al.* [8] applied a genetic clustering algorithm to five distinct malignancies [9]. Uninformative genes were removed using iterative feature elimination [10], while particle swarm and decision tree algorithms simultaneously selected genes and categorized cancers. A gene expression-based categorization method was used in [11].

Classification is a method for identifying new data categories. In recent years, many supervised learning approaches and procedures have been used to classify cancer subtypes. For example, Nguyen *et al.* [12] employed designed supervised learning hidden Markov strategies to classify cancers based on gene expression characteristics. In [13], OncoNet Explainer was developed to make explainable cancer type recognitions using Gene Expression (GE) data. However, a few flaws may limit cancer genomic data applications. On one side, these techniques often involve complex strategies that require a lot of data to train.

Section 2 of this study presents literature review and examines the approaches to cancer subtype recognition using gene selection, along with their benefits and

drawbacks. In addition, previous cancer subtype recognition algorithms are compared in the table, focusing on their merits and disadvantages. Section 3 discusses the research gap from previous works and the challenges in analyzing gene expression data in cancer research. The need for the creation of new mathematical and statistical approaches for analyzing such heterogeneous data is emphasized in Sec. 3. Section 4 discusses the experimental results and discussion with tabulated results. Section 5 briefly discusses inferences from presented studies. Section 6 concludes the study.

2. Literature review

Maulik *et al.* [14] developed TSVM to predict gene expression-based cancer subtypes. This approach also assists in identifying potential gene markers for each cancer subtype, thereby allowing for a more accurate cancer recognition. However, the limited sample size still hinders classifier design. Ordinary supervised classifiers require labeled data, but, many microarray data with insufficient follow-up are often ignored. A novel feature selection and TSVM approach is proposed. The TSVM was designed using selected genes from the microarray data. The suggested strategy outperforms standard inductive SVM (ISVM) and low-density separation (LDS) in semi-supervised cancer categorization and gene marker recognition.

Using the intricate network of miRNA-TF-mRNA regulation, Xu *et al.* [15] presented a method to detect cancer subtypes, employing Weighted Similarity Network Fusion (WSNF). First, a regulatory network was constructed with nodes representing microRNAs, transcription factors, messenger RNAs (mRNAs), and edges indicating trait connections. The network data, as well as miRNA, TF, and mRNA expression data were then used to calculate the weight of the attributes, which represents their importance. The TCGA Breast Invasive Carcinoma (BRCA) and Glioblastoma Multiforme (GBM) datasets were used in the work. The WSNF technique correctly identified five breast cancer subtypes and three GBM subtypes.

Salem *et al.* [16] proposed an innovative strategy for classifying human cancer illnesses. The developed technique integrates both Data Gain (IG) and Standard Genetic Algorithm (SGA). It initially utilizes Data Gain to select features, then GA to reduce and refine these features, and finally Genetic Programming (GP) to categorize cancer types. Applying the proposed process to cancer datasets compared to other ML methodologies demonstrates that no categorization strategy consistently outperforms all others, Nonetheless, GAs improve the recognition rate of multiple classifiers in general.

Zhou and Feng [17] proposed the gcForest approach that enables representation learning by forest. Cascade forest and multi-grained scanning are the two

components of the gcForest technology. The convolution technique used in convolutional neural networks is simplified in multi-grained scanning. When inputting high-dimension sample data, multi-grained scanning may capture numerous levels of data by segmenting the high-dimension sample data into multiple-scale sequences of characteristics by using sliding windows of multiple sizes, allowing gcForest to be contextually or structurally aware. This technique can learn more specific characteristics and produce more accurate results [17].

Guo *et al.* [18] created the Boosting Cascade Deep Forest (BCD Forest) algorithm to identify cancer subtypes using small-scale biological data. The BCD Forest technique differs from the standard deep forest strategy in two important ways: firstly, it introduces a multi-class-grained scanning technique for training multiple binary classifiers to increase ensemble diversity. Secondly, it incorporates a boosting approach to highlight more significant characteristics in cascade forests. When applied to identify cancer subtypes, comparative studies reveal that the proposed technique achieves a higher recognition rate.

Guo *et al.* [19] developed the Cancer Subtype Prediction Using RV2 (CSPRV) approach, which enhances cancer subtype identification, by integrating multi-source transcriptome expression data and multiple biological networks. Using a generalized matrix correlation technique, the approach predicts similarities between samples in each view of expression data (RV2). The presented approach may recognize clinically important cancer subgroups, through tests using TCGA cancer datasets.

Vasudevan and Murugesan [20] created a prognosis-enhanced neural network classifier to detect cancer subtypes in multigenic data. To begin, max-flow/min-cut graph clustering is utilized to discover potential cancer clusterings. They employed 215 samples with microRNA expression to define glioblastoma multiforme subgroups (12 042 genes). The samples were classified into four types based on mutations and gene expression patterns: mesenchymal, classical, proneural, and neural. Finally, A-measure and f-measure were used to assess the outcomes.

Lee *et al.* [21] created a Cancer Predictor utilizing an Ensemble Model (CPEM). The researchers also evaluated input properties such as mutation patterns, rates, spectra, and signatures. Then they tested various ML and feature selection strategies, eventually settling on one that achieved 84% precision using ten-fold cross-validation. They also employed the six most frequent malignancies out of 31 types, and the strategy correctly classified 94% of them.

Dong *et al.* [22] developed MLW-gcForest to categorize cancer subtypes. This algorithm's key contributions are: (1) allocating weights to random forests based on their categorization abilities, and (2) creating a sorting optimization technique that provides varying weights to sliding window feature vectors. The MLW-gcForest approach was trained using methylation data from five cancer genome atlases (TCGA). The MLW-gcForest method produces good preci-

sion and AUC values for categorizing cancer subtypes. Also, data on methylation can be used to diagnose cancer.

To identify cancer subtypes, Xu *et al.* [23] developed a Deep Flexible Neural Forest (DFN Forest) technique. The developed DFN Forest strategy converts a multi-categorization issue into numerous binary categorization issues for each forest. In addition to the DFN Forest technique, this study proposes a Fisher Ratio and Neighborhood Rough Set Combination for lowering the dimensionality of gene expression data. On the other hand, the developed DFN Forest strategy better categorizes cancer subtypes.

Mostavi *et al.* [24] employed three CNN algorithms to categorize tumor and non-tumor samples as malignant or normal. These three algorithms were: 1D-CNN, 2D-Vanilla-CNN, and 2D-Hybrid-CNN.The approaches were put to the test using gene expression data from TCGAs 10 340 cancer samples and 713 normal tissues. Among 34 classes, the designed strategies provided outstanding recognition precision (93.9–95%) (33 cancers and normal tissues). With the 1D-CNN approach, 2090 cancer indicators were identified using directed saliency. For the classification of breast cancer subtypes, the 1D-CNN strategy attained an average precision of 88.42% across five subtypes.

Zhong *et al.* [25] created a CFN Forest strategy to categorize cancer subtypes. To create the strategy's structure and parameters, the Flexible Neural Tree (FNT) Group Forest was developed by extending the conventional flexible neural tree structure of the CFN Forest. The FNT Group Forest strategy was built using the multilayer cascade structure, with varying characteristics between tiers, enhancing strategy effectiveness. The CFN Forest strategy improved operational efficiency and resilience by using sample selection across layers and assigning multiple weights for each layer's output. Additionally, the FNT Group Forest was employed with several feature sets to increase strategy structural variety and suitability for small sample size datasets to categorize cancer subtypes. In addition, using RNA-seq gene expression data, CFN Forest increases the accuracy of cancer subtype classification.

Zhong *et al.* [26] proposed a Laminar Augmented Cascading Flexible Neural Forest (LACFN Forest) strategy to categorize cancer subtypes. This strategy uses the DFN Forest as the basic classifier. Each layer of the forest features an output judgment mechanism to lower the strategy's computing complexity. The highly connected deep neural forest was used to improve recognition results. Using data from RNA-seq gene expression, LACFN Forest performed better in subtype categorization.

Majji *et al.* [27] proposed a distinct deep recurrent neural network (Jaya ALO-based Deep RNN) for cancer classification. An algorithmic approach was used to create the strategy. Normalization is the initial stage. Data normalization eliminates data redundancy and reduces object storage in relational databases

that store the same data in several places. Then the data is transformed using log transformation to make the patterns more understandable, reduce skew, and help satisfy the assumptions. The feature dimension is also reduced via non-negative matrix factorization. The Jaya ALO-based Deep RNN showed enhanced results with 95.97% precision, 95.95% sensitivity, and 96.96% specificity.

Yu *et al.* [28] developed weighted differentially expressed genes (weighted DEGs) by combining gene regulatory network biological relevance with differential expression analysis. A high-weight gene has a bigger biological influence since it regulates more target genes. The drastically diverse interaction topologies inspired the authors to design a Gene Ontology (GO) enrichment based on gene coexpression networks (GOEGCN). For the control and experimental groups, the GOEGCN considers a two-sided distinction analysis of gene coexpression networks. Using this technique, researchers can now investigate how regulated co-expressed gene pairings influence biological processes at the GO level.

Jaber *et al.* [29] developed a deep learning system for calculating PAM50 intrinsic subtyping in breast cancer using only whole-slide images of H&E-stained breast biopsy tissue sections, which is a simple method for diagnosing intrinsic molecular subtype (IMS) in breast cancer. This approach was trained on images of 443 previously PAM50 subtyped tumors to classify microscopic parts of the images into four major molecular subtypes: basal-like, HER2-enriched, Luminal A, and Luminal B, aa well as distinguishing between basal vs. non-basal subtypes. After then, the method was utilized to classify the subtypes of 222 previously unclassified tumors.

Chakraborty *et al.* [30] to assess complicated biological data, proposed a hybrid Machine Learning Classification Techniques based on ensemble methodology with Enhanced-Grey Wolf Optimization (E-GWO) feature selection algorithm. For the experimental, the authors merged five biological heart disease data sets: Cleveland, Long Beach, VA, Switzerland, Hungarians, and Statlog. Bagging and boosting methods are used to create new hybrid Machines Learning Classification Techniques classifiers such as Naive Bayes Bagging Technique (NBBT), Random Forest Bagging Technique (RFBT), Decision Tree Bagging Technique (DTBT), K-Nearest Neighbors Bagging Technique (KNNBT), Neural Network Bagging Technique (NNBT), Gradient Boosting Boosting Technique (GBBT), and Adaptive Boosting Boosting Technique (ABBT). To evaluate hybrid approaches, accuracy, recall, precision, F1-score, specificity, error rate, G-mean, false-negative rate (FNR), false-positive rate (FPR), and negative predictive value (NPV) are utilized. Experimental results revealed that with E-GWO feature selection, the designed hybrid classifier RFBT obtains the highest accuracy of 99.26%. The proposed strategy enhanced the accuracy of the conventional model by 11.90%.

Table 1 shows the comparative analysis of previous cancer subtype recognition methods.

Table 1. Comparative analysis of previous cancer subtype recognition methods.

No.	Authors name	Methods	Merits	Disadvantages
1.	Maulik *et al.* [14]	TSVM	It improves the recognition effectiveness	It may converge to a local optimum
2.	Xu *et al.* [15]	WSNF	Better categorization precision	It does not cover all true interactions
3.	Salem *et al.* [16]	Genetic Programming (GP)	Higher actual positive rate and minimum root mean squared error	It has an issue with time complexity and precision
4.	Zhou & Feng [17]	gcForest	Minimum execution time and better precision	It only improves the categorization effectiveness of the strategy on small samples
5.	Guo *et al.* [18]	Boosting Cascade Deep Forest (BCD Forest)	Higher categorization precision	Better gene selection mechanisms are required to achieve improved precision
6.	Guo *et al.* [19]	Cancer Subtype Recognition using RV2 (CSPRV)	Higher recognition precision	The PCA has an issue with data loss
7.	Vasudevan and Murugesan [20]	Enhanced neural network classifier	Higher precision and higher recall	Better dimension reduction mechanisms are required to achieve improved precision and precision
8.	Lee *et al.* [21]	Cancer Predictor using an Ensemble Model (CPEM)	It improves the recognition of cancer susceptibility	It only achieves 94% of categorization precision
9.	Dong *et al.* [22]	MLW-gcForest	It attains high precision and Area Under Curve (AUC) values	Computational complexity
10.	Xu *et al.* [23]	DFN Forest	High true positive rate and minimum execution time	It has low precision due to the inability to select relevant features and the lengthy time required for the categorization procedure
11.	Mostavi *et al.* [24]	Convolutional Neural Network (CNN)	The CNN is slower due to an operation such as a max pool	It only achieves an average precision of 88.42%
12.	Zhong *et al.* [25]	Cascade Flexible Neural Forest (CFN Forest)	It effectively improves the precision and displays good robustness	When dealing with tiny sample sizes and complex biology data, it suffers from overfitting and low categorization precision
13.	Zhong *et al.* [26]	LACFN Forest	Better recall and f-measure	It has an issue with the inability to choose the appropriate characteristics
14.	Majji *et al.* [27]	Jaya ALO-based Deep RNN	Higher sensitivity and specificity	Better methods are required to achieve improved categorization precision

3. Research gap

In the past, researchers have analyzed gene expression data using various mathematical and statistical methods for a variety of reasons. The discovery of relevant gene-related circuits enhanced disease classification, prediction, medication development, and individualized therapy. To achieve these objectives, various methods have been developed. However, all techniques are hindered by the complexities and high dimensionality of gene expression data. Furthermore, the number of individual cancers is vast with at least >100 molecularly different forms of cancer. Additionally, the technologies used to calculate gene expression across the genome are evolving, leading to greater precision in gene expression estimates. Instead of using DNA microarrays, RNA-seq can easily identify novel gene expression patterns (novel transcripts). As a result, technological advancement necessitates the creation of new mathematical and statistical approaches for analyzing such heterogeneous data. Another issue is the presence of other interacting elements (e.g., environmental factors). For example, smoking, asbestos, and nutritional variables are likely to interact with and alter genes associated with particular cancers. Recent studies that use the machine and deep learning methodologies to predict cancer and identify biomarker genes need to be evaluated.

4. Experimental results

MATLAB is used to conduct comparison. The subtypes were determined using RNA-seq gene expression datasets. There are four primary subgroups of BRCA in 514 BRCA samples: basal-like, HER2-enriched, Luminal-A, and Luminal-B. Furthermore, 164 GBM samples were classified as Classical, Mesenchymal, Neural, or Proneural. while 275 LUNG samples have Bronchioid, Magnoid, and Squamoid subtypes. This study evaluates the precision and recall results of various techniques such as CFN Forest, gcForest, BCD-Forest, MLW gcForest and DFN Forest. Tables 2 and 3 clearly tabulated the precision and recall results.

4.1. Precision

Precision in classification refers to the ratio of correctly predicted positive observations to the total predicted positive observations. Precision is calculated as the number of true positive predictions divided by the sum of true positive and false positive predictions:

$$\text{Precision} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})}, \tag{1}$$

where TP – true positive, TN – true negative, FP – false positive, and FN – false negative.

Table 2. Precision comparison.

Dataset	Precision [%]				
	CFN Forest	gcForest	BCD-Forest	MLW gcForest algorithm	DFN Forest
BRCA	94.4	85.2	92.0	91.5	93.6
GBM	92.1	83.6	80.6	88.5	84.2
LUNG	90.9	84.4	86.7	90.2	88.0

The precision of the CFN Forest, gcForest, BCD-Forest, MLW gcForest, and DFN Forest algorithms is assessed. Datasets are taken on the x-plane, and precision is taken on the y-axis. According to the test findings, the CFN Forest, gcForest, BCD-Forest, MLW gcForest, and DFN Forest techniques achieve precision of 94.4%, 85.2%, 92.0%, 91.5%, and 93.6%, respectively, on the BRCA dataset.

4.2. Recall

Recall in classification refers to the ratio of correctly predicted positive observations to the actual positive observations in the dataset. Recall is calculated as the number of true positive predictions divided by the sum of true positive and false negative predictions:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

Table 3. Recall comparison.

Dataset	Recall [%]				
	CFN Forest	gcForest	BCD-Forest	MLW gcForest algorithm	DFN Forest
BRCA	96.0	82.6	92.0	91.6	96.5
GBM	88.8	85.0	80.6	87.8	80.2
LUNG	90.5	81.9	86.7	85.2	85.0

In terms of recall, the CFN Forest, gcForest, BCD-Forest, MLW gcForest, and DFN Forest approaches are examined. The datasets are represented on the x-axis, and the recall is represented on the y-axis. According to the test findings, the recall of CFN Forest, gcForest, BCD-Forest, MLW gcForest, and DFN Forest techniques is 96%, 82.6%, 92.0%, 91.6%, and 96.5%, respectively, on the BRCA dataset.

5. Inferences from existing works

Cancer can be detected using various methods based on gene expression data. According to this research, it can be concluded that traditional cancer subtype classification is outperformed by gene expression-based classification. However, because this type of data has thousands of variables, categorization is challenging to execute without efficient and precise algorithms. Additionally, TSVM, BCD Forest, Enhanced Neural Network Classifier, DFN Forest strategy, CNN, CFN Forest strategy, GP, and LACFN Forest strategy encounter issues with categorization precision, actual positive rate, and time complexity.

6. Conclusion

Cancer subtype classification is critical in cancer detection. Accurate subtype categorization helps clinicians in understanding cancer etiology and primary site, which is crucial for cancer genesis research. In addition, subtype categorization of cancer has several applications in supervised learning applications. Algorithms, including TSVM, BCD Forest, Enhanced Neural Network Classifier, DFN Forest strategy, CNN, CFN Forest strategy, GP, and LACFN Forest strategy, regularly outperform current approaches in experiments. However, existing cancer subtype recognition methods still face challenges to perform well. Datasets that are routinely used are highlighted to evaluate ML models in development using gene expression data. There are still challenges in deciphering gene expression data, but those challenges can be used as stepping stones for future researchers, leading to discoveries that could aid in improved cancer categorization and, eventually, personalized treatments. According to the results of the experiments, deep learning algorithms outperform traditional machine learning techniques in analyzing cancer gene expression data.

References

1. K.H. Park, V.H. Pham, K. Davagdorj, L. Munkhdalai, K.H. Ryu, A subtype classification of hematopoietic cancer using a machine learning approach, [in:] *Recent Challenges in Intelligent Information and Database Systems Asian Conference on Intelligent Data and Database Systems, (ACIIDS 2021)*, T.P. Hong, K. Wojtkiewicz, R. Chawuthai, P. Sitek [Eds], Communications in Computer and Information Science, Vol. 1371, Springer, Singapore, pp. 113–121, 2021, doi: 10.1007/978-981-16-1685-3_10.

2. L. Zhang *et al.*, Elastic net regularized Softmax regression methods for multi-subtype categorization in cancer, *Current Bioinformatics*, **15**(3): 212–224, 2020, doi: 10.2174/1574893615666200106181112141724.

3. J.T.-H. Chang, Y.M. Lee, R.S. Huang, The impact of the Cancer Genome Atlas on lung cancer, *Translational Research*, **166**(6): 568–585, 2015, doi: 10.1016/j.trsl.2015.08.001.

4. K. Chaudhary, O.B. Poirion, L. Lu, L.X. Garmire, Deep learning-based multi-omics integration robustly predicts survival in liver cancer, *Clinical Cancer Research*, **24**(6): 1248–1259, 2018, doi: 10.1158/1078-0432.CCR-17-0853.

5. Y. Guo, X. Shang, Z. Li, Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer, *Neurocomputing*, **324**(1): 20–30, 2019, doi: 10.1016/j.neucom.2018.03.072.

6. H. Liu, J. Li, L. Wong, A comparative study on feature selection and categorization methods using gene expression profiles and proteomic patterns, *Genome Informatics*, **13**: 51–60, 2002, doi: 10.11234/gi1990.13.51.

7. D. Castillo *et al.*, Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level, *PloS ONE*, **14**(2): e0212127, 2019, doi: 10.1371/journal.pone.0212127.

8. P. García-Díaz, I. Sánchez-Berriel, J.A. Martínez-Rojas, A.M. Diez-Pascual, Unsupervised feature selection algorithm for multiclass cancer categorization of gene expression RNA-Seq data, *Genomics* **112**(2): 1916–1925, 2020, doi: 10.1016/j.ygeno.2019.11.004.

9. S. Ramaswamy *et al.*, Multiclass cancer diagnosis using tumor gene expression signatures, *Proceedings of the National Academy of Sciences*, **98**(26): 15149–15154, 2001, doi: 10.1073/pnas.211566398.

10. K.H. Chen *et al.*, Gene selection for cancer identification: A decision tree strategy empowered by particle swarm optimization algorithm, *BMC Bioinformatics*, **15**(1): 49, 2014, doi: 10.1186/1471-2105-15-49.

11. L. Goh, Q. Song, N.K. Kasabov, A novel feature selection method to improve categorization of gene expression data, [in:] *APBC '04: Proceedings of the Second Conference on Asia-Pacific Bioinformatics*, Dunedin, New Zealand, Vol. 29, pp. 161–166, 2004, doi: 10.5555/976520.976542.

12. T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, Hidden Markov models for cancer classification using gene expression profiles, *Information Sciences*, **316**: 293–307, 2015, doi: 10.1016/j.ins.2015.04.012.

13. Md.R. Karim, M. Cochez, O. Beyan, S. Decker, C. Lange, OncoNetExplainer: Explainable predictions of cancer types based on gene expression data, *IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, Athens, Greece, pp. 415–422, 2019, doi: 10.1109/BIBE.2019.00081.

14. U. Maulik, A. Mukhopadhyay, D. Chakraborty, Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM, *IEEE Transactions on Biomedical Engineering*, **60**(4): 1111–1117, 2013, doi: 10.1109/TBME.2012.2225622.

15. T. Xu, T.D. Le, L. Liu, R. Wang, B. Sun, J. Li, Identifying cancer subtypes from miRNA-TF-mRNA regulatory networks and expression data, *PloS ONE*, **11**(4): e0152792, 2016, doi: 10.1371/journal.pone.0152792.

16. H. Salem, G. Attiya, N. El-Fishawy, Classification of human cancer diseases by gene expression profiles, *Applied Soft Computing*, **50**(1): 124–134, 2017, doi: 10.1016/j.asoc.2016.11.026.

17. Z.H. Zhou, J. Feng, Deep forest, *National Science Review*, **6**(1): 74–86, 2019, doi: 10.1093/nsr/nwx118.

18. Y. Guo, S. Liu, Z. Li, X. Shang, BCDForest: A boosting cascade deep forest strategy towards the classification of cancer subtypes based on gene expression data, *BMC Bioinformatics*, **19**(Suppl. 5): 118, 2018, doi: 10.1186/s12859-018-2095-4.

19. Y. Guo, Y. Qi, Z. Li, X. Shang, Improvement of cancer subtype prediction by incorporating transcriptome expression data and heterogeneous biological networks, *BMC Medical Genomics*, **11**(Suppl. 6): 87–98, 2018, doi: 10.1186/s12920-018-0435-x.

20. P. Vasudevan, T. Murugesan, Cancer subtype discovery using prognosis-enhanced neural network classifier in multigenomic data, *Technology in Cancer Research & Treatment*, **17**(7): 1–13, 2018, doi: 10.1177/1533033818790509.

21. K. Lee, H.O. Jeong, S. Lee, W.K. Jeong, CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network, *Scientific Reports*, **9**(1): 16927, 2019, doi: 10.1038/s41598-019-53034-3.

22. Y. Dong, W. Yang, J. Wang, J. Zhao, Y. Qiang, MLW-gcForest: A multi-weighted gcForest model for cancer subtype classification by methylation data, *Applied Sciences*, **9**(17): 3589, 2019, doi: 10.3390/app9173589.

23. J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, M.M. Khan, A novel deep flexible neural forest strategy for classification of cancer subtypes based on gene expression data, *IEEE Access*, **7**(2): 22086–22095, 2019, doi: 10.1109/ACCESS.2019.2898723.

24. M. Mostavi, Y.C. Chiu, Y. Huang, Y. Chen, Convolutional neural network strategies for cancer type prediction based on gene expression, *BMC Medical Genomics*, **13**(5): 44, 2020, doi: 10.1186/s12920-020-0677-2.

25. L. Zhong, Q. Meng, Y. Chen, A cascade flexible neural forest strategy for cancer subtype categorization on gene expression data, *Computational Intelligence and Neuroscience*, **2021**: 6480456, 2021, doi: 10.1155/2021/6480456.

26. L. Zhong, Q. Meng, Y. Chen, L. Du, P. Wu, A laminar augmented cascading flexible neural forest strategy for classification of cancer subtypes based on gene expression data, *BMC Bioinformatics*, **22**(1): 475, 2021, doi: 10.1186/s12859-021-04391-2.

27. R. Majji, G. Nalinipriya, C. Vidyadhari, R. Cristin, Jaya ant lion optimization-driven deep recurrent neural network for cancer categorization using gene expression data, *Medical & Biological Engineering & Computing*, **59**(5): 1005–1021, 2021, doi: 10.1007/s11517-021-02350-w.

28. Z. Yu, Z. Wang, X. Yu, Z. Zhang, RNA-seq-based breast cancer subtypes classification using machine learning approaches, *Computational Intelligence and Neuroscience*, **2020**: 4737969, 2020, doi: 10.1155/2020/4737969.

29. M.I. Jaber *et al.*, A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that may affect survival, *Breast Cancer Research*, **22**(1): 12, 2020, doi: 10.1186/s13058-020-1248-3.

30. C. Chakraborty, A. Kishor, J.J.P.C. Rodrigues, Novel enhanced-grey wolf optimization hybrid machine learning technique for biomedical data computation, *Computers and Electrical Engineering*, **99**(6): 107778, 2022, doi: 10.1016/j.compeleceng.2022.107778.

Received March 24, 2022; revised version August 14, 2022; accepted August 22, 2022; published online June 6, 2024.