IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 March 2024, accepted 23 April 2024, date of publication 26 April 2024, date of current version 3 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3394030

RESEARCH ARTICLE

Enhancing Lung Cancer Classification and Prediction With Deep Learning and Multi-Omics Data

TEHNAN I. A. MOHAMED[1,2] AND ABSALOM EL-SHAMIR EZUGWU[3]

[1]Department of Computer Science, Faculty of Mathematical and Computer Sciences, University of Gezira, Wad Madani 11123, Sudan
[2]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg Campus, Pietermaritzburg, KwaZulu-Natal 3201, South Africa
[3]Unit for Data Science and Computing, North-West University, Potchefstroom 2520, South Africa

Corresponding author: Absalom El-Shamir Ezugwu (absalom.ezugwu@nwu.ac.za)

ABSTRACT Lung adenocarcinoma (LUAD), a prevalent histological type of lung cancer and a subtype of non-small cell lung cancer (NSCLC) accounts for 45–55% of all lung cancer cases. Various factors, including environmental influences and genetics, have been identified as contributors to the initiation and progression of LUAD. Recent large-scale analyses have probed into RNASeq, miRNA, and DNA methylation alterations in LUAD. In this study, we devised an innovative deep-learning model for lung cancer detection by integrating markers from mRNA, miRNA, and DNA methylation. The initial phase involved meticulous data preparation, encompassing multiple steps, followed by a differential analysis aimed at identifying genes exhibiting differential expression across different lung cancer stages (Stages I, II, III, and IV). The DESeq2 technique was employed for RNASeq data, while the LIMMA package was utilized for miRNA and DNA methylation datasets during the differential analysis. Subsequently, integration of all prepared omics data types was achieved by selecting common samples, resulting in a consolidated dataset comprising 448 samples and 8228 features (genes). To streamline features, principal components analysis (PCA) was implemented, and the synthetic minority over-sampling technique (SMOTE) algorithm was applied to ensure class balance. The integrated and processed data were then input into the PCA-SMOTE-CNN model for the classification process. The deep learning model, specifically designed for classifying and predicting lung cancer using an integrated omics dataset, was evaluated using various metrics, including precision, recall, F1-score, and accuracy. Experimental results emphasized the superior predictive performance of the proposed model, attaining an accuracy, precision, recall, and F1-score of 0.97 each, surpassing recent competitive methods.

INDEX TERMS Gene expression, lung cancer, mRNA, miRNA, DNA methylation, differential analysis, omics data.

I. INTRODUCTION

Non-communicable diseases (NCDs) have characteristics that make them ailments that are linked to one's lifestyle. They are highlighted as the primary factors contributing to cardiometabolic conditions like metabolic disorders such as cardiovascular diseases, obesity, and type 2 diabetes. Furthermore, behaviors like smoking and alcohol consumption are also associated with NCDs, including several types of cancer. This emphasis on lifestyle factors implies that these conditions and illnesses can be prevented, and their complications and associated health issues can be mitigated through individual behavioral modifications, including adopting a healthier diet, engaging in physical activity, and managing one's weight [1]. Cancer, which is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells is increasing due to factors like unwell lifestyles, the aging population, and commercial interests. At present approximately

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

one out of every five men and one out of every six women are diagnosed with cancer. Unfortunately, about one in eight men and one in ten women will lose their lives to this disease. Projections suggest that by 2030 around 13 million people will die from cancer annually with most of these deaths occurring in middle-income countries [2]. This makes cancer a leading cause of death which has the potential to hurt a nation's productivity. The WHO aims to reduce NCDs related deaths by 2030 requires significant advancements, in cancer treatment and control [3], [4], [5]. While lifestyle and social aspects play a great part in the growth of NCDs, it is important to note that each of these conditions also has a significant genetic component [6].

Lung cancer is ranked as the most common type of cancer leading to death and the second most commonly diagnosed cancer in both women and men [7], [8]. The two main subtypes of lung cancer are small cell carcinoma (SCLC), which is fast-growing but not common, and non-small cell carcinoma (NSCLC), the most diagnosed type of cancer, which grows slowly. Symptoms of lung cancer include weight loss, chest pain, shortness of breath, and persistent cough. The diagnostic methods for lung cancer include physical examination, different imaging techniques, and molecular testing to determine exact genetic mutations or biomarkers to help in selecting the best therapy options [9], [10], [11]. Advances in genome profiling methods in recent decades have significantly improved the comprehension of cancer development at the molecular level and helped in cancer treatments including lung cancer by identification of biomarkers [12].

Gene expression profiling has provided valuable information about gene activities and describes the current physiology of the cell. It has been used effectively to help in the early diagnosis and prognosis of cancer types [13], [14]. Single-cell RNA sequencing (scRNA-Seq), which is a powerful technique used in molecular biology to analyze gene expression at the single-cell level is a helpful tool for characterizing gene expression [15]. Gene expression is an approach that converts the genetic information contained in DNA instructions to produce proteins and different molecules. DNA transcription is a fundamental process in molecular biology where information from a DNA sequence is transcribed into RNA. This process is a key step in gene expression, during which the genetic instructions stored in DNA are used to synthesize ribonucleic acid RNA molecules [14]. RNAs can be categorized into coding and non-coding RNAs, with the latter containing small nuclear RNAs (snRNAs), small interfering RNAs (siRNAs), small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs), PIWI-interacting RNAs (piRNAs), ribosomal RNAs (rRNAs), long non-coding RNAs (lncRNAs), microRNAs (miRNAs), and circular RNAs (circRNAs) [16], [17].

Different RNA modifications have been discovered and the most recognized among them is methylation modifications. miRNAs are non-coding single-stranded small RNA. They are widely located in eukaryotes and play a significant role in post-transcriptional regulation of gene expression, which is achieved through translation inhibition and mRNA cleavage. The miRNA is an essential biomarker that helps in the diagnosis and treatment of diseases and the development of anti-tumor medicines [18], [19]. In addition, DNA methylation constitutes a fundamental chemical process concerning the binding and interchange of DNA with a methyl group. This change modifies the functionality of the DNA, recreating a pivotal part in X-chromosome inactivation, genomic imprinting, repression of repetitive elements, and the aging process. Additionally, DNA methylation is associated with many kinds of cancer [20].

A significant issue in gene expression data is that they include a high number of gene counts often referred to as the curse of dimensionality against a few observations that are referred to as data sparsity. However, high-dimensional data contains irrelevant or redundant features, class imbalance, and a high amount of noise in the genes that lead to inaccurate diagnoses of cancer [13]. For that reason, it is necessary to use techniques that decrease the dimensionality of the data while keeping the underlying gene relationships, especially when dealing with extensive gene expression datasets. Dimensionality reduction algorithms enable a summary of the data's variability to a smaller set of random variables and help to visualize datasets with tens to thousands of dimensions in 2D or 3D formats. the most commonly utilized technique is principal component analysis (PCA), it employs linear combinations of variables to develop orthogonal axes that effectively capture the data's variation utilizing a decreased number of variables [21].

Class imbalance is a common issue where the distribution of examples across different classes is not equal. In other words, some classes have significantly fewer instances or samples than others. The majority class is the class that has the highest number of instances while the minority classes are the classes with fewer data. The imbalance ratio (IR) is defined as the difference between the number of samples in the majority class and each of the minority classes. To handle this issue, it is required to rebalance the classes by modifying the data itself. This can be achieved by removing some samples from the majority class (under-sampling) or by expanding the number of minority class samples (over-sampling). In Under-sampling, the less important patterns from the majority class are removed, which may lead to the loss of essential information. Over-sampling entails copying or generating new minority class patterns to compensate for the lack of data, which can lead to overfitting [22], [23], [24]. The SMOTE (Synthetic Minority Over-sampling Technique) method is an effective resampling technique for imbalanced data classification by oversampling samples from the minority class to rebalance the gene expression dataset [25].

Analyzing the gene expression data is still challenging due to its many characteristics such as high dimensionality, complexity, and heterogeneity. Deep Learning (DL) algorithms have been applied and proven to be an effective technique

for handling gene expression data, leading to significant improvement in the predictions and diagnosis of various types of cancer [13], [25], [26], [27]. Deep Learning algorithms extract the features from the original input data. Different deep learning approaches, such as Convolutional Neural Networks (CNN), Feed-Forward Neural Networks (FFN), Recurrent Neural Networks (RNN), and Autoencoders (AE) are used to analyze the gene expression datasets. Especially, CNN is employed to learn multiple layers of kernel filters and classifier weights [26], [27]. CNN models have demonstrated outstanding classification performance in gene expression analysis because of their ability to capture local spatial relations from the input data. Therefore, CNN models have consistently classified among the top-performing deep learning models when applying gene expression data [14], [28].

In this study, we focused on LUAD, which is the primary type of lung cancer diagnosis, and introduced a novel deep-learning model for integrating mRNA, miRNA, and DNA methylation. In contrast to previous studies, our approach included a detailed differential analysis across various lung cancer stages and the meticulous integration of omics data through sample selection. This paper makes the following essential contributions:

- Preparing and combining datasets: we combined different kinds of omics data including mRNA, DNA methylation, and miRNA for lung cancer, this allows for a deeper exploration into the intricacy of epigenetic regulation and interaction.
- Learning from diverse data sources: we created progressive learning techniques that seamlessly integrate diverse datasets, mitigating potential biases related to class imbalances.
- Proposing an improved deep learning model: we have introduced an enhanced deep learning model for predicting lung cancer stages (I, II, III, and IV) using meticulously curated multi-omics data. Nevertheless, constructing the prediction model with integrated diverse datasets posed significant challenges due to the high dimensionality of the data and the presence of imbalanced class characteristics. As a result, we employed techniques to mitigate dimensionality and address class imbalance. The synergistic application of these methodologies is intended to boost the predictive performance of the model.

II. RELATED WORK

Many classification models based on gene expression data have been developed. Ismail et al. [24] proposed a hybrid stacking ensemble model with a synthetic minority oversampling technique (Stack-SMOTE) to predict the genes related to autism spectrum disorder (ASD). They proposed an ensemble learning method using a gradient boosting technique based on random forest (GBBRF). The results of the proposed hybrid Stacking-SMOTE model achieved an accuracy of 95.5%. Another study by Mulla et al. [29] introduced a method that combines SMOTE oversampling with PCA. Three classification algorithms were used namely K-Nearest Neighbor (KNN), Logistic Regression (LR), and Decision Tree (DT) based on two datasets. The experimental results of the PCA+SMOTE method showed an improvement in the classification results. the KNN model demonstrated high performance. A study by Sakib et al. [30], used various classification methods including DT, Support Vector Machine (SVM), LR, KNN, Random Forest (RF), and Naïve Bayes (NB) to detect blood cancer using gene expression data. PCA and SMOTE were applied in the data preprocessing. The result indicated that NB, LR, and SVM outperformed other methods with an accuracy of 100%.

Almarzouki [44] proposes a deep-learning approach for cancer classification using gene expression profiling data. It introduces feature selection techniques like mutual information difference and mutual information quotient used to reduce the dimensionality of gene expression data and select important genes. Various classification algorithms like convolutional neural networks are then trained on the selected features to classify cancers like lung, kidney, and brain tumors. The CNN model achieves up to 96.43% accuracy on test data using k-fold cross-validation. Overall, the study presents an effective framework involving feature selection and deep learning for cancer profiling and classification from high-dimensional gene expression data, which can help improve cancer diagnosis and outcomes.

Xu et al. [45] propose a deep flexible neural forest (DFNForest) model for cancer subtype classification based on gene expression data. It combines the fisher ratio and neighborhood rough set for dimensionality reduction of genes to select the most informative genes. Fisher ratio is first used to eliminate invalid genes, then a neighborhood rough set is applied to reduce redundant genes. DFNForest is then proposed as an ensemble of flexible neural trees to solve multi-classification problems. Each forest contains binary classification problems. The model depth is increased through a cascade structure without additional parameters. Experiments on three cancer datasets show the gene selection method achieves high accuracy with fewer genes compared to other methods. DFNForest also outperforms other classification methods on the gene expression data, demonstrating its effectiveness for cancer subtype classification.

Dwivedi [46] presents a framework for classifying cancers using machine learning techniques on microarray gene expression data. Specifically, it evaluates six different machine learning algorithms (artificial neural network, support vector machine, logistic regression, k-nearest neighbor, classification tree, naive Bayes) for their ability to classify acute lymphoblastic leukemia and acute myeloid leukemia samples based on expression levels of 7,129 genes. The results show that an artificial neural network approach achieved the highest classification accuracy of 98%, outperforming the other methods. Validation on independent test samples also achieved 100% accurate classification with

some methods. Therefore, the study demonstrates the potential of machine learning to effectively classify cancers using gene expression profiling data.

Tarek et al. [47] propose an ensemble-based gene expression classification system for cancer diagnosis and analysis. It introduces an ensemble of 5 classifiers using different feature selection methods and a 3-NN algorithm. Three datasets - Colon, Leukemia, and Breast cancer - are used to evaluate the approach. Their results show that the proposed ensemble system improves over individual classifiers and a related baseline approach, achieving higher accuracy and lower error rates on all three datasets. In particular, the ensemble is able to perfectly classify the Breast cancer dataset, representing an improvement over individual techniques. Overall, the study demonstrates the effectiveness of the ensemble-based gene expression classification approach for cancer applications.

In previous studies, different datasets such as mRNA data, miRNA data, and DNA methylation were used. In some research, different types of datasets were combined to detect various diseases. Yang et al. [31] identified and validated key genes during the progression and development of Lung adenocarcinoma (LUAD). They applied various kinds of analyses including survival analysis, enrichment analysis, and protein-protein interaction (PPI) networks. The results identified nine genes that play essential roles in the development of LUAD. Another study by Park et al. [20] proposed a deep learning-based model to predict Alzheimer's disease (AD) based on the combination of gene expression and DNA methylation data. The results of the proposed model showed an accuracy of 0.82%. Kutlay and Son [32] used multiple machine-learning models including RF, SVM, artificial neural networks (ANN), NB, and AdaBoost to determine the metastasis by integration of DNA methylation, miRNA, and mRNA. The proposed method achieved an F1 score of 92%.

A study by Su et al. [33] considered differentially expressed miRNAs and methylated using molecular and cellular function analysis. Their results showed that the interaction between miRNA and DNA methylation plays essential functions in lung cancer. Tomeva et al. [34] examined mutations and methylation in cell-free DNA (cfDNA), as well as miRNAs, in plasma samples collected from a total of 97 cancer patients. From the results, an accuracy of 95.4% was achieved by the proposed model. Another study by Shujaat et al. [35] presented a convolution neural network model called iProm-Sigma54 based on a grid search algorithm to produce a CNN-based predictor. Their results demonstrated that the proposed model outperformed previous methods. Varghese et al. [36] assessed the epigenetic mechanisms of ethical heterogeneity in hepatocellular carcinoma (HCC) via integration of miRNA, DNA methylation, and gene expression by applied mix ANOVA and Pathway analysis. The experimental results showed that, important differentially expressed genes in HCC were identified through the integrative analysis.

A study by Guan et al. [37] investigated the combined competitive endogenous RNA (ceRNA) and DNA methylation in esophageal carcinoma through enrichment analysis using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and gene ontology (GO). The results revealed that four key long non-coding RNAs (lncRNAs) were identified as associated with esophageal carcinoma. Rong et al. [38] proposed a CC2DT model combining convolutional autoencoders and convolutional neural network methods for classification and early lung cancer diagnosis. They integrated three lung cancer gene expression datasets, including miRNAseq, mRNA expression, and DNA methylation. Their experimental results show that an accuracy of 0.824 was obtained by the CC2DT model. Albaradei et al. [39] introduced a deep learning model (MetaCancer) to distinguish pan-cancer metastasis status. They used data from 400 patients including microRNA, DNA methylation, and RNA. The proposed model obtained an accuracy of 0.888. Wang et al. [40] developed a model for lung cancer subtype diagnosis utilizing weakly paired multi-omics data (LungDWM). According to the results obtained, the LungDWM model achieved an accuracy of 0.942.

III. MATERIAL AND METHODS

In this section, we introduce the proposed deep learning model, providing a description of the employed dataset and how the integration process for three various datasets was accomplished.

A. DATASET AND PRE-PROCESSING

We have downloaded three omics data types, namely, RNASeq, miRNA, and DNA methylation data from the cancer genome atlas (TCGA) (https://portal.gdc.cancer.gov/) using the TCGABiolinks package in R. Different techniques accompanied the data preparation. The RNASeq dataset had 60660 genes and 531 samples; the genes located on sex chromosomes (X and Y) were removed, reducing the total number of genes to 57670. Furthermore, filtering was performed using the rowSums function to exclude genes with low expression or variation. Genes were retained only if the sum of counts across all samples was greater than two and if at least four samples had counts greater than two. This process resulted in 46615 genes that have high variability between the samples. After that, the DESeq2 technique was applied to find the differentially expressed genes (DEGs) between the cancer stages (Stages I, II, III, and IV). Thus, 5271 genes were found to be differentially expressed between the cancer stages at a 0.05 threshold level.

Also, we downloaded miRNA Expression Quantification data with 1881 features. The Count per million was used to filter the miRNA features leading to 1065 features. Moreover, differential miRNA was accomplished using the linear models for microarray data (LIMMA) package in R to find the essential features that distinguish the cancer stages. This process reduced the miRNA features to 118, which significantly discriminated the stages at 0.05 level.

Additionally, we downloaded the Methylation data with 485577 features (probes), which is a very high dimensional

TABLE 1. Summary description of the datasets.

Datasets	Total number of probes	Number of probes after reduction	Total number of samples
RNA-Seq	60660	5271	531
miRNA	1881	118	531
DNA methylation	485577	2838	531
Integrated dataset	8227	-	448

TABLE 2. Summary of major model hyperparameters.

Hyperparameter	Value
Batch size	32
Learning rate	0.001
Activation function	ReLU
Number of PCA	400
Number of epochs	500

data. However, probes with missing data were removed, leaving 331959 probes. Also, features that matched the sex chromosomes were excluded leaving 325128 probes. The overlapped probes were removed, and single nucleotide polymorphisms (SNPs) with a minor allele frequency of 5% were kept. This process led to 299834 features. In addition, we removed probes that have been demonstrated to map multiple places in the genome [41], reducing the probes to 298588. The probes were further reduced to 260077 by filtering those with low expression levels using the count per million (CPM) method. These were further reduced using LIMMA to find the differentially methylated probes between the cancer stages. The differentially methylated analysis using LIMMA resulted in 2838 probes that significantly discriminate between the cancer stages. All the prepared omics data types were integrated based on the common samples, making a new dataset with 448 samples and 8227 features. **Table 1** displays the description of the datasets.

B. PROPOSED CNN MODEL

After data preprocessing, we addressed the problems of high dimensionality and class imbalance in the dataset using the PCA model. This reduced the dimensionality of the dataset by applying a different number of components to determine which was better. After that, we employed the data after the dimensionality reduction as input for the SMOTE algorithm to rebalance the classes.

We developed a CNN architecture based on the lung cancer datasets within the TensorFlow environment using the Keras library in Python programming language. Table 1 shows the PCA-SMOTE-CNN model hyperparameters that were used in this study. The structure of the proposed CNN model starts with three 1D convolutional layers that are used for the handling of one-dimensional data such as DNA/RNA expression data to extract features (probes) from the input data with a kernel size of 3, ReLU activation, and filters of 16, 32, and 64, followed by the dropout layers of values of 0.7, 0.6, and 0.5 after each 1D convolutional layer, respectively. The last 1D convolutional layers are followed by a max-pooling layer. The pooling layers with a size equal to 2 utilized, reduce the dimensionality of the input features, compress the number of parameters and data, and enable overcoming overfitting. We flatten the output of the convolutional layers to create a single long feature vector. Then dense layers with 64 units and ReLU activation function were incorporated to allow each neuron to interact with all the neurons in the previous layer. This helped to capture complex patterns in the data, along with a dropout layer to reduce overfitting with the value of 0.5. Finally, the output layer consisting of four units with softmax activation function was used for multi-class classification. The proposed CNN was designed to extract features from the input data, reduce dimensionality, and make predictions based on the learned patterns. The proposed CNN model architecture applied in this study is shown in Figure 1 as detailed above. Figure 1 demonstrates the steps we followed in this study. Initially, we obtained the datasets, subsequently processed each dataset and fed the preprocessed data to the proposed model. Table 2 displays a summary of the key hyperparameters employed in the implementation of the proposed model for the study.

C. PERFORMANCE MEASURES

We evaluated the performance of our constructed CNN model using the accuracy, precision, recall, and F1-measure. The accuracy which measuring the percentage of correctly classified cases, may not sufficiently evaluate a classifier's performance, especially with imbalanced data. Precision computes the proportion of observations predicted as positive by the model that are actual positives. On the other hand, Recall calculates the proportion of actual positive observations that the model correctly predicts as positive. Moreover, we employed the weighted average which is more advantageous over a regular average because it provides a better level of detail. It assigns different weights to the data points, reducing the influence of less important data and allowing more significant data to have a more pronounced impact on the results. This can lead to a more nuanced and accurate assessment. The equations for these metrics are provided below. Where i and j indicate the different classes. The weighted accuracy is computed using the equation 1.

$$Balanced\_Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (1)$$

In this context, TP represents instances of True Positive, TN represents instances of True Negative, FP represents instances of False Positive, and FN represents instances of

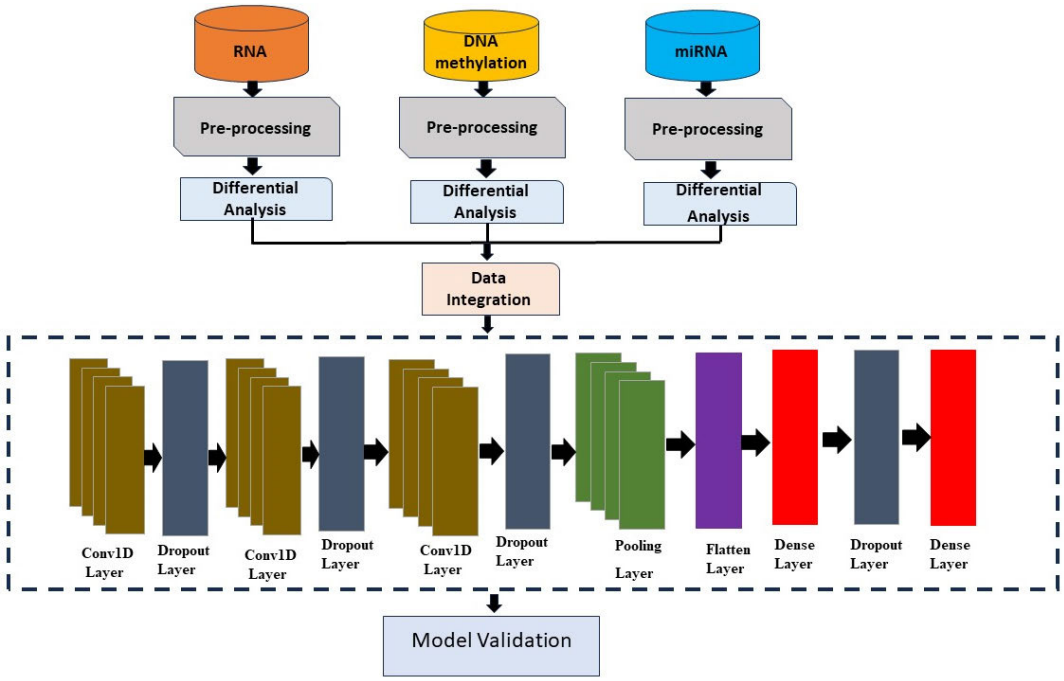

FIGURE 1. The proposed methodology.

False Negative.

$$Recall_i = M_{ii} \Big/ M_{ii} + \sum_{j=1:n;i\neq j} M_{ji} \tag{2}$$

$$Precision_i = M_{ii} \Big/ M_{ii} + \sum_{j=1:n;i\neq j} M_{ij} \tag{3}$$

Class-wise precision and recall scores were combined using micro-averaged, macro-averaged, and weighted-averaged precision and recall to derive global precision and recall values for the overall model. The weighted-average metrics represent a sample-weighted mean of class-wise precision and recall, making them suitable for assessing model performance on imbalanced datasets. The three global scores are computed as follows:

$$Micro\_Recall_i = \sum_{i:n} M_{ii} \Big/ \sum_{i:n} M_{ii} + \sum_{i:n} \left(\sum_{j=1:n;i\neq j} M_{ji} \right) \tag{4}$$

$$Micro_{Precision_i} = \sum_{i:n} M_{ii} \Big/ \sum_{i:n} M_{ii} + \sum_{i:n} \left(\sum_{j=1:n;i\neq j} M_{ij} \right) \tag{5}$$

$$Macro\_Recall_i = \sum_{i:n} Recall_i \Big/ n \tag{6}$$

$$Macro\_Precision_i = \sum_{i:n} Precision_i \Big/ n \tag{7}$$

$$Weighted\_Recall = \sum_{i:n} w_i \times Recall_i \tag{8}$$

$$Weighted\_Precission = \sum_{i:n} w_i \times Precision_i \tag{9}$$

$$\text{where } w_i = \frac{Number\ of\ samples\ in\ class\ i}{Total\ number\ of\ samples} \tag{10}$$

Ideally, we aim to assign class weights in the range of [0, 1], ensuring that the sum of weights across all classes equals 1. $0 \leq w_i \leq 1$, and $\sum_{i=1} w_i = 1$. Additionally, in imbalanced data scenarios, the less frequent class is often more significant, leading us to assign weights accordingly. A widely used formulation for this purpose is the Normalized Inverse Class Frequency.

$$w_i = 1 \Big/ \left(f_i \times \sum_{j=1}^{n} f_j \right). \tag{11}$$

$$F1score_i = \frac{2 \times Recall_i \times Precision_i}{Recall_i + Precision_i} \tag{12}$$

IV. EXPERIMENT CONFIGURATION

The conducted experiments were carried out utilizing a Lenovo computer system that is powered by an 11th Generation Intel(R) Core i5 processor, providing robust computational capabilities. Additionally, the storage component of the system is noteworthy, with a hard disk size reaching up to 952.69 gigabytes, and a storage memory capacity of up to 16 gigabytes, ensuring ample space for data processing and storage. The development of the experimental models was implemented within the Spyder platform, leveraging the programming capabilities of Python 3.10.9. This integrated development environment (IDE) facilitated the coding and implementation processes, ensuring a seamless and efficient development environment for the models under investigation.

A. RESULTS AND DISCUSSION

Table 3 shows the performance metrics per class across different numbers of PCA (100, 200, 300, and 400) employed

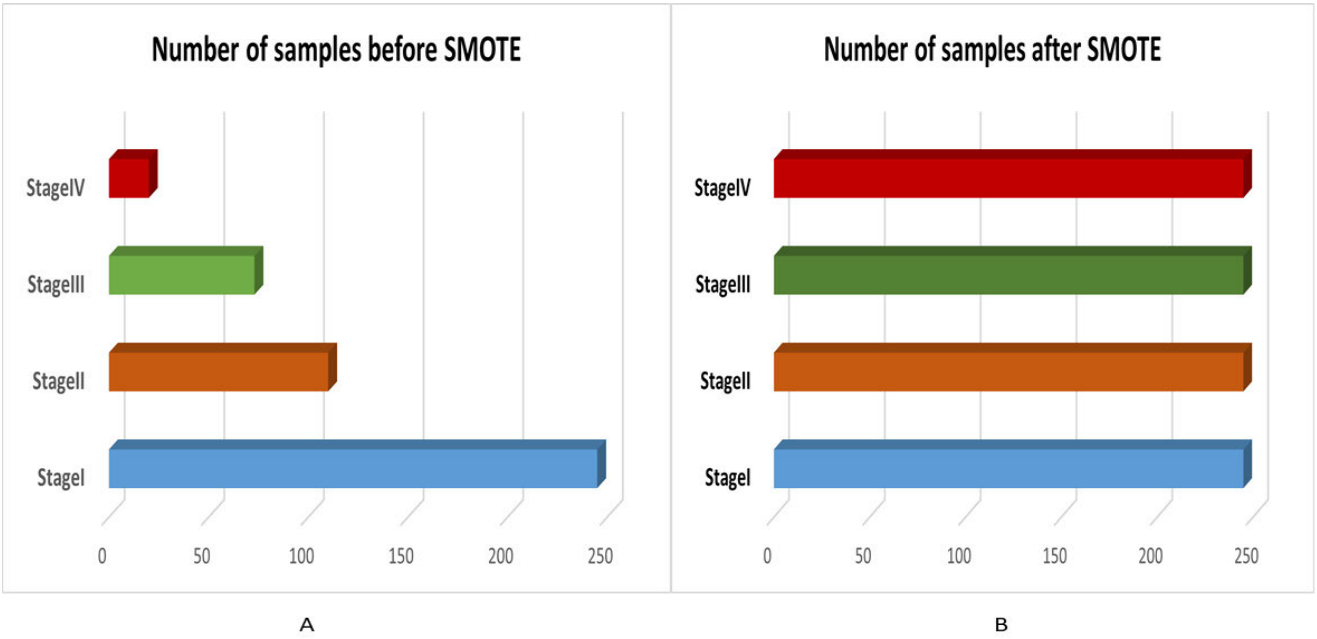

FIGURE 2. Bar graph of the number of samples within a training dataset before and after the SMOTE algorithm.

in our experiment. In stage I, it becomes clear that employing 400 PCA yields outstanding results for the proposed model, with a recall of 0.91, f1-score of 0.95, and a precision of 100% across all PCA values. In Stage II, the performance of 400 PCA surpasses other PCs, showcasing precision, recall, and f1-score of 100%. Stage III displays 100% precision, recall, and f1-score when applying 400 PCA. Meanwhile, stage IV exhibits similar outcomes when employing 200, 300, and 400 PCA. In summary, the process of picking the number of PCA has a significant impact on the model's performance in different stages, and the optimal number of PCA may vary depending on the specific stage of the experiment. It is also reported that a lower number of PCA can still result in an excellent performance in some stages. In addition, the results show that 400 PCA performed better in all stages.

Figure 2 presents bar plots of the number of samples within a training dataset before and after the SMOTE algorithm. In section A of Figure 2, it is indicated that the data utilized has a significant imbalance, with 448 samples of different stages of lung cancer. Specifically, stage I is a majority class with 245 samples, while stages II, III, and IV have 110, 73, and 20 samples, respectively. Section B of Figure 2 illustrates the data after the implementation of the SMOTE algorithm, and it is completely balanced with all classes now possessing an equal number of 245 samples.

Table 4 displays the test results across different batch sizes, highlighting a trend where increasing the batch size correlates with a reduction in performance metrics. We utilized the weighted average for the recall, precision, and f1-score. Notably, the proposed model excelled and achieved identical results for a batch size of 32. Additionally, the performance decreases when a batch.

Figure 3 illustrates the results of the proposed model with and without PCA. Based on the analysis of the plot,

TABLE 3. Classification metrics per class based on different numbers of PCA.

Class (Stages)	Number of PCA	Performance Metrics		
		Precision	Recall	F1-score
Stage I	100	100	0.73	0.84
	200	100	0.64	0.78
	300	0.88	0.64	0.74
	400	100	0.91	0.95
Stage II	100	0.60	0.75	0.67
	200	0.80	100	0.89
	300	0.75	0.75	0.75
	400	100	100	100
Stage III	100	0.70	0.88	0.78
	200	0.80	100	0.89
	300	0.80	100	0.89
	400	100	100	100
Stage IV	100	100	100	100
	200	0.88	100	0.93
	300	0.88	100	0.93
	400	0.88	100	0.93

TABLE 4. Test results of the proposed model with 400 PCA across various batch sizes.

Batch size	Performance Metrics			
	Accuracy	Precision	Recall	F1-score
32	0.97	0.97	0.97	0.97
64	0.87	0.91	0.87	0.87
128	0.83	0.90	0.83	0.83
256	0.83	0.88	0.83	0.82

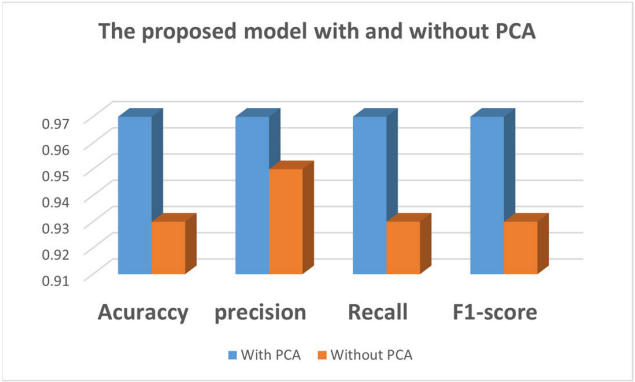

FIGURE 3. The proposed model with and without PCA.

significant differences in performance metrics are observed between the models. The proposed model achieved a recall of 0.97, while without PCA, achieved a recall of 0.93% which is lower. Furthermore, the model with PCA demonstrated a high precision of 97%, compared to 95% for the model without PCA. The accuracy of the model with and without PCA was 97% and 93%, respectively. Moreover, the F1-score improved to 97% with PCA compared to 93% without PCA. More so, it is concluded that the usage of PCA enhances the model's ability to precisely identify positive cases and reduce false negatives simultaneously, thus increasing the model's accuracy.

Figure 4 illustrates the outcomes of the proposed model in terms of the training and validation accuracy, as well as the loss values across different batch sizes (32, 64, 128, and 256) with 400 PCA to determine the optimal batch size over 500 epochs. In parts A, B, C, and D of Figure 3, it is shown that the loss values of the training and validation decreased continuously approaching zero. This is a good indication that the proposed model learned effectively from the training set. Furthermore, the training and validation accuracy in part A surpasses that of parts B, C, and D. This indicates that the proposed model shows strong generalization capabilities when making predictions on new data. These graphical representations distinctly illustrate that increasing the batch size leads to an elevation in loss values. Therefore, it is evident that the most suitable batch size for the proposed model is 32.

Figure 5 shows the confusion matrix of our proposed model based on the 400 selected PCs for various batch sizes, where part (A) indicates the model performance when a batch size of 32 is utilized, while batch sizes of 64, 128, and 256 are depicted in parts (B), (C), and (D), respectively. Overall, it is observed that the model with a batch size of 32 surpasses other batch sizes.

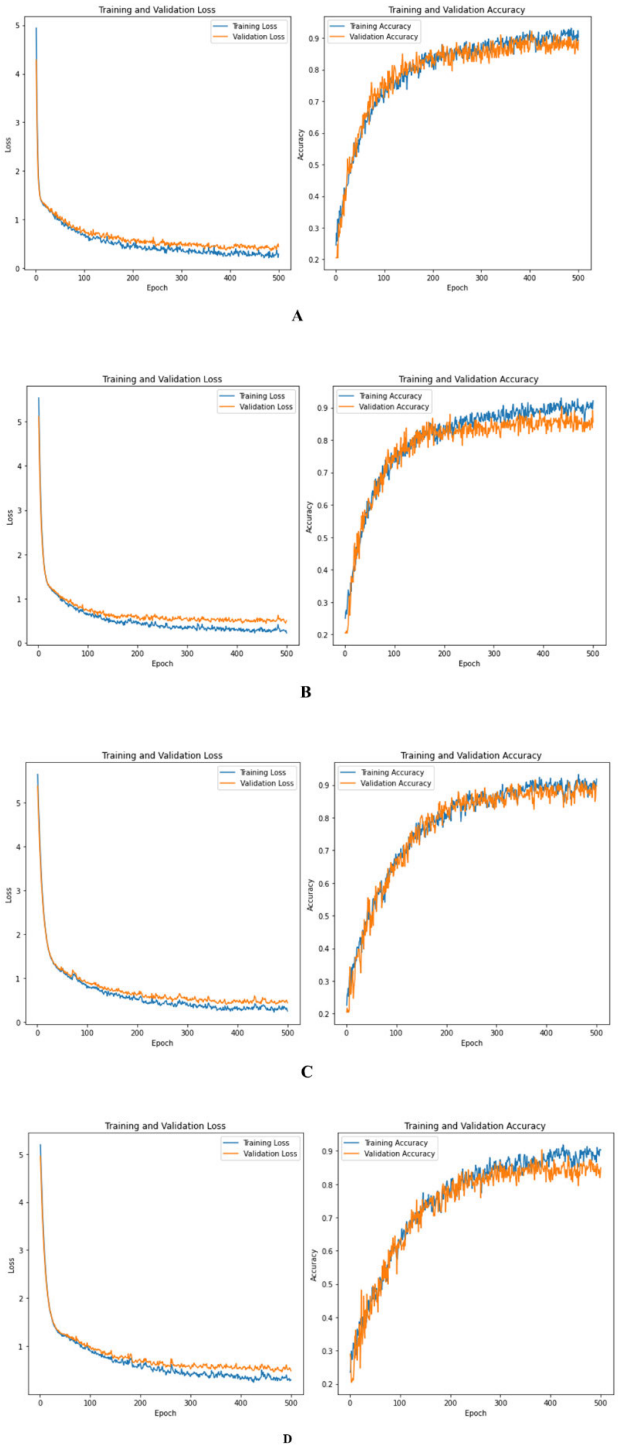

FIGURE 4. Accuracy and loss during training and validation across varying batch sizes. In Figures A, B, C, and D of Figure 3, it is illustrated that the loss values for both training and validation consistently decreased, approaching zero.

B. COMPARISON WITH VARIOUS MODELS

Table 5 provides a comparative analysis of our PCA-SMOTE-CNN model against other solutions such as LungDWM (Lung cancer subtype Diagnosis using Weakly paired Multiomics data), convolutional variational autoencoder (CVAE) or CVAE-based (MetaCancer), CC2DT, and SVMEnsemble that were discussed in the studies presented in [37], [38], and [40], respectively. All models utilized the integrated mRNA, miRNA, and DNA methylation dataset.

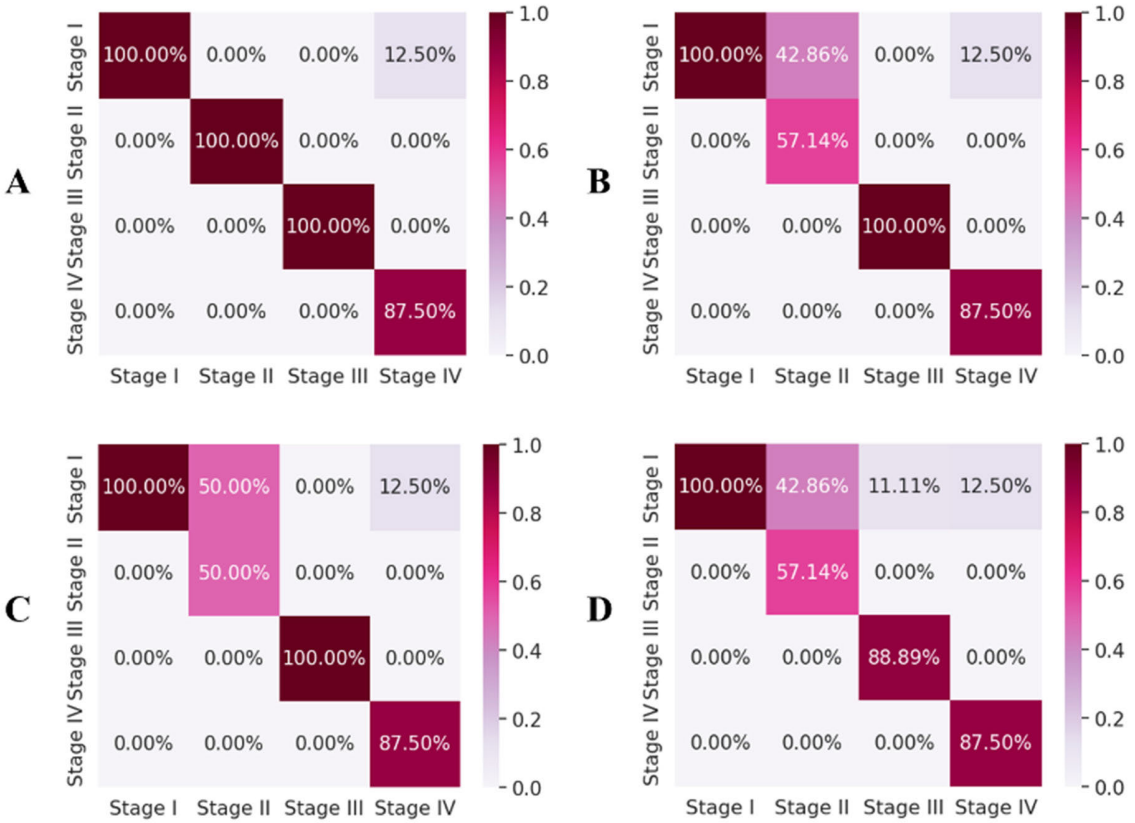

FIGURE 5. Confusion matrix of the proposed model based on 400 PCA using 32, 64, 128, and 256 batch sizes in (A), (B), (C), and (D), respectively.

For fair comparisons with literature findings, we had to adjust our model parameters and experimental conditions to match closely with the aforementioned approaches. Our method shows an accuracy of 0.97 outperforming LungDWM (0.942), CVAE-based (0.888), CC2DT (0.824), and SVMEnsemble (0.825). Furthermore, PCA-SMOTE-CNN achieved a precision and f1-score of 0.97 surpassing the f1-score for LungDWM (0.937), CVAE-based (0.904), CC2DT (0.887), and SVMEnsemble (0.829). In addition, the proposed model achieved a precision of 0.970 outperforming the precision score for CVAE-based (0.916), CC2DT (0.855), and SVMEnsemble (0.810). The proposed model exhibited a recall of 0.97 outperforming the CVAE-based (0.876), CC2DT(0.899), and SVMEnsemble (0.850). Moreover, we also implemented and utilized the long short-term memory (LSTM) and multi-layer perceptron (MLP) models on our integrated dataset. LSTM model achieved an accuracy of 0.70, precision of 0.71, recall of 0.76, and F1-score of 0.69. However, the MLP model scored an accuracy of 0.93, precision of 0.92, recall of 0.95, and F1-score of 0.93.

In summary, the proposed PCA-SMOTE-CNN model appeared as a promising and effective approach for lung cancer classification using integrated multi-omics data, proving excellent performance compared to the alternative models.

Table 6 demonstrates the comparison between the proposed model with various machine learning models including K-nearest neighbors (KNN), Support Vector Machines (SVM), decision trees (DT), GaussianNaive Bayes (GNB), and Random Forests (RF) using the same integrated multi-omics data and experimental condition. The

TABLE 5. Comparative evaluation of our proposed approach to other deep learning methods that integrate RNA, miRNA, and DNA methylation data.

Method	Performance Metrics			
	Accuracy	Precision	Recall	F1-score
PCA-SMOTE-CNN	**0.970**	**0.970**	**0.970**	**0.970**
LungDWM [40]	0.942	-	-	0.937
CVAE-based (MetaCancer) [39]	0.888	0.916	0.876	0.904
CC2DT [38]	0.824	0.855	0.899	0.887
SVMEnsemble [39]	0.825	0.810	0.850	0.829
MLP	0.93	0.92	0.95	0.93
LSTM	0.70	0.71	0.76	0.69

TABLE 6. Comparison between the proposed model and different machine learning models based on the integrated data.

Models	Accuracy	Precision	Recall	F1-score
PCA-SMOTE-CNN	0.97	0.97	0.97	0.97
SVM	0.93	0.92	0.93	0.94
RF	0.53	100	0.533	0.61
DT	0.80	0.84	0.80	0.80
GNB	0.90	0.94	0.90	0.91
KNN	0.57	0.47	0.57	0.47

PCA-SMOTE-CNN model performed better than other models achieving an accuracy, precision, recall, and an F1-score of 0.970. Notably, SVM, GNB, and DT, also performed well, with an accuracy of 0.93, 0.90, and 0.80 respectively. In contrast, RF observed low performance with an accuracy of 0.53 compared to other models. Overall, the PCA-SMOTE-CNN model showcased outstanding performance, highlighting its effectiveness in comparison to other models. The model performance demonstrates that incorporating RNASeq, miRNA, and DNA methylation data types further allows the machine learning algorithm to understand detailed patterns in the dataset. This incorporation of different molecular data gives the model a more complex understanding of those underlying biological factors that contribute to lung cancer. Therefore, this can lead to more precise classifications and predictions.

To fully test this model, three separate experiments were conducted for comparison as illustrated in **Figure 6**. Next, the applied DeepMO [42] computational model employs deep neural networks structured on multi-omics data. This involved the integration of three diverse omics data types: Data on copy number variation (CNV) data, mRNA, and DNA methylation. The result was an accuracy of 0.77. In contrast, multi-omics graph convolutional networks (MOGONET) [43] proposed a novel multi-omics approach that integrated mRNA expression data and DNA methylation information with microRNAs to discriminate breast invasive carcinoma and Alzheimer's Disease patients against normal controls. The accuracy of the MOGONET model was 0.81. However, the final accuracy of an ensemble of decision trees with gradient boosting (XGBoost) for applying to LUAD on the integrated training dataset was 0.85 %. The most important observation is that the proposed model seemed to work better than the other models and in fact had a higher efficiency.

Figure 7 presents the results of performance scores obtained with the PCA-SMOTE-CNN model using single omics data or integrated omics (miRNA, RNASeq, and DNA methylation). The results indicate that the PCA-SMOTE-CNN model demonstrates superior performance when integrated with multiple omics data compared to using single omics with accuracy, precision, recall, and F1 scores of 0.93, 0.96, 0.93, and 0.94 respectively. On the other hand, RNASeq

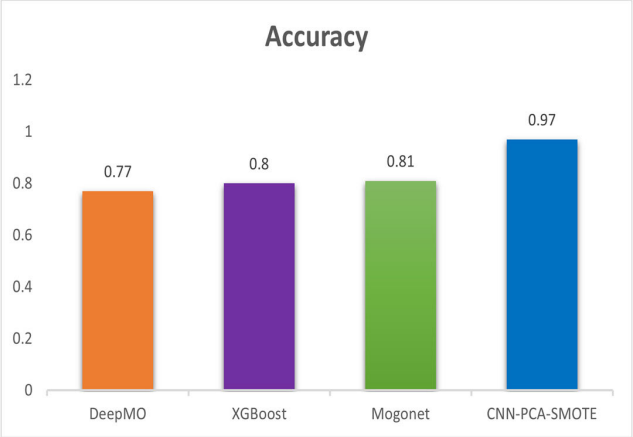

FIGURE 6. Comparison between the proposed model with different model.

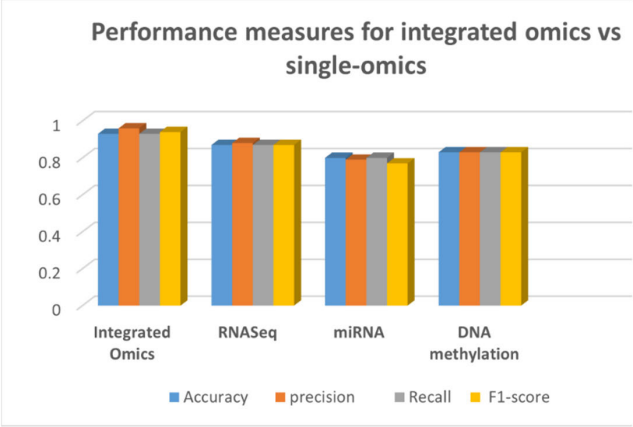

FIGURE 7. PCA-SMOTE-CNN performance when using the integrated omics (mRNA, DNA methylation, RNASeq) dataset or Single-omics data.

accomplished an accuracy of 0.87, precision of 0.89, recall of 0.87, and F1 score of 0.87. DNA methylation performed 0.83 for all metrics. miRNA had the lowest performance with accuracy, precision, recall, and F1-score of 0.8, 0.79, 0.8, and 0.77 respectively. To compare, we utilized 100 PCA components because there are relatively few probes (118) for miRNA data.

Figure 8 illustrates the area under the curve (AUC) evaluation metric when utilizing a single or a combination of omics types. When using a single-omics type, miRNA performed the best with AUC = 0.98, DNA methylation had a performance with AUC = 0.92, and RNASeq ranked second with AUC = 0.95. However, using integrated multi-omics data enhanced the performance and achieved AUC = 1.00.

C. PROPOSED MODEL STRENGTHS AND LIMITATIONS

The strengths of the proposed architecture of the CNN model lie in its ability to leverage multiple classification models through a combination of PCA and SMOTE. PCA helps to reduce the data dimensionality, potentially improving model performance and generalization. Conversely, SMOTE

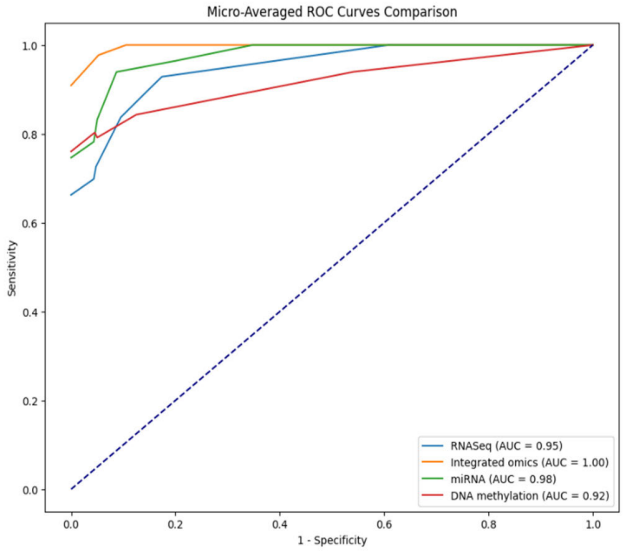

FIGURE 8. **Micro-averaged ROC curves using single-omics vs integrated omics.**

addresses class imbalances, which leads to a more balanced dataset. However, despite the aforementioned merit, the limited availability of high-quality multi-omics data for lung cancer may impact the robustness of the model. Moreover, inconsistencies or inaccuracies in the collected data may introduce bias and affect the reliability of the results.

Further suggested improvement can concentrate on combining more classification models with the CNN model, optimizing the hyperparameters, utilizing different techniques for class imbalances, combining more methods for data reduction, and identifying Hub Genes. Despite these limitations, the proposed PCA-SMOTE-CNN model significantly outperformed other models developed in previous studies. Further research is necessary to evaluate the PCA-SMOTE-CNN model's performance on other medical problems.

V. CONCLUSION

This study focused on the comprehensive analysis of LUAD development and progression by applying combined datasets. To improve lung cancer classification and detection, we designed a deep learning model that utilizes integrated data from RNASeq, miRNA, and DNA methylation markers. The experimental results indicated the effectiveness of the proposed method in classifying and predicting lung cancer using the integrated dataset. Comparative analysis with recent competitive techniques demonstrated that our proposed method has an outstanding prediction performance, as indicated by various evaluation metrics such as accuracy, precision, recall, and F1-score. This indicates the potential of our integrated approach to improve the diagnosis and understanding of LUAD, contributing significant insights to the domain of lung cancer-related research.

Future research in the domain of deep learning models for improved classification and prediction of lung cancer using multi-omics data could explore several promising directions to enhance the field. For example, researchers can investigate the inclusion of additional omics data types, such as proteomics and metabolomics, to create a more comprehensive and holistic view of the molecular landscape associated with lung cancer. This expansion could potentially provide richer insights into the underlying mechanisms and facilitate more accurate predictions. Moreover, the focus on developing methods to enhance the interpretability of deep learning models in the context of multi-omics data can be looked into or investigated. Furthermore, transparent and interpretable models are crucial for gaining trust in the clinical application of these models and can aid researchers and clinicians in understanding the biological significance of model predictions. As regards the implementation of a more robust model, researchers could explore the option of transfer learning techniques that leverage pre-trained models on related cancer types or datasets. Transfer learning has the potential to improve model performance, especially when faced with limited labeled data for lung cancer, by transferring knowledge gained from other well-annotated datasets. Another interesting study would be to conduct longitudinal studies to capture the dynamic changes in omics profiles over time. Longitudinal data can provide insights into the progression of lung cancer and aid in the development of models that consider temporal aspects, potentially leading to more accurate predictions and personalized treatment strategies.

DATA AVAILABILITY

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

CONFLICT OF INTEREST STATEMENT

The authors declare that no conflict of interest would hinder the publication of this paper.

AUTHOR CONTRIBUTIONS

Tehnan I. A. Mohamed and Absalom El-Shamir Ezugwu : Conceptualization; Tehnan I. A. Mohamed: Methodology; Tehnan I. A. Mohamed: Formal Analysis; Tehnan I. A. Mohamed: Writing original draft; Absalom El-Shamir Ezugwu : Supervision; Absalom El-Shamir Ezugwu : Writing review and editing. All authors read and approved the final manuscript.

Abbreviations	Meaning
NCDs	Non-communicable diseases
SCLC	Small Cell Carcinoma
NSCLC	Non-Small Cell Carcinoma
scRNA-seq	Single-cell RNA sequencing
RNAs	Cellular ribonucleic acids
mRNA	messenger RNA
PCA	principal component analysis
SMOTE	Synthetic Minority Over-sampling Technique

CNN	Convolutional Neural Networks
FFN	Feed Forward Neural Networks
RNN	Recurrent Neural Networks
AE	Autoencoders
LUAD	Lung adenocarcinoma
PPI	protein-protein interaction
cfDNA	cell-free DNA
TOO	tumor tissue-of-origin
CUP	carcinoma of unknown primary
ceRNA	competitive endogenous RNA
KEGG	Kyoto encyclopedia of genes and genomes
GO	gene ontology
SVM	Support Vector Machine
ASD	autism spectrum disorder
TCGA	The Cancer Genome Atlas
SNPs	single nucleotide polymorphisms
CPM	count per million
CNV	copy number variation
MOGONET	multi-omics graph convolutional networks
LSTM	long short-term memory
MLP	multi-layer perceptron

REFERENCES

[1] L. Manderson and S. Jewett, "Risk, lifestyle and non-communicable diseases of poverty," *Globalization Health*, vol. 19, no. 1, pp. 1–9, Mar. 2023.

[2] T. Lancet, "GLOBOCAN 2018: Counting the toll of cancer," *Lancet*, vol. 392, no. 10152, p. 985, Sep. 2018.

[3] L. Davies, D. A. Milner, L. N. Shulman, L. Kyokunda, A. Bedada, P. Vuylsteke, N. Masalu, P. Jackson, N. Jennings, A. Odunlami, P. Mtshali, and U. Dugan, "Analysis of cancer research projects in sub-saharan africa: A quantitative perspective on unmet needs and opportunities," *JCO Global Oncol.*, vol. 9, Jun. 2023, Art. no. e2200203.

[4] R. L. Siegel, "Cancer statistics," *Ca Cancer J. Clin.*, vol. 73, no. 1, pp. 17–48, 2023.

[5] U. G. Assembly, "Political declaration of the third high-level meeting of the General Assembly on the prevention and control of non-communicable diseases," in *Resolution Adopted by the General Assembly*. New York, NY, USA: United Nations Digital Library, Oct. 2018. [Online]. Available: https://digitallibrary.un.org/record/1645265?ln=en&v=pdf

[6] H. Fitipaldi and P. W. Franks, "Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005–2022," *Human Mol. Genet.*, vol. 32, no. 3, pp. 520–532, Jan. 2023.

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

[8] K. Huang, Y. Zhang, X. Shi, Z. Yin, W. Zhao, L. Huang, F. Wang, and X. Zhou, "Cell-type-specific alternative polyadenylation promotes oncogenic gene expression in non-small cell lung cancer progression," *Mol. Therapy Nucleic Acids*, vol. 33, pp. 816–831, Sep. 2023.

[9] Organization. (2023). *W.H. Lung Cancer.* [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[10] J. N. Bodor, Y. Boumber, and H. Borghaei, "Biomarkers for immune checkpoint inhibition in non–small cell lung cancer (NSCLC)," *Cancer*, vol. 126, no. 2, pp. 260–270, Jan. 2020.

[11] L. Ginn, L. Shi, M. La Montagna, and M. Garofalo, "LncRNAs in non-small-cell lung cancer," *Non-Coding RNA*, vol. 6, no. 3, p. 25, Jun. 2020.

[12] L. Cai, S. Lin, L. Girard, Y. Zhou, L. Yang, B. Ci, Q. Zhou, D. Luo, B. Yao, H. Tang, J. Allen, K. Huffman, A. Gazdar, J. Heymach, I. Wistuba, G. Xiao, J. Minna, and Y. Xie, "LCE: An open web portal to explore gene expression and clinical associations in lung cancer," *Oncogene*, vol. 38, no. 14, pp. 2551–2564, Apr. 2019.

[13] S. Osama, H. Shaban, and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118946.

[14] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, p. 173, Jan. 2023.

[15] Q. Xue, W. Peng, S. Zhang, X. Wei, L. Ye, Z. Wang, X. Xiang, P. Zhang, and Q. Zhou, "Promising immunotherapeutic targets in lung cancer based on single-cell RNA sequencing," *Frontiers Immunol.*, vol. 14, Apr. 2023, Art. no. 1148061.

[16] W. Ren, Y. Yuan, J. Peng, L. Mutti, and X. Jiang, "The function and clinical implication of circular RNAs in lung cancer," *Frontiers Oncol.*, vol. 12, Oct. 2022, Art. no. 862602.

[17] X. Li, L. Yang, and L.-L. Chen, "The biogenesis, functions, and challenges of circular RNAs," *Mol. Cell*, vol. 71, no. 3, pp. 428–442, Aug. 2018.

[18] M. H. Alderman and A. Z. Xiao, "N (6)-Methyladenine in eukaryotes," *Cellular Mol. Life Sci.*, vol. 76, pp. 2957–2966, Aug. 2019.

[19] X. Han, J. Guo, and Z. Fan, "Interactions between m6A modification and miRNAs in malignant tumors," *Cell Death Disease*, vol. 12, no. 6, p. 598, Jun. 2021.

[20] C. Park, J. Ha, and S. Park, "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset," *Expert Syst. Appl.*, vol. 140, Feb. 2020, Art. no. 112873.

[21] M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, and C. Trapnell, "Dimensionality reduction by UMAP to visualize physical and genetic interactions," *Nature Commun.*, vol. 11, no. 1, p. 1537, Mar. 2020.

[22] K. Ghimire, "Machine learning approach to distinguish ulcerative colitis and Crohn's disease using SMOTE (Synthetic minority oversampling Technique) methods," *SMU Data Sci. Rev.*, vol. 5, no. 2, p. 9, 2021.

[23] L. Y. Venkataramana, "Geometric SMOTE-based approach to improve the prediction of Alzheimer's and Parkinson's diseases for highly class-imbalanced data," in *AI, IoT, and Blockchain Breakthroughs in E-Governance*. Hershey, PA, USA: IGI Global, 2023, pp. 114–137.

[24] E. Ismail, W. Gad, and M. Hashem, "A hybrid stacking-SMOTE model for optimizing the prediction of autistic genes," *BMC Bioinf.*, vol. 24, no. 1, p. 379, Oct. 2023.

[25] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, "A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare," *BioData Mining*, vol. 16, no. 1, p. 15, Apr. 2023.

[26] U. Ravindran and C. Gunavathi, "A survey on gene expression data analysis using deep learning methods for cancer diagnosis," *Prog. Biophys. Mol. Biol.*, vol. 177, pp. 1–13, Jan. 2023.

[27] R. K. Mondol, E. K. A. Millar, P. H. Graham, L. Browne, A. Sowmya, and E. Meijering, "Hist2RNA: An efficient deep learning architecture to predict gene expression from breast cancer histopathology images," *Cancers*, vol. 15, no. 9, p. 2569, Apr. 2023.

[28] R. Majji, B. Maram, and R. Rajeswari, "Chronological horse herd optimization-based gene selection with deep learning towards survival prediction using PAN-cancer gene-expression data," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104696.

[29] G. A. A. Mulla, Y. Demir, and M. Hassan, "Combination of PCA with SMOTE oversampling for classification of high-dimensional imbalanced data," *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 10, no. 3, pp. 858–869, Sep. 2021.

[30] S. Sakib, A. K. Tanzeem, I. K. Tasawar, F. Shorna, Md. A. B. Siddique, and S. B. Alam, "Blood cancer recognition based on discriminant gene expressions: A comparative analysis of optimized machine learning algorithms," in *Proc. IEEE 12th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2021, pp. 0385–0391.

[31] Z. Yang, B. Liu, T. Lin, Y. Zhang, L. Zhang, and M. Wang, "Multiomics analysis on DNA methylation and the expression of both messenger RNA and microRNA in lung adenocarcinoma," *J. Cellular Physiol.*, vol. 234, no. 5, pp. 7579–7586, May 2019.

[32] A. Kutlay and Y. A. Son, "Integrative predictive modeling of metastasis in melanoma cancer based on MicroRNA, mRNA, and DNA methylation data," *Frontiers Mol. Biosci.*, vol. 8, Sep. 2021, Art. no. 637355.

[33] Y. Su, A. Shetty, and F. Jiang, "Integrated analysis of miRNAs and DNA methylation identifies miR-132-3p as a tumor suppressor in lung adenocarcinoma," *Thoracic Cancer*, vol. 11, no. 8, pp. 2112–2124, Aug. 2020.

[34] E. Tomeva, O. J. Switzeny, C. Heitzinger, B. Hippe, and A. G. Haslberger, "Comprehensive approach to distinguish patients with solid tumors from healthy controls by combining androgen receptor mutation p.H875Y with cell-free DNA methylation and circulating miRNAs," *Cancers*, vol. 14, no. 2, p. 462, Jan. 2022.
[35] M. Shujaat, H. Kim, H. Tayara, and K. T. Chong, "iProm-sigma54: A CNN base prediction tool for σ^{54} promoters," *Cells*, vol. 12, no. 6, p. 829, Mar. 2023.
[36] R. S. Varghese, M. E. Barefoot, S. Jain, Y. Chen, Y. Zhang, A. Alley, A. H. Kroemer, M. G. Tadesse, D. Kumar, Z. A. Sherif, and H. W. Ressom, "Integrative analysis of DNA methylation and microRNA expression reveals mechanisms of racial heterogeneity in hepatocellular carcinoma," *Frontiers Genet.*, vol. 12, Sep. 2021, Art. no. 708326.
[37] X. Guan, Y. Yao, G. Bao, Y. Wang, A. Zhang, and X. Zhong, "Diagnostic model of combined ceRNA and DNA methylation related genes in esophageal carcinoma," *PeerJ*, vol. 8, p. e8831, Mar. 2020.
[38] Z. Rong, D. Lingyun, L. Jinxing, and G. Ying, "Diagnostic classification of lung cancer using deep transfer learning technology and multi-omics data," *Chin. J. Electron.*, vol. 30, no. 5, pp. 843–852, Sep. 2021.
[39] S. Albaradei, F. Napolitano, M. A. Thafar, T. Gojobori, M. Essack, and X. Gao, "MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4404–4411, 2021.
[40] X. Wang, G. Yu, J. Wang, A. M. Zain, and W. Guo, "Lung cancer subtype diagnosis using weakly-paired multi-omics data," *Bioinformatics*, vol. 38, no. 22, pp. 5092–5099, Nov. 2022.
[41] Y.-A. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg, "Discovery of cross-reactive probes and polymorphic CpGs in the illumina infinium HumanMethylation450 microarray," *Epigenetics*, vol. 8, no. 2, pp. 203–209, Feb. 2013.
[42] Y. Lin, W. Zhang, H. Cao, G. Li, and W. Du, "Classifying breast cancer subtypes using deep neural networks based on multi-omics data," *Genes*, vol. 11, no. 8, p. 888, Aug. 2020.
[43] T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature Commun.*, vol. 12, no. 1, p. 3445, Jun. 2021.
[44] H. Z. Almarzouki, "Deep-learning-based cancer profiles classification using gene expression data profile," *J. Healthcare Eng.*, vol. 2022, pp. 1–13, Jan. 2022, doi: 10.1155/2022/4715998.
[45] J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, and M. M. Khan, "A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data," *IEEE Access*, vol. 7, pp. 22086–22095, 2019.
[46] A. K. Dwivedi, "Artificial neural network model for effective cancer classification using microarray gene expression data," *Neural Comput. Appl.*, vol. 29, no. 12, pp. 1545–1554, Jun. 2018.
[47] S. Tarek, R. A. Elwahab, and M. Shoman, "Gene expression based cancer classification," *Egyptian Informat. J.*, vol. 18, no. 3, pp. 151–159, Nov. 2017.

TEHNAN I. A. MOHAMED received the B.Sc. degree (Hons.) in computer science from the University of Gezira, Sudan, and the M.Sc. degree in computer science from the University of KwaZulu-Natal, South Africa, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include applied artificial intelligence, specifically in using machine learning models for medical image analysis.

ABSALOM EL-SHAMIR EZUGWU received the B.Sc. degree in mathematics with computer science and the M.Sc. and Ph.D. degrees in computer science from Ahmadu Bello University, Zaria, Nigeria. He currently holds the position of a Full Professor of computer science with the Unit for Data Science and Computing, North-West University, Potchefstroom, South Africa. He has contributed significantly to the academic community through the publication of numerous papers in internationally refereed journals, edited books, conference proceedings, and local journals. His research interests include artificial intelligence, swarm intelligence, and nature-inspired algorithm design, with a specific emphasis on computational intelligence and metaheuristic solutions for real-world global optimization problems. He is an Active Member of prominent organizations, such as the Association for Computing Machinery (ACM), the International Association of Engineers (IAENG), and the Operations Research Society of South Africa (ORSSA).