IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 July 2024, accepted 5 September 2024, date of publication 16 September 2024, date of current version 27 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3461878

RESEARCH ARTICLE

Multimodal Non-Small Cell Lung Cancer Classification Using Convolutional Neural Networks

MARIAN MAGDY AMIN, AHMED S. ISMAIL, AND MASOUD E. SHAHEEN
Faculty of Computers and Artificial Intelligence, Fayoum University, Faiyum 2933110, Egypt
Corresponding author: Marian Magdy Amin (mm6252@fayoum.edu.eg)

ABSTRACT Lung cancer is the leading cause of death worldwide. Early detection of lung cancer is a hard mission. New Small Cell Lung Cancer (NSCLC) is the most prevalent sub-type of lung cancer. Differentiating between several NSCLC subtypes is important for making the right decision of treatment plan for the patient. Despite the focus of recent researchers on single modality approach, multi-omics modalities have many underlying influences and discoveries in the cancer detection and classification area. Through this research multi-omics modalities are used. Previous efforts have been focused either on multimodality using traditional machine learning classifiers or single modality using deep learning. Also, for the molecular sources (RNA-seq and miRNA-Seq) traditional machine learning approaches are usually used. For this work, deep learning using Convolutional Neural Networks (CNNs) is used and applied on the above-mentioned multimodalities. The classification accuracy results obtained for RNA-Seq, miRNA-Seq, WSIs are 96.79%, 98.59%, 89.73% respectively. The F1 scores obtained for RNA-Seq, miRNA-Seq, WSIs are 95.238%,99.67%,89.76% respectively. Moreover, the Area Under Curve obtained for RNA-Seq, miRNA-Seq, WSIs are 100%, 99.41%,97,54% respectively. These results improves the results obtained by other related works as will be explained. According to these improvements in the results, the lung cancer classification could be better and the disease would be discovered at early stages which is the goal for the research field efforts.

INDEX TERMS Lung cancer, convolutional neural networks, multimodality, molecular data, whole slide images, deep learning, multiomics.

I. INTRODUCTION

Lung cancer ranks as the primary cause of death among men and the second leading cause among women globally [1], [2]. It is widely acknowledged as one of the most lethal forms of cancer, surpassing breast, prostate, colorectal, stomach, and liver cancer in terms of severity [3].

Efforts to detect lung cancer early have become paramount due to its aggressive nature. Non-Small Cell Lung Cancer (NSCLC) accounts for the majority of cases, comprising approximately 80-85% of diagnoses [2], [4]. Despite the high mortality rates and the aggressive nature of cancer, many patients could be saved. As declared by the World Health Organization 30-50% of cases can be avoided if they can be early detected and received the appropriate treatment [3], [5]. However, cancer is often detected at late stages where effective treatment and cures is unachievable [3]. The complex nature of cancer molecular biology made it difficult for professionals to detect early signs and symptoms using traditional diagnosis and screening techniques. Moreover, one major challenge when using chemotherapy is maximizing the drug efficiency while minimizing the toxic effects of healthy cells. So, for obtaining successful treatment, accurate classification for the tumor is needed. But classification of laboratory screenings depends on the insights of experts and morphological appearance of the tumor. But this approach inherits serious limitations because of the similar histopathological appearances of the tumors which have significant different courses and responses to therapy [3]. As a result, deeper approaches should be followed for accurate classification.

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

Studies have been focused recently on the genetic molecular level for early diagnosis and early treatment plans.

Within NSCLC, two predominant subtypes exist: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). LUAD typically develops in the peripheral lung tissue, whereas LUSC tends to manifest in central locations [2]. Accurate differentiation between these subtypes is crucial as they necessitate distinct treatment approaches [3], [6].

Next Generation Sequencing methods such as whole-genome DNA sequencing and total RNA sequencing, are considered revolutionary technologies for studying genetic changes in cancer [7]. These technologies are promising in accurate classification of tumor cells because of their ability of sequencing thousands of genes and detecting genomic and transcriptomic alterations [3]. NGS achieves that through comparing DNA and RNA in normal and cancer cells. It discovers the genetic changes that leads to abnormal activity that leads to the presence of different levels of genes and consequently proteins within the cells [3], [8]. This is known as gene expression. Differentially expressed genes within the cell gives great insight of the motive of tumor growth [3], [8].

Gene expression has been the focus of recent researches in the field of cancer classification [3], [9]. Cell functions are determined by individual proteins and the synthesis of these proteins depends on which genes are expressed by the cell. Therefore, the gene expression gives some insight about the cell function [3]. Gene expression is the process of translating DNA into proteins and non-coding RNA. Microarray shows limitation because of the incomplete snapshot of the transcriptome. Also, it cannot detect previously unidentified genes or transcripts [3], [9]. Gene expression quantification, therefore, is an effective alternative. It identifies which genes are preferentially expressed in different tissues. So, in this paper STAR counts data provided by TCGA program is used.

Through this paper two types of gene expression data are used mRNA-Seq and miRNA-Seq. mRNA is involved in conveying genetic information from DNA to ribosome, where it is translated into proteins. While miRNA is a type of non-coding RNA which plays a role in gene regulation by binding to target mRNAs and inhibiting their translation or promoting their degradation. Any abnormalities in these processes of mRNA or miRNA leads to different gene expressions which accordingly activates tumor growth [3], [10]. However, cancer classification using gene expression is very challenging given the complexity and massive amount of genetic data that is produced [3], [9], [11], [12]. The magnitude of variant obtained from RNA-Sequencing for example is exponential which makes it difficult for traditional machine learning methods and bioinformatics tools to approach genetic variants for disease prediction [3], [13], [14]. Gene expression is known by high dimensionality, with w very large number of features representing genes and very small number of training data representing patient samples [3], [13], [15]. Deep learning has been used recently for dealing with this problem of high dimensionality [16].

Deep learning has been massively used using single modality (images in most of cases) for detecting different types of cancer. In this paper, multi-omics modalities are used for lung cancer classification such as: RNA-Seq [2], [17]and miRNA-Seq [2], [18]. Moreover, Whole Slide Images (WSIs) [2], [19] are used. So, in this paper multimodalities are used RNA-Seq, miRNA-Seq, and WSIs.

Despite the focus of this paper on lung cancer, the idea itself is effective when being applied on multiple large number of tumors (multi-scale, multi-modalities). The architecture of the deep learning model requires minimal preprocessing for multi-layer networks [20]. It utilizes the spatial relationships within the data to decrease the dimensions that need to be learned [15]. This approach speeds up the learning process while yielding more precise results.

So, NGS using deep learning techniques is the suitable solution for the problem of complexity and high dimensionality. Convolutional Neural Networks are used in this work for both slide images and molecular sources (RNA-Seq (mRNA), miRNA-seq) for dealing with high dimensionality problem.

Through this research several past efforts will be shown through the related works to highlight the importance of reaching an accurate classification for this type of cancer. Also, the dataset used and the preprocessing done will be discussed in the methodology section. Then, a detailed explanation of the approach used of the convolutional neural networks will be illustrated in the implementation and discussion section. After that, the results will be shown to compare with the previously obtained results by other researches. Also, future work points that could be made will be shown in the future work section.

II. RELATED WORKS

In recent years, there has been significant interest in using machine learning models with biological data to aid in the diagnosis and prognosis of cancer patients, particularly in the context of lung cancer. Gene expression data has been widely explored for lung cancer classification, with studies achieving high accuracy rates. For instance, Smolander et al. achieved a 95.97% accuracy in distinguishing LUAD from control samples using deep learning models applied to coding RNA data [21]. Similarly, Fan et al. used support vector machines (SVMs) with a 12-gene signature to achieve a 91% accuracy in the same classification task [22].

Multiclass classification of lung cancer subtypes has also been addressed in the literature. Gonzales et al. developed a model for classifying small-cell lung cancer (SCLC), LUAD, LUSC, and large-cell lung carcinoma (LCLC) using differentially expressed genes (DEGs) as input, achieving an accuracy of 88.23% with the random forest (RF) algorithm [23]. Additionally, Castillo-Secilla et al. attained a 95.7% accuracy in subtype classification of non-small cell lung cancer (NSCLC) using random forest [24]. Studies utilizing miRNA-Seq data for lung cancer classification have also been conducted. Ye et al. identified a 10-miRNA signature for

TABLE 1. Summary of the related works.

Author	Journal	Modalities	Problem	Model	Metrics	Results
Carrillo-Perez, F.[2], 2022	Journal of Personalized Medicine	RNA-seq miRNA-seq CNV metDNA WSI	Classification of lung cancer subtypes: (LUAD vs LUSC) vs Control (benign samples)	CNN-SVM-Late fusion	Accuracy	95.53%
Smolander et al.[21],2019	BMC Cancer	RNA-seq	Classification of lung cancer subtype LUAD vs control	DNN	Accuracy	95.97%
Fan et al.[22],2018	J. Transl. Med	RNA-seq	Classification of lung cancer subtype LUAD vs control	SVM	Accuracy	91%
Castillo-Secilla et al.[24],2021	Comput. Biol. Med	RNA-seq	Classification of lung cancer subtypes : (LUAD vs. LUSC) vs. control	RF	Accuracy	95.7%
Ye et al.[18],2020	Gene	miRNA	Classification of lung cancer subtype LUSC vs. control	SVM	F1 Score	99.4%
Coudray et al.[19],2018	Nature medicine	WSI	Classification of lung cancer subtypes: (LUAD vs. LUSC) vs. control	CNN	Area Under Curve (AUC)	0.978
Kanavati et al.[25] 2020	Sci. Rep	WSI	Classification of lung cancer subtype Lung carcinoma vs. control	CNN	AUC	0.988
Graham et al.[26],2018	Medical Imaging 2018: Digital Pathology. International Society for Optics and Photonics	WSI	Classification of lung cancer subtypes: (LUAD vs. LUSC) vs. control	CNN	Accuracy	81%

distinguishing LUSC from control samples, achieving an F1 score of 99.4% [18].

Deep learning approaches, particularly convolutional neural networks (CNNs), have shown promise in combination with whole-slide imaging (WSI) for NSCLC subtype classification. Coudray et al. utilized CNNs with tiles extracted from WSI to classify LUAD, LUSC, and control samples, achieving an impressive AUC score of 0.978 [19]. Furthermore, Kanavati et al. used transfer learning with CNNs on manually labeled images to distinguish lung carcinoma from control samples, obtaining an AUC score of 0.988 [25].

At last, hybrid approaches combining deep learning with traditional statistics have been explored. Graham et al. utilized tiles extracted from images along with summary statistics to classify LUAD, control, and LUSC samples, achieving an accuracy of 81% [26].

III. METHODOLOGY

Through this section data preparation and pre-processing methodology will be discussed. Data is collected from the Genomic Data Commons (GDC) portal and different preprocessing techniques are used for preparing genomic and slide images for the training phase

A. DATA ACQUISITION AND PRE-PROCESSING

In this work two molecular modalities and one imaging modality are considered: RNA-Seq, miRNA-Seq and Whole Slide Images (WSIs). The data were collected from the Cancer Genome Atlas (TCGA) program [27] which is easily accessible from the Genomic Data Commons (GDC) portal. GDC is the National Cancer Institute’s (NCI) data sharing platform. The GDC contains NCI-generated data from some of the largest and most comprehensive cancer genomic datasets, including the Cancer Genome Atlas (TCGA) program. The data provided covers real-world scenarios which is important while dealing with such fatal disease.

WSI images were downloaded as a zip file and SVS files were extracted for the pre-processing phase. For the molecular data (RNA-Seq and miRNA-Seq), TCGAbiolinks Bioconductor R backage was used for downloading the data and extracting the cases with the corresponding expressed genes for pre-processing. STAR Counts from TCGA project are used for analysis.

For avoiding small test set or data imbalance no anomaly detection is used, but stratified kfold cross validation is used to obtain unbiased results. Stratified kfold cross validation ensures gaining the same proportion of data across the splits.

The molecular models follows a binary classification either the patient sample is Tumor or Normal while the WSI model follows 3 classes classification (LUAD, LUSC, Control). The data sample of each patient can be fed to these individual models for accurate diagnosis. First, the sample is tested on RNA-Seq (mRNA) model and miRNA-Seq model to be classified as normal or tumor. For Whole Slide Images, it goes for further classification as LUAD, LUSC or Normal for deciding the appropriate treatment plan.

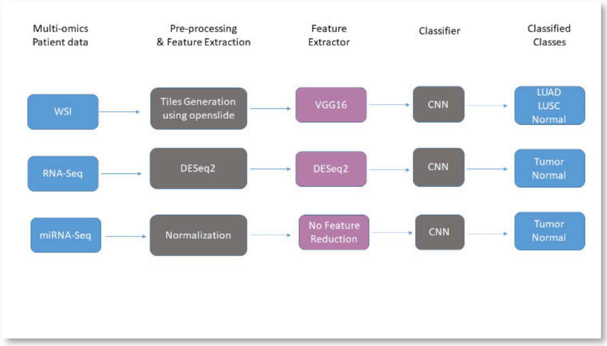

FIGURE 1. **The multimodal architecture for NSCLC classification.**

The multimodal architecture is shown as in figure1. For the whole slide images the tiles are first generated, then feature extraction is performed using the pre-trained model VGG16, after that the training phase is processed through a CNN model. Also for RNA-Seq DESeq2 is used for the preprocessing and for the feature reduction and then a CNN model is used for training. For miRNA-Seq, normalization is used for the preprocessing and no feature reduction is needed, and then a CNN model is used for the training phase.

1) WSI PRE-PROCESSING

As shown in figure2, tiles are first generated and separated in different folders for training, validation and testing. Then a pre-trained model VGG16 is used as a feature reduction. After that a CNN model is used for the training and classification.

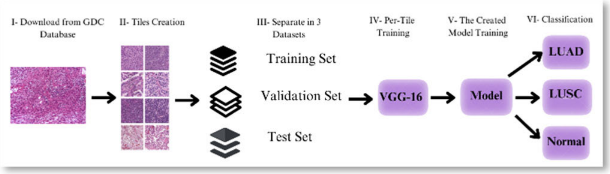

FIGURE 2. **Deep learning system architecture for the whole slide images.**

For the pre-processing stage of WSI, python package openslide was used. Images from directories for training and validation are loaded using OpenCV ('CV2'). Labels for each image is extracted from the directory path. Label encoding is performed using 'LabelEncoder' from scikit-learn to convert categorical labels into numerical values. The data is split into separate folders for training and validation to make sure that our model is never trained and tested on the same set of tiles. Normalization for the pixel values is performed between 1 and 0. This normalization helps the neural network converge faster and perform better during training. No data augmentation is performed.

WSI images are known for their large size that they cannot be fed directly to any neural networks. So, a magnification factor 20x was used for obtaining 512 × 512 non-overlapping pixel-tiles. Depending on the original WSI image, tens to thousands of tiles are generated as in the literature review [19].

TABLE 2. **Number of tiles generated per class.**

	# Tiles
LUAD	7225
LUSC	3380
Normal	1705
Total	12310

Table2 shows the number of tiles generated for each class that are used for the training. For implementing the classification models, the Python packages Tensorflow and Scikit-Learn were used.

2) OMICS DATA PRE-PROCESSING

As shown in Figure3, gene expression quantification STAR Counts data is used from the GDC portal for both mRNA and miRNA samples. first preprocessing and feature reduction for RNA-Seq is performed using DESeq2 and then a CNN model is used for training. For miRNA-Seq normalization is performed as the preprocessing and no feature reduction is needed and then a CNN model is used for training. Classification is binary meaning it is either Normal or Tumor.

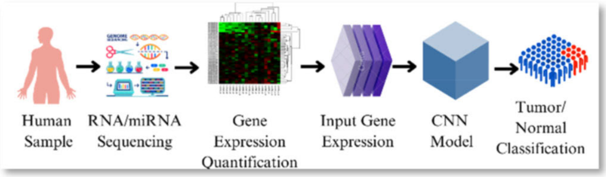

FIGURE 3. **Deep learning system architecture for the molecular samples.**

Choosing the right pre-processing tools and packages was important for accurate extraction of differentially expressed genes. Noise and missing Values are two common challenges faced when dealing with gene expression from next generation sequencing. Choosing the right subset of genes that affects the most in tumor growth is essential in the preprocessing phase. According to the fact that RNA-Seq does not follow Normal Distribution but follows Negative Binomial Distribution makes DESeq2 the best choice for RNA-Seq pre-processing. DESeq2 is an R package that is used for the analysis of high-throughput RNA-Seq data to extract the Differentially Expressed Genes (DEGs) which are usually responsible for causing cancer.

One essential step in RNA-Seq data analysis involves normalizing the data, with DESeq2 employing the ''size factor'' method to account for variants in sequencing depth among samples. This normalization ensures gene expression values are comparable across samples, facilitating accurate identification of DEGs. Alongside factor normalization, DESeq2 utilizes a variance-stabilizing transformation for further enhance data quality by stabilizing variance across expression levels. This combined approach minimizes bias and enhances the accuracy of differentially expressed genes.

Moreover, DESeq2 offers negative binomial distribution models to address the over-dispersion often observed in RNA-Seq data, acknowledging variability not adequately explained by a simple Poisson distribution. By incorporating the negative binomial distribution, DESeq2 effectively models gene expression count dispersion, yielding more reliable estimates of differential expression.

DESeq2 was used over 60660 genes to extract the most important differentially expressed genes. As parameters, a Log_2 Fold Chain (LFC) value of 1.2 and a p-value of 0.05were set. 33 DEGs were extracted for the training and classification phase.

For the case of miRNA, there was no need for feature reduction because TCGA provides information for only 1881 miRNAs. Only normalization was applied using MinMaxScaler from scikit-learn library in python. For both of the genetic data the categorical labels were encoded using 'LabelEncoder' to convert them into numeric format.

Figure4 shows the gene expression data reshaping process for feeding into a CNN model. The gene expression can be presented in 2D format where rows represent genes and columns represent expression profiles of patient tumor samples. The value in cell X_{ij} represents the gene (i) in patient sample (j). Then, in order to feed the neural networks with this data, it has to be reshaped in a format suitable for the network to be understood, analyzed and classified. A common 3D shape for convolutional neural network (*number of samples, sequence length, number of channels*) is used for the molecular data. The number of samples represents the data points or samples in the dataset. Each sample represents an individual instance or observation in the dataset. The sequence length represents the number of genes measured (length of gene sequence) in each sample. The number of channels represents the number of features or gene expression levels measured for each gene. So, reshaping input features array to (*number of samples, sequence length, number of channels*) means organizing your data such that each sample consists of a sequence of gene expression values (sequence length) across different genes (number of channels). This shape allows the CNN to effectively learn spatial patterns in the gene expression data across different genes while considering the sequential nature of the data.

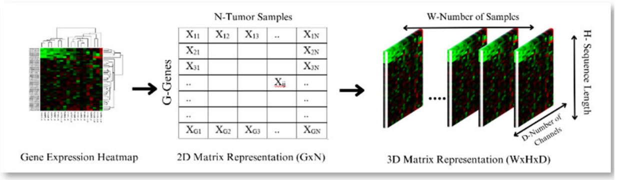

FIGURE 4. Gene expression 3D reshaping for feeding into CNNs.

Here in Table3 the number of samples for each data modality are indicated.

TABLE 3. Number of samples per cass for each data modality.

	RNA-Seq(mRNA)	miRNA-Seq
Normal	110	91
Tumor	1041	762
Total	1151	853

IV. IMPLEMENTATION AND DISCUSSION

Through this section a detailed explanation for the training phase is given. Also, figures of the obtained results are shown.

A. MODEL SELECTION AND TRAINING

For the training phase different methodologies are used for the image and genetic data. For the whole slide images, pre-trained model is first used as a transfer learning for feature reduction and then a CNN model is used for training and classification. RNA-Seq and miRNA samples are trained and classified through a CNN model.

1) WSI TRAINING

The VGG16 architecture was used for the WSIs. Different pre-trained models were tried for training such as Incepri-onV3 and Resnet-50, but VGG16 gives better results. VGG16 network is a specific implementation within the VGG family of convolutional neural networks, designed to enhance the depth of the architecture while maintaining simplicity in the design by using small 3 × 3 convolutional filters. These small filters ensures that the network captures intricate patterns while keeping the number of parameters manageable. Moreover, the increased depth of the VGG16 network, with 16 weight layers in total, allows it to learn complex features and achieve high performance in image classification tasks. In addition to that, the consistent use of the same filter size and doubling the number of filters after each max-pooling layer provides a systematic approach to increasing the network capacity and depth.

The pre-trained weights on ImageNet are used as the starting point. Transfer learning is also applied through freezing the layers of VGG16 and set them to non-trainable for preventing weights from being applied during training. All 18 convolutional and pooling layers in the VGG16 model

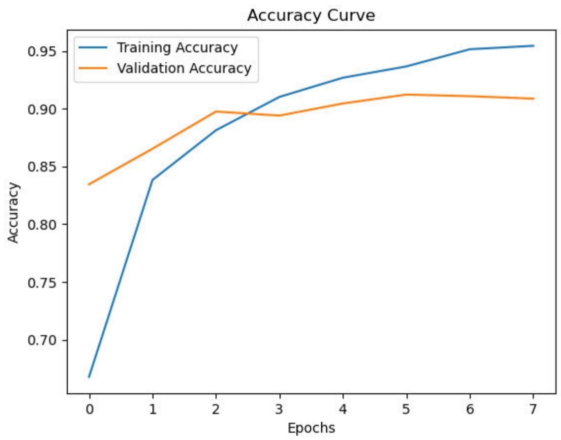

FIGURE 5. WSI training and validation accuracy.

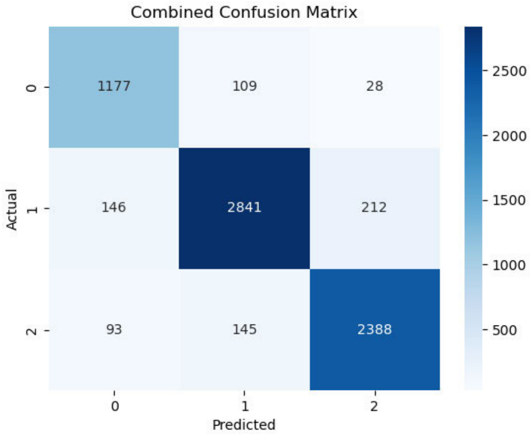

FIGURE 6. WSI confusion matrix.

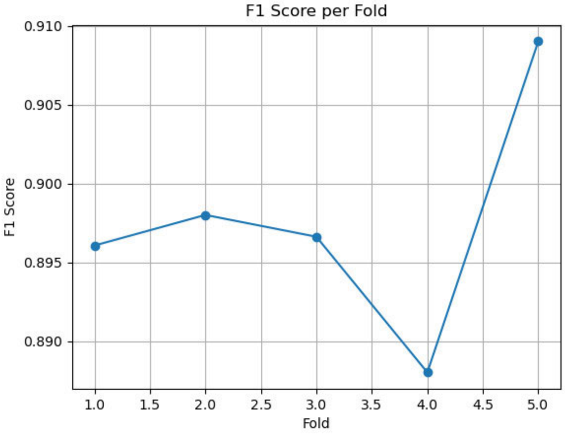

FIGURE 7. WSI F1 score.

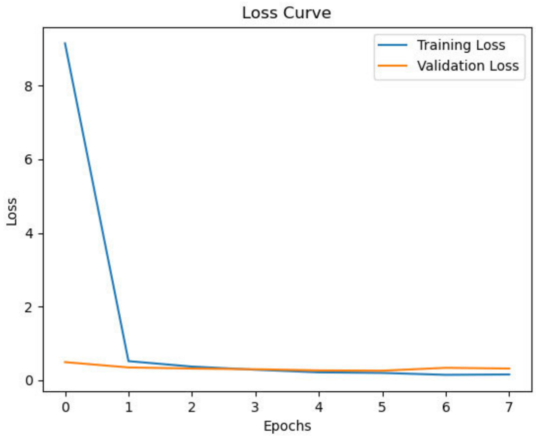

FIGURE 8. WSI training and validation loss curve.

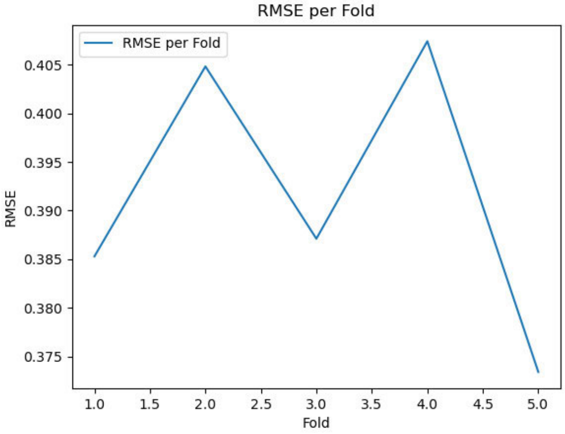

FIGURE 9. WSI RMSE for each fold.

were frozen during training. The training images are passed through the VGG16 model to extract features and convert them to feature vectors.

Stratified k-fold cross validation is used to split the training data into 5 folds while ensuring that each fold has approximately the same proportion of samples for each class to avoid small test set and imbalance to obtain unbiased results.

A custom neural networks model is defined and compiled for classification. It consists of a Flatten layer to flatten the feature maps, followed by dense layers with ReLU activation, 0.4 dropout for regularization, and a softmax output layer. Adam optimizer was used with its default parameters. Categorical cross entropy as a loss function and accuracy as evaluation metric.

The network was trained during 8 epochs for each fold. The following metrics were calculated during each fold (accuracy, F1 score, AUC, RMSE, and confusion matrix) and the average of each matrix was calculated at the end of the training and validation phase.

2) MOLECUAR SOURCES TRAINING

The implementation of molecular data training was performed using Keras and scikit-learn and tensorflow. The input features (gene expression data) and labels are loaded and the dataset are split into 80% training and 20% testing sets.

A CNN model was defined using keras 'sequential' API which consists of:

- One dimensional convolutional layer with 64 filters, kernel size of 3, and ReLU activation function.
- Max pooling layer with pool size of 2.

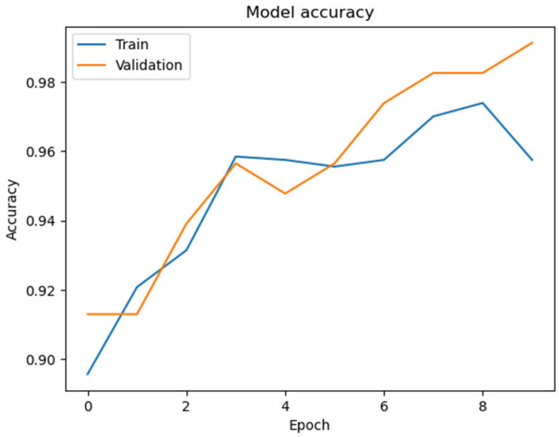

FIGURE 10. mRNA model accuracy.

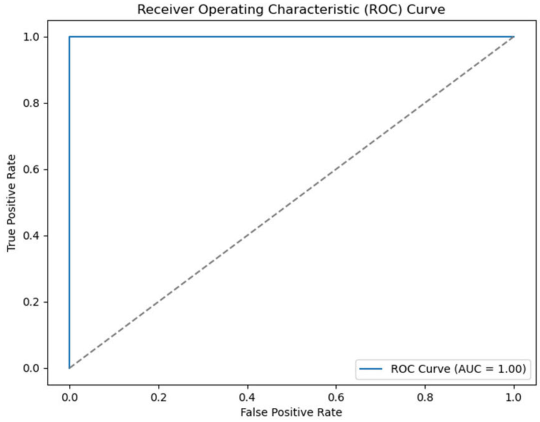

FIGURE 11. mRNA model ROC curve.

- Flattening layer to flatten the output of the output of the convolutional layer.
- Fully connected dense layer with 128 neurons and RelU activation function.
- Output layer ('Dense') with 2 neurons and softmax activation function for binary classification.

Then the model was compiled with default parameters of Adam optimizer, sparse-categorical cross-entropy loss function and accuracy metric.

For RNA-Seq (mRNA) process runs for stratified 10 fold cross validation and 10 epochs with batch size of 32, while for miRNA-Seq the process runs for 5 folds and 10 epochs with batch size of 32.

V. RESULTS

The results obtained by each modality can be observed in Table4. As observed, the best results are achieved according to the following order: miRNA-Seq, RNA-Seq, WSI. Binary classification is performed on both of the molecular data. Classification is based on being tumor or normal. Based on the results of the binary classification, the samples go for one more classification step on WSI images for more precise results. The WSI assures the result if it is normal and goes for further classification. Otherwise, if classification is tumor the samples are to be classified either for LUAD or LUSC. The followed method helps in determining the right treatment plan.

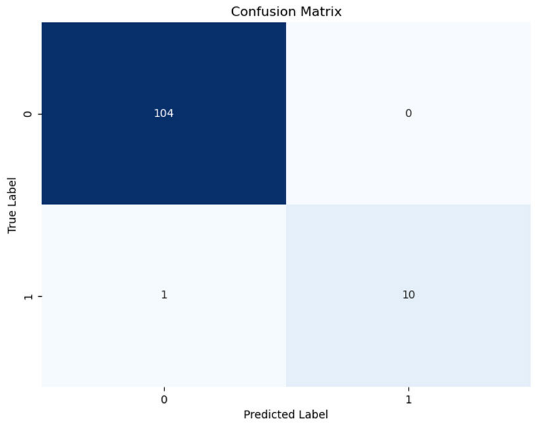

FIGURE 12. mRNA model confusion matrix.

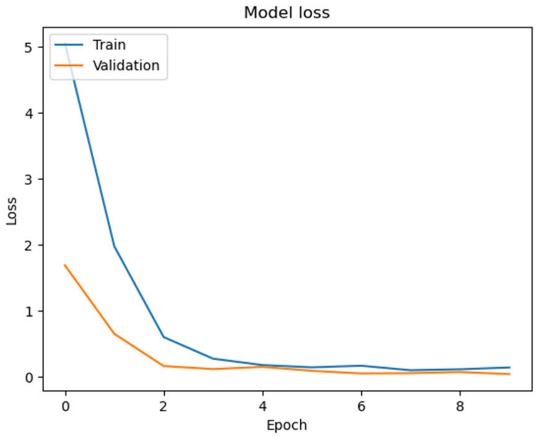

FIGURE 13. mRNA model loss for the training and validation.

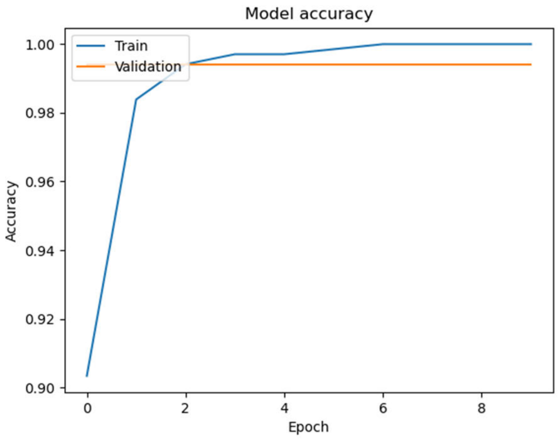

FIGURE 14. miRNA model accuracy for the training and validation.

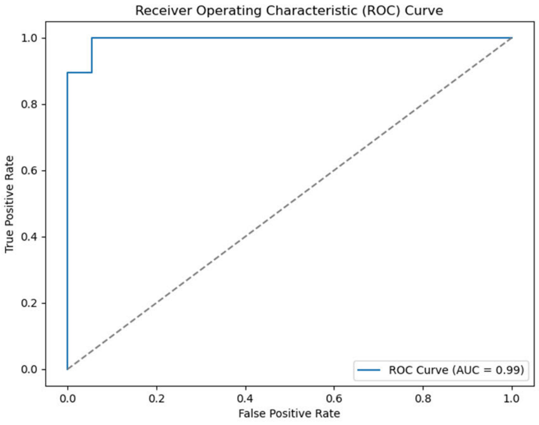

FIGURE 15. miRNA model ROC curve.

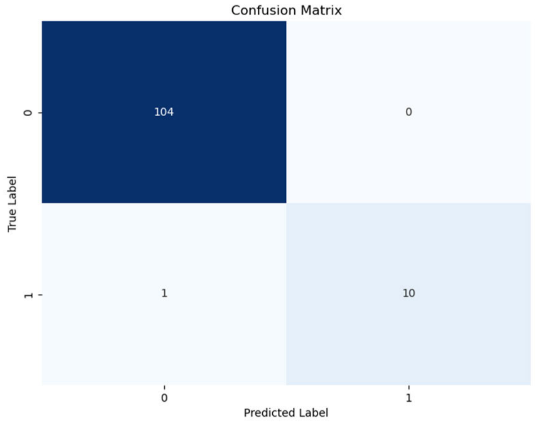

FIGURE 16. miRNA model confusion matrix.

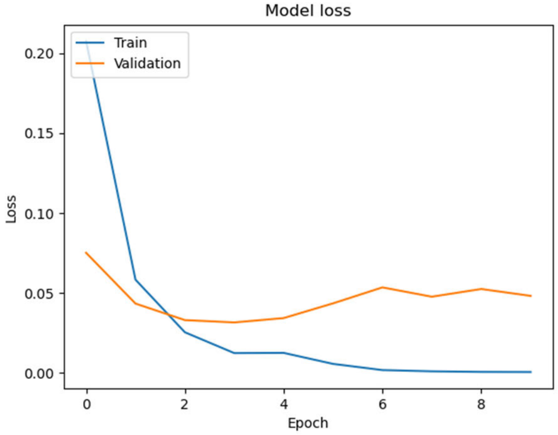

FIGURE 17. miRNA model loss for the training and validation.

VI. DISCUSSION

The results observed in table exceeds those gained by other previous studies as previously mentioned in Table 1. The mRNA-Seq classification accuracy is 96.79% which exceeds the results obtained by Smolander et al. [21], Fan et al. [22], and Castillo-Secilla et al. [24]. Also, for the miRNA-seq classification, F1score of 99.67% is improved comparing to previous related works obtained by Ye et al. [18]. Moreover, WSI classification accuracy of 89.73% exceeds the ones obtained by Graham et al. [26], and Carrillo-Perez et al. [2]. All efforts exerted in the research field are for improving the results obtained which reflects in faster discovery and treatment of the disease. So, through better results gained in this research, the hope for better treatment could be possible. Also, applying the methodology performed on other types of cancer could be beneficial.

TABLE 4. Results obtained by each data modality.

Modality	ACC	F1 Score	AUC
mRNA	96.79%	95.238%	100%
miRNA	98.59%	99.67%	99.41%
WSI	89.73%	89.76%	97,54%

VII. CONCLUSION

To summarize the previous work, deep learning using convolutional neural networks has already proven its great impact on the bioinformatics field and will lead the next revolution of detecting rare diseases at very early stages. Convolutional neural networks will confront the growing problem of high dimensionality especially in the omics data. Moreover, it is already known for its accurate classification of high resolution slide images. Throughout this research we demonstrated the great benefit of using convolutional neural networks in the classification of non-small cell lung cancer using omics data such as mRNA-Seq and miRNA-Seq and images like Whole Slide Images achieving higher result than obtained by previous works.

VIII. FUTURE WORK

Different omics modalities could be used for new insights and new biomarkers to be discovered such as DNA methylation (metDNA), copy number vartaion (CNV). Moreover, different pre-trained models could be used to test the best models for higher results such as InceptionV3, Resnet18, Resnet50. Also, difficulty types of cancer data are available in the GDC portal. This methodology could be applied on other 32 types of cancer as a unified tool for classification. In addition to that, late fusion is proved to be a great tool for more insights and biological meanings. It may be applied for the classification of these different types for more accurate treatment.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] F. Carrillo-Perez, J. C. Morales, D. Castillo-Secilla, O. Gevaert, I. Rojas, and L. J. Herrera, "Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis," *J. Personalized Med.*, vol. 12, no. 4, p. 601, Apr. 2022, doi: 10.3390/jpm12040601.

[3] T. Khorshed, M. N. Moustafa, and A. Rafea, "Deep learning for multi-tissue cancer classification of gene expressions (GeneXNet)," *IEEE Access*, vol. 8, pp. 90615–90629, 2020, doi: 10.1109/ACCESS.2020.2992907.

[4] *Types of Lung Cancer | Cancer Research U.K.* Accessed: May 23, 2024. [Online]. Available: https://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/types

[5] A. M. Bode and Z. Dong, "Cancer prevention research—Then and now," *Nature Rev. Cancer*, vol. 9, no. 7, pp. 508–516, Jun. 2009, doi: 10.1038/nrc2646.

[6] N. Hanna, D. Johnson, S. Temin, S. Baker, J. Brahmer, P. M. Ellis, G. Giaccone, P. J. Hesketh, I. Jaiyesimi, N. B. Leighl, G. J. Riely, J. H. Schiller, B. J. Schneider, T. J. Smith, J. Tashbar, W. A. Biermann, and G. Masters, "Systemic therapy for stage IV non–small-cell lung cancer: American society of clinical oncology clinical practice guideline update," *J. Clin. Oncol.*, vol. 35, no. 30, pp. 3484–3515, Oct. 2017, doi: 10.1200/jco.2017.74.6065.

[7] Z. Wang, M. Gerstein, and M. Snyder, "RNA-seq: A revolutionary tool for transcriptomics," *Nature Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009, doi: 10.1038/nrg2484.

[8] J. M. Rizzo and M. J. Buck, "Key principles and clinical applications of 'next-generation' DNA sequencing," *Cancer Prevention Res.*, vol. 5, no. 7, pp. 887–900, Jul. 2012, doi: 10.1158/1940-6207.capr-11-0432.

[9] J. M. Knight, I. Ivanov, K. Triff, R. S. Chapkin, and E. R. Dougherty, "Detecting multivariate gene interactions in RNA-seq data using optimal Bayesian classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 2, pp. 484–493, Mar. 2018, doi: 10.1109/TCBB.2015.2485223.

[10] S. Sleijfer, J. Bogaerts, and L. L. Siu, "Designing transformative clinical trials in the cancer genome era," *J. Clin. Oncol.*, vol. 31, no. 15, pp. 1834–1841, May 2013, doi: 10.1200/jco.2012.45.3639.

[11] L. E. MacConaill, "Existing and emerging technologies for tumor genomic profiling," *J. Clin. Oncol.*, vol. 31, no. 15, pp. 1815–1824, May 2013, doi: 10.1200/jco.2012.46.5948.

[12] *The Cancer Genome Atlas Program (TCGA)—NCI.* Accessed: May 23, 2024. [Online]. Available: https://www.cancer.gov/ccg/research/genome-sequencing/tcga

[13] K. R. Kukurba and S. B. Montgomery, "RNA sequencing and analysis," *Cold Spring Harbor Protocols*, vol. 2015, no. 11, Nov. 2015, Art. no. pdb-top084970, doi: 10.1101/pdb.top084970.

[14] J. E. Dancey, P. L. Bedard, N. Onetto, and T. J. Hudson, "The genetic basis for cancer treatment decisions," *Cell*, vol. 148, no. 3, pp. 409–420, Feb. 2012, doi: 10.1016/j.cell.2012.01.014.

[15] C. Peng, X. Wu, W. Yuan, X. Zhang, Y. Zhang, and Y. Li, "MGRFE: Multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 621–632, Mar. 2021, doi: 10.1109/TCBB.2019.2921961.

[16] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[17] D. Castillo, J. M. Galvez, L. J. Herrera, F. Rojas, O. Valenzuela, O. Caba, J. Prados, and I. Rojas, "Leukemia multiclass assessment and classification from microarray and RNA-seq technologies integration at gene expression level," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0212127, doi: 10.1371/journal.pone.0212127.

[18] Z. Ye, B. Sun, and Z. Xiao, "Machine learning identifies 10 feature miRNAs for lung squamous cell carcinoma," *Gene*, vol. 749, Jul. 2020, Art. no. 144669, doi: 10.1016/j.gene.2020.144669.

[19] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Med.*, vol. 24, no. 10, pp. 1559–1567, Oct. 2018, doi: 10.1038/s41591-018-0177-5.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[21] J. Smolander, A. Stupnikov, G. Glazko, M. Dehmer, and F. Emmert-Streib, "Comparing biological information contained in mRNA and non-coding RNAs for classification of lung cancer patients," *BMC Cancer*, vol. 19, no. 1, Dec. 2019, Art. no. 1176, doi: 10.1186/s12885-019-6338-1.

[22] Z. Fan, W. Xue, L. Li, C. Zhang, J. Lu, Y. Zhai, Z. Suo, and J. Zhao, "Identification of an early diagnostic biomarker of lung adenocarcinoma based on co-expression similarity and construction of a diagnostic model," *J. Transl. Med.*, vol. 16, no. 1, Jul. 2018, Art. no. 205, doi: 10.1186/s12967-018-1577-5.

[23] S. González, D. Castillo, J. M. Galvez, I. Rojas, and L. J. Herrera, "Feature selection and assessment of lung cancer sub-types by applying predictive models," in *Proc. Int. Work-Conf. Artif. Neural Netw.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11507, 2019, pp. 883–894, doi: 10.1007/978-3-030-20518-8_73.

[24] D. Castillo-Secilla, J. M. Gálvez, F. Carrillo-Perez, M. Verona-Almeida, D. Redondo-Sánchez, F. M. Ortuno, L. J. Herrera, and I. Rojas, "KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104387, doi: 10.1016/j.compbiomed.2021.104387.

[25] F. Kanavati, G. Toyokawa, S. Momosaki, M. Rambeau, Y. Kozuma, F. Shoji, K. Yamazaki, S. Takeo, O. Iizuka, and M. Tsuneki, "Weakly-supervised learning for lung carcinoma classification using deep learning," *Sci. Rep.*, vol. 10, no. 1, Jun. 2020, Art. no. 9297, doi: 10.1038/s41598-020-66333-x.

[26] S. Graham, M. Shaban, T. Qaiser, N. A. Koohbanani, S. A. Khurram, and N. Rajpoot, "Classification of lung cancer histology images using patch-level summary statistics," *Proc. SPIE*, vol. 10581, pp. 327–334, Mar. 2018, doi: 10.1117/12.2293855.

[27] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuar, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013, doi: 10.1038/ng.2764.

MARIAN MAGDY AMIN received the B.Sc. degree in computer science from The American University in Cairo, in 2014. She is a currently pursuing the master's degree (by research) with Fayoum University, Egypt.

AHMED S. ISMAIL received the master's degree in computer science and information systems from the Department of Information Systems, Faculty of Computers and Information, Fayoum, Egypt, in 2012, and the Ph.D. degree from the Department of Information Systems, Faculty of Computer Science and Information Systems, Fayoum University, in January 2021. He is an Assistant Professor. He has authored/co-authored several scientific researches in various technical fields, such as semantic web, data science, big data, the Internet of Things, and blockchain.

MASOUD E. SHAHEEN received the B.Sc. degree in science from the Department of Mathematics and Computer Science, Minia University, in 1996, the M.S. degree in computer science from the Faculty of Science, Fayoum University, Egypt, in 2005, and the Ph.D. degree in computer science from The University of Southern Mississippi, Hattiesburg, MS, USA, in 2013. He was the Vice-Dean of postgraduate studies and research with the Faculty of Computers and AI, Fayoum University, from May 2021 to December 2022, where he is an Associate Professor with the Computer Science Department. He is the Project Portal Manager with Fayoum University. He is also the Vice-Dean for Community Service and Environmental Development Affairs.