OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf115
https://doi.org/10.1093/bib/bbaf115
Problem Solving Protocol

DOMSCNet: a deep learning model for the classification of stomach cancer using multi-layer omics data

Kasmika Borah[1], Himanish Shekhar Das[1,*], Ram Kaji Budhathoki[2,*], Khursheed Aurangzeb[3], Saurav Mallik[4,5,*]

[1]Department of Computer Science and Information Technology, Cotton University, Hem Baruah Rd, Panbazar, Guwahati, Kamrup Metropolitan district, Assam 781001, India
[2]Department of Electrical and Electronics Engineering, School of Engineering, Kathmandu University, Kavrepalanchok district, Dhulikhel 45200, Nepal
[3]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, P. O. Box 51178, Riyadh district, 11543, Saudi Arabia
[4]Department of Environmental Health, Harvard T. H. Chan School of Public Health, 665 Huntington Avenue, Boston, MA 02115, United States
[5]Department of Pharmacology & Toxicology, University of Arizona, 1295 N Martin Ave, Pima district, Tucson, AZ 85721, United States

[*]Corresponding authors. Ram Kaji Budhathoki, Department of Electrical and Electronics Engineering, School of Engineering, Kathmandu University, Dhulikhel 45200, Nepal. E-mail: ram.budhathoki@ku.edu.np; Himanish Shekhar Das, Department of Computer Science and Information Technology, Cotton University, Panbazar, Guwahati 781001, India. E-mail: himanish.das@cottonuniversity.ac.in; Saurav Mallik, Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, MA 02115, United States. E-mail: sauravmtech2@gmail.com.

Abstract

The rapid advancement of next-generation sequencing (NGS) technology and the expanding availability of NGS datasets have led to a significant surge in biomedical research. To better understand the molecular processes, underlying cancer and to support its development, diagnosis, prediction, and therapy; NGS data analysis is crucial. However, the NGS multi-layer omics high-dimensional dataset is highly complex. In recent times, some computational methods have been developed for cancer omics data interpretation. However, various existing methods face challenges in accounting for diverse types of cancer omics data and struggle to effectively extract informative features for the integrated identification of core units. To address these challenges, we proposed a hybrid feature selection (HFS) technique to detect optimal features from multi-layer omics datasets. Subsequently, this study proposes a novel hybrid deep recurrent neural network-based model DOMSCNet to classify stomach cancer. The proposed model was made generic for all four multi-layer omics datasets. To observe the robustness of the DOMSCNet model, the proposed model was validated with eight external datasets. Experimental results showed that the SelectKBest-maximum relevancy minimum redundancy-Boruta (SMB), HFS technique outperformed all other HFS techniques. Across four multi-layer omics datasets and validated datasets, the proposed DOMSCNet model outdid existing classifiers along with other proposed classifiers.

Keywords: hybrid feature selection; hybrid deep learning; multi-layer omics; molecular signature; classification

Introduction

Multi-layer omics profiles play a crucial role in classifying varied cancer tumor stages for disease diagnosis and provide different layers of biological molecular insights to understand the disease pathophysiology against treatment [1]. Cancer is a heterogeneous disease that deregulates cellular functions across various cell molecules, where these molecules interact and mutually influence each other in reprogramming cellular functions. Recently, advanced next-generation sequencing (NGS) technologies have increased tremendously in interpreting cancer classification, prediction, and treatment [1]. However, NGS multi-layer omics datasets are inherently complex, characterized by high dimensionality, extensive redundancy, and significant heterogeneity, making it difficult for machine learning and deep learning (DL) models to generalize well [2–4]. Reducing the dimensionality of data generated by NGS technology is a complex task. Feature selection (FS) is one of the important approaches that can reduce dimensionality from the high-dimensional NGS data and improve the model performance during classification, and identification [3, 5, 6]. The FS techniques not only enhance model performance but it can also able to identify molecular signatures or markers or biomarkers and putative target agents and help in understanding the molecular signatures function for specific diseases.

The field of DL has developed a multitude of techniques for diverse downstream analysis, and recently, these methods have been applied to a broad spectrum of problems, exhibiting considerable promise in biomedical applications [7]. DL in multi-omics data classification etc. has emerged as a powerful approach to integrate and analyze different layers of biological information, to better understand diseases like cancer [8, 9]. The “curse of dimensionality” presents a major obstacle when applying DL models to categorize NGS data [9, 10]. However, various existing methods face challenges in accounting for diverse types of cancer omics data and struggle to effectively extract or detect informative features for the integrated identification of core units.

In this study, we set out to address the challenge of building a DL-based model for classification between stomach cancer (SC) and normal tissue samples using multi-layer omics data. The SC or gastric cancer is the second dominant cause of

Received: December 8, 2024. **Revised:** January 31, 2025. **Accepted:** February 20, 2025

cancer-related mortalities, with the majority of diagnoses and fatalities occurring in Asia [11]. The Lauren Classification and the World Health Organization (WHO) Classification are the two most popular methods for histologically classifying SC [12]. Based on tissue form and growth patterns, the Lauren Classification classifies stomach adenocarcinomas into three primary categories: mixed, diffuse, and intestinal. More specifically, the WHO classification categorizes SC into the following histological subtypes: tubular adenocarcinoma, mucinous adenocarcinoma, papillary adenocarcinoma, and signet-ring cell carcinoma. Four molecular subtypes of SC have been identified by the Cancer Genome Atlas (TCGA) project: chromosomal instability (CIN), Epstein–Barr virus (EBV), microsatellite instability (MSI), and genomic stability [13–16]. EBV-positive tumors have significant levels of DNA hypermethylation, amplification of PD-L1 and PD-L2, as well as recurrent mutations in *PIK3CA* and *ARID1A*. MSI tumors are more prevalent in the elderly and patients and are characterized by a lack of mismatch repair. Low somatic copy number changes, and *RHOA* and *CDH1* mutations are found in stomach malignancies. RTK/Ras-focused amplification and non-diploidy are characteristics of CIN tumors, which are more prevalent in the gastroesophageal junction. In this study, the primary tumor of stomach adenocarcinomas over adjacent normal solid tissue dataset of the TCGA-Stomach Adenocarcinoma (STAD) project was used to conduct the experiments.

The initial goal of this study is to identify robust feature patterns through hybrid feature selection (HFS) techniques for classifying primary tumors of SC and identifying molecular signatures using multi-layer omics datasets namely Exon, mRNA, miRNA expression, and DNA CpG site methylation, respectively. Reducing the number of features upfront minimizes the risk of overfitting, improves computational efficiency, and allows the model to focus on the most relevant information for classification and prediction. In the proposed study, we proposed a hybrid deep recurrent neural network (HDRNN)–based model for the classification of primary tumor of SC over normal samples by integrating the HFS technique. The proposed HFS techniques combine filter–filter–wrapper and filter–filter–embedded approaches. The hybrid techniques are applied to improve the robustness of the model during classification by identifying the most relevant features from the multi-layer omics data. Thereafter, we employed existing classification models of long short-term memory (LSTM), grated recurrent unit (GRU), and their extension with the features obtained from HFS techniques. The experimental results of existing and proposed models' (LSTM, GRU, and their extensions) performance with HFS techniques did not show any improvement in terms of accuracy score for all multi-layer omics data. Consequently, we proposed three hybrid DRNN models namely (i) Bidirectional Long Short-Term Memory (BiLSTM) - Bidirectional Gated Recurrent Unit (BiGRU) as (DOMSCNet), (ii) BiLSTM–BiGRU-attention, and (iii) BiLSTM-attention–BiGRU-attention to classify the SC versus normal solid tissue samples. To observe the robustness of the proposed model, validation experiments were performed with eight external multi-layer omics datasets of NCBI Geodataset and TCGA-Liver Hepatocellular Carcinoma (LIHC) project.

Related works

Recently, the DL approach entered into a transcriptomic, epigenetic, and genomic era to evaluate cancer diseases. In this section, the methodology of the existing works is surveyed based on two criteria: (i) HFS techniques for dimensionality reduction of omics data and (ii) DL-based algorithms for the classification of cancer utilizing various omics data. Several DL-based classification methods, including deep neural networks (DNNs), convolutional neural networks (CNNs), Recurrent Neural Network (RNN), CNN–RNN integration, etc., were introduced in the existing study. In recent years, Dutta *et al.* [9] developed a self-attention-based deep multi-criteria model for the analysis of disease prognosis using transcriptomic and proteomics datasets. Bhonde *et al.* [17] used hybrid particle swarm optimization, principal component analysis (PCA), and random forest to reduce dimensionality from gene expression data. Bhonde *et al.* [17] again applied CNN and Bidirectional LSTM (BiLSTM) classifiers for the classification of cancer and achieved an accuracy of 96.89. Metipatil *et al.* [18] proposed a CNN–BiLSTM model for predicting cancer using a gene expression dataset and achieved an accuracy of 98.3. Susmi [19] developed an optimal BiLSTM with a self-attention mechanism and remora optimization algorithm for the classification of gene expression data. Babichev *et al.* [20] used GRU and LSTM models for the classification of gene expression data and GRU achieved the highest accuracy of 97.2. Mallick *et al.* [21] employed LSTM, BiLSTM, and feed forward neural network for analyzing the gene–gene interaction dependencies for cancer disease. The MRMR-mv FS technique was introduced for consolidative analyses and predictive models from multi-omics data and achieved an average area under curves (AUC) score of 0.53 (methylation data) and 0.64 (RNA-Seq data) [22]. Sahu proposed an HFS framework that effectively combines filter and wrapper methods to enhance cancer classification accuracy in multi-omics data by identifying relevant miRNA subsets [23]. Tabakhi and Lu [24] proposed a multi-agent architecture model to integrate all omics data and enhance cancer classification through methylation data. Li *et al.* [25] proposed the Adversarial Variational Bayes AutoEncoder - Multi-Omics Dual-net Feature importance Ranking (AVBAE-MODFR) as HFS method, effectively enhancing classification performance by evaluating the importance of multi-omics features. Wang *et al.* merges the heuristic search method and FS method to identify suitable subsets for classification and detection of biomarkers [26]. Mahto *et al.* [27] proposed a Cuckoo Search and Spider Monkey Optimization (CSSMO)-based FS technique that effectively combines spider monkey optimization and cuckoo search algorithms and performs cancer classification. Zhou *et al.* [28] used the LSTM algorithm for the prediction of transcriptomic gene expression data. Bhonde *et al.* [17] proposed a method that utilizes BiLSTM–CNN for the classification of cancer using a gene expression dataset, with an accuracy of 96.89%. In 2024, Chai *et al.* created a novel contrastive-learning technique for identifying important cancer subtypes. The results revealed that the proposed strategy effectively grouped cancer subtypes and identified 14 important genes associated with cancer, 12 of which (85.7) were confirmed by a study of the literature [29]. The existing study combined the deep residual network and VGG-16 architecture to create the Ensemble Residual-VGG-16 model for cancer diagnosis from mammography images [30]. Huang *et al.* [31] proposed a differential sparse canonical correlation analysis network for classifying breast cancer subtypes. Ren *et al.* [32] proposed a multi-view graph neural network (MVGNN) for the classification of breast cancer using mRNA and DNA methylation datasets, while Chi-square and mRMR FS techniques were used to identify optimal features. The model MVGNN obtained a classification accuracy of 0.906. A method developed by Mohamed and Ezugwu [33] based on PCA–Synthetic Minority Oversampling Technique (SMOTE)–CNN for RNA-Seq, miRNA, and DNA methylation datasets of the TCGA-Lung Adenocarcinoma (LUAD) project, achieved an accuracy of 0.97. Lan *et al*. [34] proposed a DL-based model DeepKEGG for cancer recurrence prediction and biomarker detection through the integration of mRNA-pathway, Simple Nucleotide Variation (SNV)-pathway, and miRNA-pathway multi-omics datasets of TCGA-Breast Invasive Carcinoma (BRCA), LIHC, Prostate Adenocarcinoma (PRAD), Bladder Urothelial Carcinoma (BLCA) projects with the highest accuracy achieved of 0.961. Divate *et al*. [35] proposed a DNN algorithm for classifying Pan-cancer gene expression data and obtained an accuracy of 0.97. Using the TCGA-BRCA dataset (mRNA, miRNA, gene mutations, DNA methylation, and MRI images), Yang *et al*. [36] creatively optimized the neural network topology of a DL model using Bayesian optimization, achieving 0.91 accuracy.

The existing study revealed that the study was not conducted for the Exon expression omics dataset, while the proposed study is based on the Exon expression dataset along with mRNA, miRNA, and DNA methylation omics datasets. The proposed method has robustness for the different individual sample datasets with improved performance for all datasets of NGS and microarray platforms. This integrated approach of HFS-hybrid DRNN is especially suited for high-dimensional datasets like multi-layer omics data, where the combination of HFS- and DL-based models enables precise and scalable classification. The proposed method has enough parameters with reduced memory space, as compared to the basic CNN, transfer learning, and graph neural network, which are computationally costly because of their large parameter sizes.

Even though these techniques have been effective in classifying cancer samples through omics data, they still have disadvantages such as (i) While features are an essential component of performance in multi-omics data; existing methods do not focus on FS. (ii) The underlying correlations between samples and features are typically ignored by some multi-omics data integration techniques, which basically merge multi-omics data as input data.

Materials and methods

In this proposed study, HFS techniques are initially applied to identify important features from multi-layer omics data relevant to SC over normal tissue. Thereafter, we proposed three HDRNN models for the classification of SC using multi-layer omics datasets of Exon expression, mRNA expression, miRNA expression, and DNA methylation. The proposed methodology consists of five different parts. In part 1, we retrieved Exon, mRNA, and miRNA expression, and DNA methylation datasets of SC and adjacent normal tissue. We preprocessed the raw data to remove noise, normalized the data, and identified statistical significance features using the Linear Models for Microarray Analysis or LIMMA package. Part 2 projected four HFS techniques to find relevant features from the multi-layer omics data of SC. Subsequently, part 3 conducted classifications of SC over normal tissue samples with the existing DL-based models, and part 4 projected the proposed HDRNN models. The part 5 included the proposed model validation with eight external datasets of the multi-layer omics profile. Finally, part 6 provided an extensive bioinformatics analysis to validate obtained important top features as molecular signatures for SC from the mentioned multi-layer omics data. Figure 1 represents the overall process flow diagram of the proposed methodology.

Dataset

In the present study, we have used four different layers of omics profile data from TCGA-STAD projects such as Exon, mRNA, miRNA expression, and DNA methylation data of primary tumor of SC over adjacent normal solid tissue. The mRNA expression and DNA methylation datasets were retrieved in 2024 from the TCGA-GDC portal (https://portal.gdc.cancer.gov/) and Exon, and miRNA expression datasets were retrieved in 2024 from the UCSC Xena browser (https://xenabrowser.net/datapages/) [37]. The Exon expression profile was restrained through the Illumina HiSeq 2000 RNA Sequencing platform with Reads Per Kilobase of exon model per Million mapped reads values. DNA methylation or cytosine–guanine dinucleotide (CpG) profile was measured experimentally using the Illumina Infinium HumanMethylation450 with beta value measured unit. The miRNA mature strand expression is measured by RNA-Seq (IlluminaHiseq) technology. The details of the dataset such as several SC samples and normal samples along with several features that were carried out for this study are shown in Table 1.

Dataset preprocessing

Preprocessing of the dataset is a vital step for ensuring accurate model execution [38]. Here, we eliminated low-expressed features and applied KNN imputation to the CpG methylation data to address missing values. After that, we removed low-expressed features of four multi-layer omics data which have low expression values with a threshold (zero_counts_per_row>60) using Python language. Subsequently, we applied min–max normalization to scale all expression values to a standard range between 0 and 1, ensuring that the relationships between the original values were preserved [38]. The normalization datasets were taken as input for statistical significance or differential analysis using the LIMMA statistical R environment package [39] and it measured the significance of features with *P*-value, fold changes, and adjusted *P*-value. LIMMA is suitable for the proposed work due to its ability to analyze datasets containing microarray and bulk RNA-Seq data, ensuring consistent analysis across sequencing platforms. LIMMA is a popular and computationally efficient statistical framework that was created mainly for differential expression analysis of high-dimensional microarray data. However, with the voom function, which modifies LIMMA for count-based RNA-Seq data. Here, we identified different numbers of significant features with a threshold of the adjusted *P*-value (Padj) <.001, using the Benjamini–Hochberg (BH) correction threshold. The statistically significant features obtained from LIMMA of all four multi-layer omics data were used for further analysis.

Dataset balancing

The imbalance of the primary tumor of SC and normal solid tissue samples mentioned in Table 1 was resolved using the hybrid SMOTE Tomek link resampling method. We generated the synthetic normal solid tissue sample data with the SMOTE Tomek link algorithm to balance the specimens with primary tumor SC sample data for the classification task. The potential risk of synthetic data uses such that (i) the data balancing technique creates synthetic samples by incorporating preexisting minority class samples. If there are noise, outliers, or sparse points in the actual data, the generated samples might not fairly represent the distribution; (ii) the approach might give samples in the majority class region when class boundaries overlap, which could confound the model; (iii) additionally, synthetic data can accentuate noise or create false clusters, which can lead to overfitting, particularly in small datasets. On the other hand, dataset imbalance creates model training bias, while the model ignores the minority class and becomes unduly focused on the dominant class.

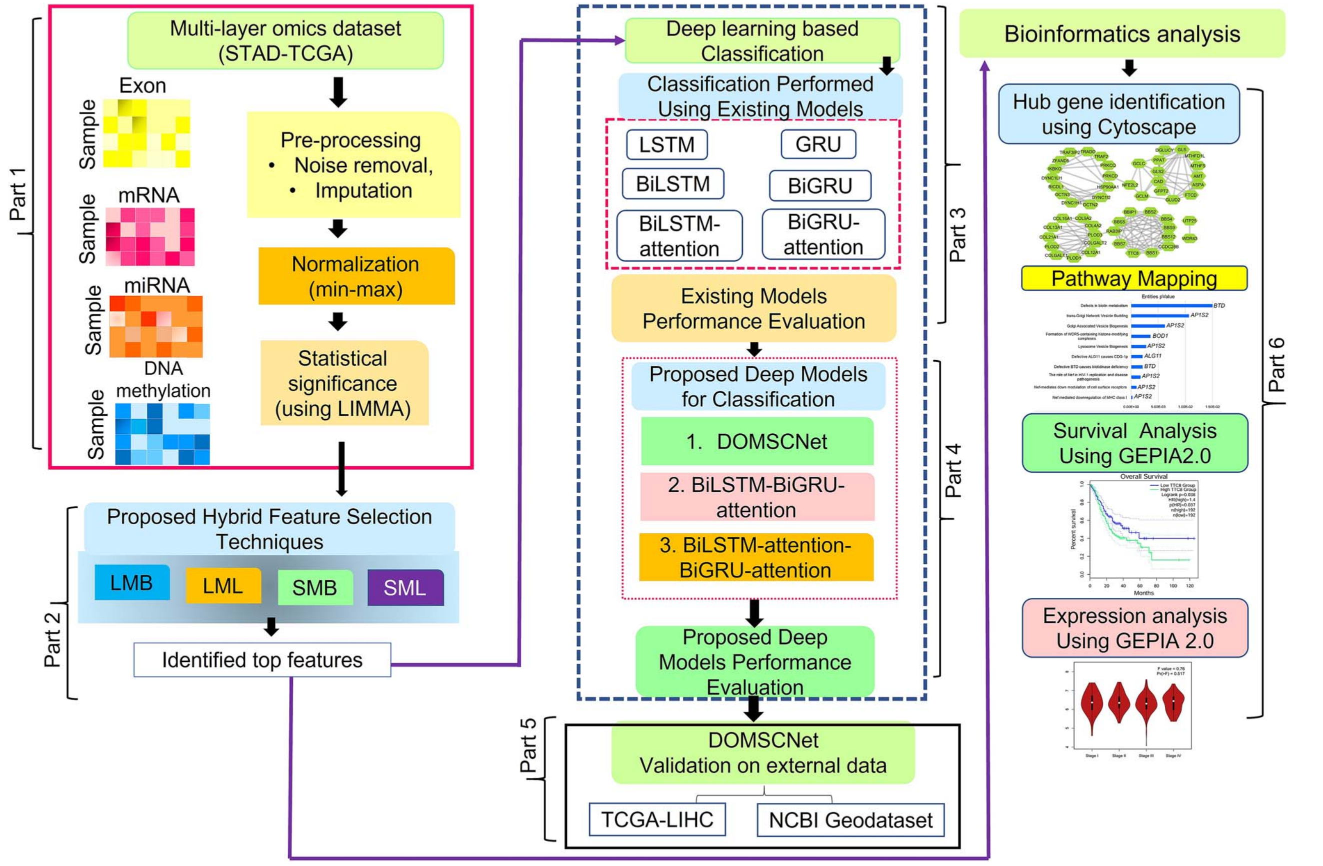
Part 1
Multi-layer omics dataset (STAD-TCGA)
Exon
Sample
mRNA
Sample
miRNA
Sample
DNA methylation
Sample
Pre-processing
• Noise removal,
• Imputation
Normalization (min-max)
Statistical significance (using LIMMA)
Part 2
Proposed Hybrid Feature Selection Techniques
LMB
LML
SMB
SML
Identified top features
Deep learning based Classification
Classification Performed Using Existing Models
LSTM
GRU
BiLSTM
BiGRU
BiLSTM-attention
BiGRU-attention
Existing Models Performance Evaluation
Part 3
Proposed Deep Models for Classification
1. DOMSCNet
2. BiLSTM-BiGRU-attention
3. BiLSTM-attention-BiGRU-attention
Proposed Deep Models Performance Evaluation
Part 4
Part 5
DOMSCNet Validation on external data
TCGA-LIHC
NCBI Geodataset
Bioinformatics analysis
Hub gene identification using Cytoscape
Pathway Mapping
Survival Analysis Using GEPIA2.0
Overall Survival
Percent survival
Months
Expression analysis Using GEPIA 2.0
Part 6

Figure 1. Overall process flow diagram of the proposed study.

Table 1. The details of the used multi-layer omics dataset

Dataset	SC sample	Normal sample	Number of features	Platform
Exon expression	415	35	2 39 322	NGS
mRNA expression	359	29	60 659	NGS
miRNA expression	438	39	1881	NGS
DNA methylation	113	25	25 975	Microarray

Initially, we experimented without balancing the dataset but it showed underfitting results. However, by addressing the limitations of both oversampling and undersampling, the SMOTE Tomek hybrid technique struck a balance. Tomek Links ensures a cleaner dataset by eliminating noise and overlapping occurrences, while SMOTE boosts minority class representation. The synthetic dataset of normal samples generated by the SMOTE Tomek algorithm is depicted in Table 2.

Proposed hybrid feature selection techniques

The FS technique is an effective technique for diminishing the dimensionality of extensive, intricate NGS data [5, 40]. HFS is essential for high-dimensional data as it combines the strengths of various FS techniques, addressing the challenges presented by large and complex datasets. HFS significantly reduces the issue of high dimensionality efficiently by decreasing the feature space.

In the present work, the dimensionality of NGS data was decreased to identify complicated patterns or features that enhance model performance and facilitate the identification of possible molecular signatures. FS techniques are categorized as unsupervised, supervised, and semi-supervised with filter, wrapper, and embedded fashion [5, 6]. Here, we proposed four HFS techniques with a combination of two-fold supervised filter techniques with wrapper and embedded techniques. Hybrid approaches can initially eliminate irrelevant features using a filter method and subsequently evaluate feature interactions through a wrapper or embedded method. Finally, HFS techniques confirm the selection of important, non-redundant features, enhancing both algorithm clarity and overall performance [5].

The first strategy of the FS method combined the Laplacian score (LS) [41], maximum relevancy minimum redundancy (MRMR) [42], and Boruta [43]. The second strategy combined LS, MRMR, and LASSO [44]. Subsequently, SelectKBest (SKB), MRMR, and Boruta were all integrated into the third FS method. The fourth strategy is FS methods, a combination of SKB, MRMR, and LASSO. The final combination of proposed HFS techniques is LMB (LS–MRMR–Boruta), LML (LS–MRMR–LASSO), SMB (SKB–MRMR–Boruta), and SML (SKB–MRMR–LASSO). These four HFS techniques are used to identify relevant features from statistically significant multi-layer omics data. We started by using the LS as a filter technique to reduce the initial number of features. After the initial filtering, we applied MRMR as another filtering method to refine and further select the most relevant features. Lastly, we used the Boruta algorithm, a wrapper technique, and the LASSO embedded

Table 2. The details of the used multi-layer omics dataset after the SMOTE Tomek algorithm were applied

TCGA-STAD dataset	SC sample	Normal sample	Number of features
Exon expression	415	415	2 39 322
mRNA expression	359	359	60 659
miRNA expression	438	438	1881
DNA methylation	113	113	25 975

technique to identify the features most effective at distinguishing between normal and SC classes. In a similar process for SMB and SML FS, we replaced LS with the SKB method as the initial filter technique. This allowed us to reduce features before further refining and selecting key features.

LS reduces dimensionality while preserving essential information, resulting in models that are both efficient and interpretable [5, 45]. The advantage of MRMR is that it helps to simplify model training and interpretation while increasing computational efficiency by lowering dimensionality from high-dimensional input [6, 43]. SKB is an upfront and speedy FS technique by reduces dimensionality from the data [46]. Here, SKB first identified the top K features and calculated the score of features with the F-value. Boruta is an effective wrapper algorithm for identifying relevant features in large datasets and producing accurate results [3]. LASSO improves both accuracy and model interpretability by identifying the most relevant subset of features [44].

Classification model long short-term memory, grated recurrent unit, and their extensions

Nowadays, DNNs have established exceptional performance in the biomedical fields, particularly in cancer classification, and early detection. In this present study, we aim to classify SC and adjacent normal solid tissue using LSTM [47] and GRU [48] with their extension models such as BiLSTM, BiGRU, etc.

In this phase of our analysis, we restructured the expression count matrix data for Exons, mRNA, miRNA, and methylation into a time series sequence format. The count matrix data are in a 1 × 1 dimensional matrix while the input data is converted into a 3 × 3 dimensional matrix (samples, time step, features). This transformation involved sequentially organizing the data, to capture temporal relationships and patterns inherent in the biological processes. By converting the data into time series sequences, we aimed to leverage models that excel in handling sequential data, thereby enhancing the overall performance and predictive capabilities of our analysis. At first, we observed the performance of existing LSTM, GRU, BiLSTM, BiGRU, BiLSTM-attention, and BiGRU-attention mechanisms in classifying SC over normal tissue samples.

Proposed models

After the performance of the existing model, we proposed three different HDRNN-based models such as (i) BiLSTM with BiGRU (DOMSCNet/Model1), (ii) BiLSTM with BiLSTM with attention (Model2), and (iii) BiLSTM-attention with BiGRU-attention (Model3). We perceived the performance of three proposed models on the best-selected feature subset obtained from the best FS technique with the existing model. Eventually, all three proposed models were implemented using the optimal features selected from the best HFS technique. The automation of transformers and other deep advanced models has recently increased their significance in the biological domains. Because of experimental constraints, omics databases frequently contain a limited number of samples. In contrast to advanced deep models, which usually require huge datasets to prevent overfitting, while DOMSCNet requires fewer samples as compared to transformers to obtain improved performance. For biological data, deep models can hinder the understanding of underlying mechanisms. In consideration of these challenges, we presented proposed methods for the classification of cancer samples in comparison to healthy human samples.

Bidirectional LSTM–BiGRU/DOMSCNet/Model1

BiLSTM and BiGRU are RNN architectures that are particularly effective for sequences in both forward and backward directions and for capturing information from both past and future contexts [28, 49]. Here, it is applied to classify SC over normal tissue samples from omics expression data due to its ability to capture sequential patterns in complex high-dimensional data. The proposed model DOMSCNet integrated BiLSTM and BiGRU layers into the classification of SC. The architecture of DOMSCNet included a BiLSTM layer (512 units). The output from the final BiLSTM layer is fed to the BiGRU layer (512 units). A dense (fully connected) layer (64 units) and Rectified Linear Unit (ReLU) activation function are applied. A dropout layer is added with a rate of 0.25, which randomly drops 25% of the neurons during training to prevent overfitting. Batch normalization is applied, which normalizes the activation to improve training stability and speed. The final output layer is a dense layer with a sigmoid activation function, making it suitable for binary classification. Figure 2 represents the architecture of the proposed model DOMSCNet. The algorithm of BiLSTM–BiGRU is defined as Equations (1)–(6) in Algorithm 1.

The output for each time step is h_t, combining both forward and backward contextual information. At time t, the hidden state of the BiLSTM model is obtained by weighted summation of the backward hidden layer state h_t^f and forward hidden layer state h_t^b as shown in Equation (6). The parameters applied in the proposed DOMSCNet architecture are tabulated in Table 2.

BiLSTM–BiGRU-attention/Model2

The proposed Model2 integrated BiLSTM with BiGRU and attention mechanism to classify SC and normal tissue samples. Figure 2b depicts the proposed Model2 architecture which includes a BiLSTM layer (512 units), batch normalization is applied to the BiLSTM output to normalize the activations, a dense layer (64 units), and ReLU activation is applied to the output from the BiLSTM model. The output from the final BiLSTM layer is fed to the BiGRU layer. Similarly, again BiGRU layer was added with the same parameters, viz 512 neuron units, and Batch normalization, and a second dense layer (64 units) with ReLU activation function. The output from the final BiGRU layer is fed to the attention module. After that, an attention layer is added and next a custom layer of the ‘Lambda’ function is added to reduce the attention output (context vector) by summing across time steps of the features. The input of the attention mechanism layer is obtained from the

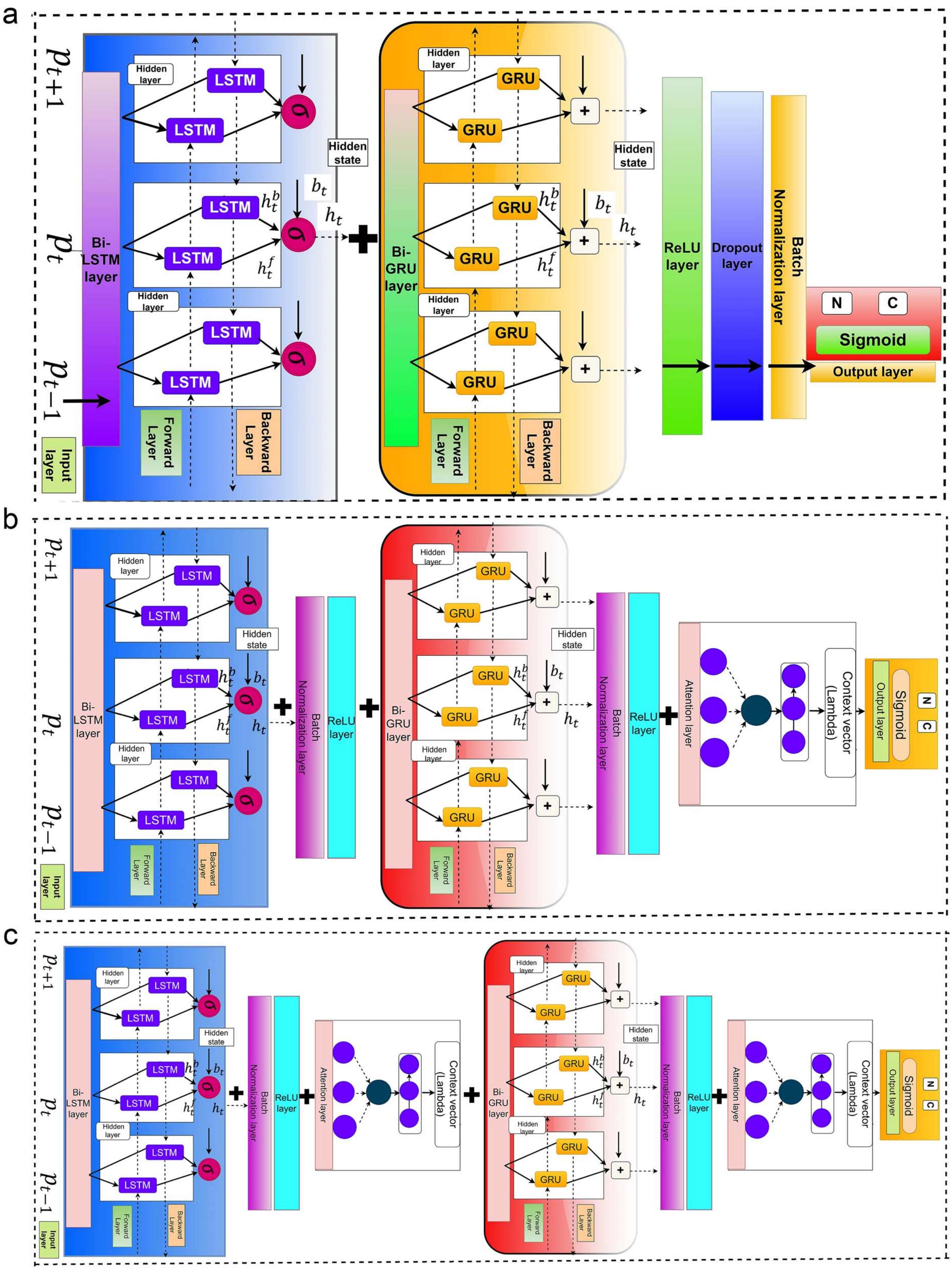

Figure 2. (a) Overall process flow diagram of DOMSCNet. (b) The architecture of Model2. (c) The architecture of Model3.

output vector of the top layer triggered by the BiLSTM–BiGRU network. The final output layer is a fully connected dense layer with a 'sigmoid' activation function, as shown in Fig. 2b. The parameter used in DOMSCNet are tabulated in Table 3.

Attention mechanism

The major intention of the attention mechanism applied in the BiLSTM–BiGRU model for the classification task is to utilize the hidden layer output of the input sequence by the BiLSTM–BiGRU

Algorithm 1. Algorithm of proposed model BiLSTM–BiGRU.

BiLSTM and BiGRU	
Input	$P = \{p_{1,}p_{2,\ldots\ldots\ldots}p_T\}$
Initialize	Hidden states h_0^f, h_T^b (for backward and forward direction), cell states C_0^f, C_T^b initialize to 0.
Forward pass for BiLSTM	For each time step t **from** 1 **to** T, compute the forward LSTM states: $f_t^f = \sigma\left(W_f^f p_t + U_f^f h_{t-1}^f\right)$ $i_t^f = \sigma\left(W_i^f p_t + U_i^f h_{t-1}^f\right)$ $O_t^f = \sigma\left(W_o^f p + U_o^f h_{t-1}^f\right)$ $\overset{\sim f}{C_t} = \mathbf{tanh}\left(W_C^f p_t + U_C^f h_{t-1}^f\right)$ $C_t^f = f_t^f \odot C_{t-1}^f + i_t^f \odot \overset{\sim f}{C_t}$ $h_t^f = \mathbf{tanh}\left(C_t^f\right)$..(1)
Backward pass for BiLSTM	For each time step t **from** T **to** 1, compute the backward LSTM states: $f_t^b = \sigma\left(W_f^b p_t + U_f^b h_{t+1}^b\right)$ $i_t^b = \sigma\left(W_i^b p_t + U_i^b h_{t-1}^b\right)$ $O_t^b = \sigma\left(W_o^b p_t + U_o^b h_{t-1}^b\right)$ $\overset{\sim b}{C_t} = \mathbf{tanh}\left(W_C^b p_t + U_C^b h_{t-1}^b\right)$ $C_t^b = f_t^b \odot C_{t-1}^b + i_t^b \odot \overset{\sim b}{C_t}$ $h_t^b = O_t^b \odot \mathbf{tanh}\left(C_t^b\right)$..(2)
Hidden state	$h_t = \left[h_t^f, h_t^b\right]$...(3)
Forward pass for BiGRU	For each time step t **from** 1 **to** T, compute the forward BiGRU hidden state: $Z_t^f = \sigma\left(W_z^f p_t + U_z^f h_{t-1}^f\right)$ $r_t^f = \sigma\left(W_r^f p_t + U_r^f h_{t-1}^f\right)$ $\tilde{h}_t^f = \mathbf{tanh}\left(W_h^f p_t + U_h^f\left(r_t^f \odot h_{t-1}^f\right)\right)$ $h_t^f = \left(1 - z_t^f\right) \odot h_{t-1}^f + z_t^f \odot$ $\tilde{h}_t^f$...(4)
Backward pass for BiGRU	For each time step t **from** T **to** 1, compute the backward BiGRU states: $Z_t^b = \sigma\left(W_z^b p_t + U_z^b h_{t+1}^b\right)$ $r_t^b = \sigma\left(W_r^b p_t + U_r^b h_{t+1}^b\right)$ $\tilde{h}_t^b = \mathbf{tanh}\left(W_h^b p_t + U_h^b\left(r_t^b \odot h_{t-1}^b\right)\right)$ $h_t^b = \left(1 - z_t^b\right) \odot h_{t+1}^b + z_t^b \odot \tilde{h}_t^b$..(5)
Hidden state	$h_t = \left[h_t^f, h_t^b\right]$...(6)

encoder. Integrating attention mechanisms with BiLSTM and BiGRU models can significantly enhance their performance by allowing the model to focus on important input features of the sequence when making the classification [48, 49]. By focusing on key features rather than processing all features equally, the attention mechanism helps the model generalize better on new data. This reduces the likelihood of overfitting to irrelevant omics data expression patterns. Understanding which features are most influential in the classification decision is essential in biomedical research. The attention mechanism highlights important features or feature sets, allowing researchers to interpret and validate which features are critical for the outcome. The attention mechanism is computed as follows:

$$e_a = \mathbf{tanh}\left(W_h j_a + b_h\right), e_a \in [-1, 1]$$

The attention weight is calculated as,

$$k_a = \frac{\mathbf{exp}\, e_a}{\sum_{t=1}^{N} \mathbf{exp}\left(e_t\right)}$$

The sum of the weight is calculated as,

$$\sum_{i=1}^{N} k_a = 1$$

The context vector is calculated as,

$$r = \sum_{i=1}^{N} k_a j_a, v \in R^{2L} \tag{7}$$

We included an attention layer to assess the importance of each feature in the context of the entire sequence. The attention mechanism assigns a weight k_a to each feature j_a, highlighting the most pertinent elements. Subsequently, the hidden states are combined to create a sentence feature vector v through a weighted sum [48]. The parameters used in Model2 during training are tabulated in Supplementary Table S2.

Table 3. The parameters used in the proposed model DOMSCNet

Layers	Output shape	Units	Activation function
Input	(None, features, 1)	—	—
BiLSTM	(None, features, 512)	512	—
Dense	(None, features, 64)	64	ReLU
BiGRU	(None, features, 512)	512	—
Dense	(None, 64)	64	ReLU
Dropout (0.25)	(None, 64)	64	—
Batch Normalization	(None, 64)	64	—
Output	(None, 1)	—	Sigmoid

Bidirectional LSTM-attention–BiGRU-attention/Model3

The architecture of Model3 is represented in Fig. 2c. Here, we added two attention layers after both the BiLSTM and BiGRU layers. The proposed Model3 consists of a BiLSTM layer (512 units). The output from the final BiLSTM layer is fed to the attention module. Batch normalization is applied to the BiLSTM output to standardize the activation, followed by a dense layer (64 units) along with the ReLU activation function. After the ReLU layer, an attention layer is added, and a custom 'Lambda' function layer is applied to reduce the attention output (context vector) by summing across the time steps of the features. Similarly, a BiGRU layer is added next to the attention layer 1 with 512 units, batch normalization, and a dense layer (64 units), and ReLU activation is included. The output from the final BiGRU layer is fed to the attention module. An attention layer with a custom 'Lambda' function is again employed to sum the attention output across the time steps. The final output layer is a fully connected dense layer with a sigmoid activation function. The 'softwares' optimizer was used for the proposed three models during training. Supplementary Table S3 depicts the parameters of Model3 architecture. The algorithm of BiLSTM, BiGRU, and attention mechanism is represented in Equations (1)–(6) in Algorithm 1 and Equation (7), respectively.

Validation study of DOMSCNet on external datasets

For validation of the proposed model, datasets were retrieved from NCBI Geodataset (https://www.ncbi.nlm.nih.gov/geo/). From Geodataset, we collected Gene or mRNA expression SC and normal samples dataset of (Accession id: GSE36968), DNA methylation dataset of SC and normal (Accession id: GSE30601), miRNA expression dataset of Lynch syndrome with cancer, rectal cancer, and healthy samples (Accession id: GSE198834), DNA methylation dataset of Uterine cervix cancer and normal (Accession id: GSE30760). Similarly, the TCGA-LIHC dataset of the primary tumor and adjacent normal tissue samples of the liver cancer was retrieved from the UCSC Xena portal. From TCGA-LIHC, mRNA or Gene expression, miRNA expression, Exon expression, and DNA methylation four multi-layer omics datasets were collected. Similarly, min–max normalization was used to normalize all eight datasets, while statistical significance analysis of features was executed with the LIMMA R package. The best HFS technique was applied to identify optimal features for the proposed model training as well as testing purposes. To handle the dataset imbalance issue, we used the SMOTE Tomek hybrid algorithm. The details of the external validation dataset are depicted in Supplementary Table S1.

Bioinformatics analysis

From the HFS experimental results, we shortlisted top genes from the obtained feature subset and conducted different analysis of shortlisted genes to understand their biological functions in SC progression and detection. The Exons, mRNA, miRNA expression, and DNA methylation site profiles are all integral components of gene regulation and play vital roles in controlling gene expression. The disease pathway analysis of selected top genes was performed using the publicly available Reactome database with a *P*-value <.05 [50]. The gene–protein interaction study was performed using the STRING database (https://string-db.org/) to identify hub genes of top genes [51]. The retrieved protein–gene interaction network was visualized using Cytoscape software and the top 15 highest degree scores gene network was extracted using CytoHubba plugins.

The expression and survival analysis of top genes were employed for the identification of prognostic signatures of SC. Here, we performed gene expression analysis between normal and primary tumor samples, pathological disease stage-wise expression analysis, and overall disease survival analysis of low-group and high-group expression conditions using the GEPIA 2.0 (Gene Expression Profiling Interactive Analysis) web tool (http://gepia2.cancer-pku.cn/#index) [52]. For GEPIA 2.0 analysis, we considered only overlapped genes between the hub gene interaction study and top-selected genes. Initially, GEPIA 2.0 was employed to analyze the selected gene expression in normal and SC tissue samples from the TCGA project. Following this, we observed how selected gene expression increases or decreases in different pathological stages. The statistical significance of the variations was assessed using an ANOVA test with *P*-value threshold <.05 in GEPIA 2.0. Additionally, Kaplan–Meier curve plots were utilized to evaluate survival rates based on lifespan data, providing insights into how genes influence cellular structure and molecular function.

Software and libraries

Python (version 3.11) and R (version 4.1.3) languages were utilized in this study to carry out the experiment and analyze the findings. Utilizing the Tensorflow Keras API (version 3.5.0) on the Google Colab platform, the Tomek link library is utilized for data balancing. LS, LASSO (Lasso, LassoCV, KFold, GridserachCV library), and SKB-FS (f_classif, selectkbest library) were used during FS. The proposed model performance used Numpy, Pandas, matplotlib, Sequential, LSTM, GRU, Bidirectional, dropout, batch normalization, dense, Model, classification_report, Input, tf, layers, Lambda, attention, etc. The LIMMA package (version 3.50.3) was utilized in the R environment of the local machine (Intel Core i3, 7th Gen, Windows) for statistical significance analysis. The Boruta-FS

Table 4. Assessment of performance matrices

Matrices	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1-score	$\frac{(1+\alpha^2)\text{precision}*\text{recall}}{(\alpha^2)\text{precision}*\text{recall}}$, where $\alpha = 1$

performed FS of all 12 datasets using the Boruta package (version 8.0.0) and the MRMR-FS used the mRMRe package (version 2.1.2) at R environment local system (Intel Core i3, 7th Gen, Windows).

Results and discussion

In the execution of the proposed models, we divided the input dataset of Exon, mRNA, and miRNA expression into 70% for training, 10% for validation, and 20% for testing. Additionally, the methylation dataset was split into 70% for training, 15% for validation, and 15% for testing due to its small number of samples. After the training process, we assessed model performance using various metrics including accuracy, precision, recall, F1-score, and confusion matrices for the validation and testing samples for each dataset. We also computed the AUC score and generated ROC plots for the proposed models utilizing the best features. The different matrices are calculated as shown in Table 4.

Results of hybrid feature selection techniques

The relevant feature subset of multi-layer omics data is obtained from HFS techniques. Initially, we identified optimal 2000 features using LS and SKB filter techniques from Exon, mRNA, and DNA CpG methylation data. After that, we identified 1000 important features from MRMR, and finally, using the Boruta and LASSO algorithm, we identified a final optimum number of features, as shown in Table 5. However, the raw feature of miRNA data is less than Exon, mRNA, and DNA methylation data, and we changed the criteria for the selection of features of miRNA data. We obtained 1000 features from LS and SKB. The MRMR identifies 500 features and Boruta and LASSO identified the final optimum features as shown in Table 5.

Existing models performance

The identified features were used for model training. Initially, we evaluated the performance of four existing and models and their extensions: LSTM, GRU, BiLSTM, BiGRU, BiLSTM-attention, and BiGRU-attention, respectively. Among these, LSTM does not produce an improved accuracy score compared to GRU, BiLSTM, BiGRU, BiLSTM-attention, and BiGRU-attention across all identified features. Similarly, GRU performed better specifically for the Exon and mRNA expression datasets. However, neither LSTM nor GRU effectively classified all four multi-layer omics datasets.

Additionally, existing BiLSTM and BiGRU models do not achieve improved accuracy scores in the training, validation, or test data for the DNA methylation dataset, as shown in Fig. 3 and Supplementary Table S4. Similarly, for the mRNA and miRNA expression datasets, the features obtained led to better performance with BiLSTM, BiGRU, BiLSTM-attention, and BiGRU-attention models as compared to the DNA methylation data. Overall, the SMB FS method showed moderate improvements across all datasets with all six classifiers. The performance of the existing models did not improve across all four multi-layer omics datasets. These existing models were unable to accurately classify data for all four omics layers. In contrast, we proposed three HDRNN models, specifically designed for SC classification from normal solid tissue using the four multi-layer omics datasets.

Performance of the proposed models

The experimental results indicate that none of the existing LSTM, GRU, and their extension models achieved strong performance on the DNA methylation data. Consequently, we introduced three HDRNN models namely DOMSCNet, Model2, and Model3 designed to perform consistently across all four multi-layer omics datasets. The FS technique SMB, showed improved results on four datasets with existing models. As a result, we opted to use the SMB HFS technique for further analysis. In our study, we executed the proposed models such as DOMSCNet, Model2, and Model3 on SMB-identified features for multi-layer omics datasets. The accuracy, precision, recall, and F1-score of the proposed (a) DOMSCNet, (b) Model2, and (c) Model3 on the validation and testing dataset are represented in Fig. 4a.

The proposed model DOMSCNet showed the highest accuracy (0.99) and precision (0.995) score on the Exon testing dataset while achieving the highest accuracy (0.99) score of testing data on miRNA expression data. The proposed Model2 obtained the highest precision score result on mRNA expression data with 0.995. Model2 achieved the highest accuracy, recall, and F1-score on mRNA and miRNA expression while it showed the lowest score on DNA methylation data with 0.940 on testing data, as shown in Fig. 4a. Figure 4a also illustrates that Model3 achieved maximum validation accuracy, precision, recall, and F1-score on mRNA expression data. Similarly, the proposed Model3 showed the highest recall and F1-score of 0.995 on the mRNA testing dataset while achieving a validation score on the mRNA expression dataset. Figure 4b showed that DOMSCNet accurately classified SC and normal tissue samples in Exon and miRNA expression testing, and validation datasets. The confusion matrices' results of test and validation data, as shown in Fig. 4c and d for Model2 and Model3, respectively. In Fig. 4b, 4c, and 4d. (a-d) represents confusion matrixes of test data and (e-f) depicts the results of validation data for DOMSCNet, Model2 and Model3 respectively.

The AUC score and ROC curve of validation and testing data for the proposed DOMSCNet, as shown in Fig. 5a. DOMSCNet achieved an improved AUC score for all four-layer omics datasets. The proposed DOMSCNet achieved the highest AUC score on the

Table 5. The results of the feature subset selected from four hybrid FS techniques

FS techniques	Exon	mRNA	miRNA	DNA methylation
LMB	106	98	54	54
LML	88	103	115	46
SMB	203	193	56	187
SML	78	141	112	48

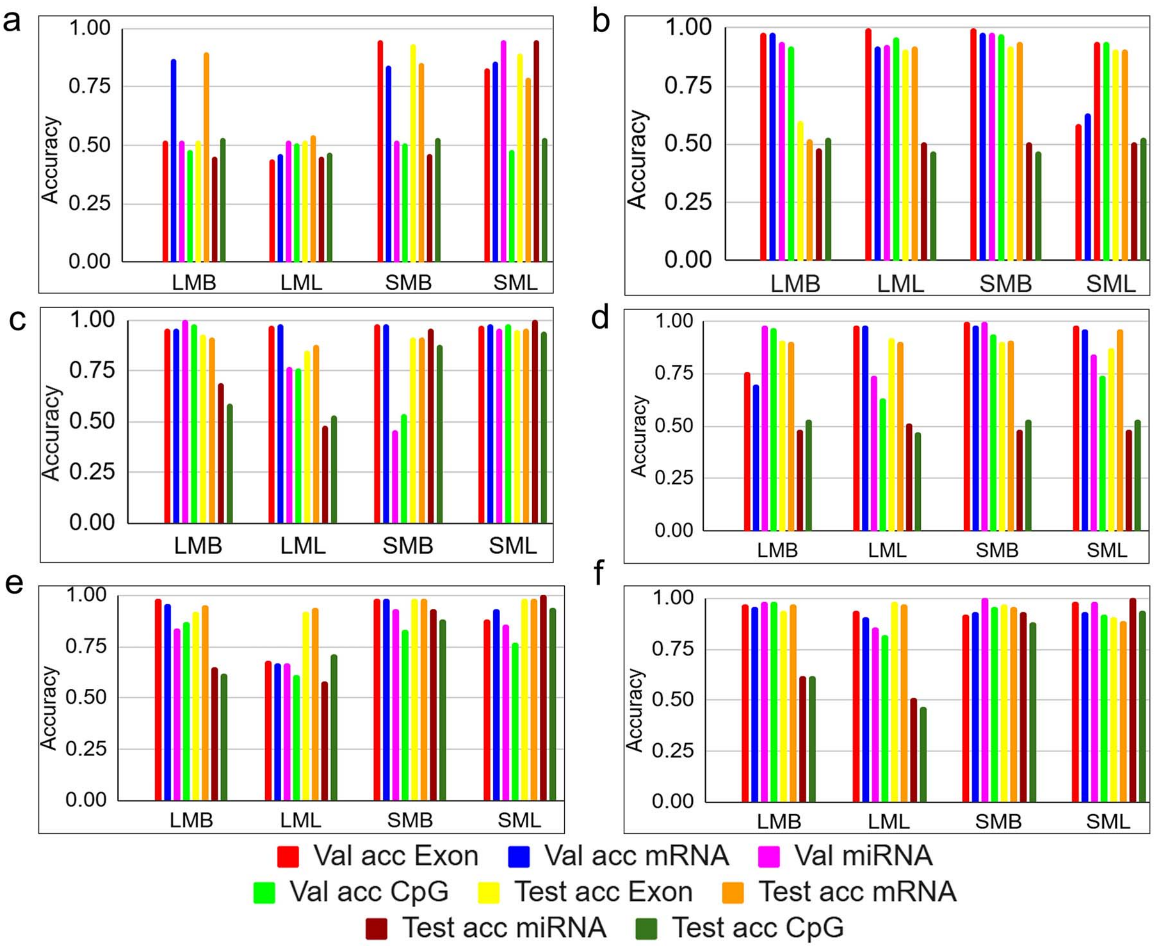

Figure 3. Validation and testing accuracy score of: (a) LSTM, (b) GRU, (c) BiLSTM, (d) BiGRU, (e) BiLSTM-attention, and (f) BiGRU-attention.

Exon, miRNA, and DNA methylation test dataset. The proposed Model2 (Fig. 5b) and Model3 (Fig. 5c) produced an improved AUC score for mRNA datasets than Model1, while Models 2 and 3 obtained poorer results than DOMSCNet for DNA methylation test data.

Table 6. MCDA results of proposed three models

Model	TOPSIS score	Rank
DOMSCNet	1	1
Model2	0.0497	3
Model3	0.0348	2

Multi-criteria decision analysis

Drawing conclusions when multiple objectives are involved can be challenging and prone to inaccuracies [53]. Multi-criteria decision analysis plays a promising role in solving this multi-objective issue by generating conclusions with the best reasonable solution. At present, multi-criteria decision analysis (MCDA) is widely used in performance assessment methods. In the present study, from the evaluation measures of the proposed three models such as accuracy, precision, Recall, F1-score, and AUC score, it is observed that no single model could achieve the best results for all four multi-layer omics data to reach our aim of this study. Thus, selecting an effective classification model for multiple datasets becomes challenging. To find out the best model, we have executed one MCDA technique known as the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). Here, the Information Entropy Weighting technique is applied with TOPSIS to calculate the measures of TOPSIS. Equation (8) represents the TOPSIS evaluation metric. In this study, accuracy, precision, F-score, recall scores, and AUC scores are used as multiple criteria. MCDA ranks based on the TOPSIS technique are shown in Table 6. From Table 6, it is observed that the proposed model DOMSCNet has the top-rank TOPSIS score followed by Model2 and Model3. DOMSCNet performed better for all four datasets while Model2 and Model3 performed poorly for the DNA methylated dataset. The TOPSIS score is calculated as a calculated normalized matrix, each element N_{ij} is calculated as:

$$N_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} {x^2}_{ij}}};$$

for $i = 1, 2, 3 \ldots\ldots\ldots, m; j = 1, 2, 3.\ldots\ldots\ldots, n$

Calculate the weighted normalized matrix, multiply each element N_{ij} by its associated weight w_j,

$$w_{ij} = w_j * N_{ij},$$

Distance calculated on ith row,

$$d^p = \sqrt{\sum_{j=1}^{n} \left(W_{ij} - K_j^{+}\right)^2}; d^n = \sqrt{\sum_{j=1}^{n} \left(W_{ij} - K_j^{-}\right)^2}$$

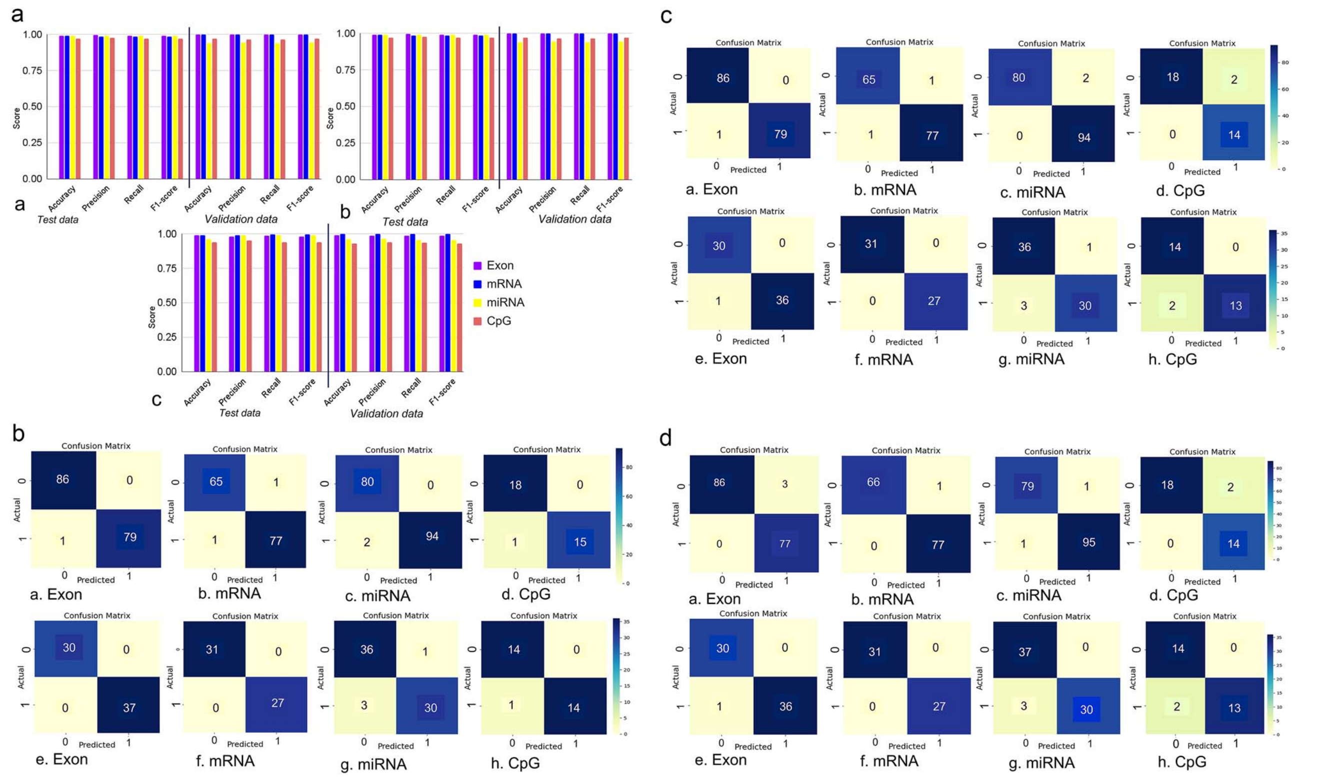

Figure 4. (a) The accuracy, precision, recall, and F1-score and (b-d) represents confusion matrixes of proposed models DOMSCNet, Model2, and Model3 respectively.

Calculate TOPSIS for each row,

$$T_i = \frac{d^n}{d^p + d^n} \tag{8}$$

The major advantage of DOMSCNet can accurately classify multi-layer omics datasets. Additionally, the best HFS technique was applied to identify optimal features for the proposed model training as well as testing purposes. To handle the dataset imbalance issue, we used the SMOTE Tomek hybrid algorithm. The proposed HFS method is fast and precise for identifying SC molecular signatures. The overall results beat the existing model for four types of multi-layer omics (Exon, mRNA, miRNA, and DNA methylation profiles) high-dimensional dataset of stomach adenocarcinomas. The proposed HFS methods validate the function of SC and discover new molecular signatures. One of the drawbacks is that, in comparison to large-size sample datasets, model performance is relatively poor in small sample datasets of DNA methylation of STAD project.

Performance of DOMSCNet on external datasets

To observe the robustness of the DOMSCNet performance, validation was executed on eight additional datasets collected from NCBI Geodatabase and UCSC Xena platform (TCGA, GDC dataset). The eight-dataset split with different train, validation, and test sets due to its varied number of sample sizes present on each multi-layer omics dataset. The mRNA dataset of SC splits with 80% training, 10% testing, and 10% validation set; the miRNA expression dataset splits with 70% training, 15% testing, and 15% validation set; DNA methylation CpG sites dataset of SC and uterine cervix cancer 70% training, 15% validation, and 15% testing. The four multi-layer omics datasets of TCGA-LIHC split with 70% training, 10% validation, and 20% testing set. For all eight multi-layer omics validation datasets used in this study, the proposed straightforward layer architecture model is able to reliably and efficiently classify the samples while maintaining computational efficiency. The obtained feature numbers obtained from SMB HFS are similar to above-mentioned TCGA-STAD dataset. With varying sample sizes across independent data sets, the model does not exhibit overfitting or underfitting. The accuracy, precision, Recall, F1-score, and confusion matrix results of DOMSCNet on an additional eight datasets are shown in Fig. 6a–c and Supplementary Fig. S1a and b. The DOMSCNet achieved the highest results on the TCGA-LIHC four multi-layer dataset. In the TCGA-STAD DNA methylation dataset, the DOMSCNet performance is quite poor (an AUC score of 0.971) compared to the other three datasets due to the presence of the small number of samples. However, in the TCGA-LIHC dataset, DOMSCNet outperforms with an AUC score on Exon 0.994, mRNA 0.987, miRNA 0.988, and DNA methylation 0.993 as shown in Fig. 6a and b and Supplementary Fig. S1a. In Fig. 6a represents the results of DOMSCNet on (A) TCGA-LIHC dataset for testing data, (B) TCGA-LIHC dataset for validation data, and (C) Results of DOMSCNet on testing data and validation set data of Geodataset. Fig. 6b and Fig. 6c depicts the confusion matrices, AUC score, and ROC plots for test set data results of DOMSCNet on multi-layer omics dataset of TCGA-LIHC and Geodataset respectively. In Fig. 6b and Fig. 6c confusion matrices, 0 indicates normal samples and 1 indicates cancer samples. In Geodataset, the DOMSCNet achieved the highest AUC score of 0.978 on the DNA methylation CpG site dataset of cervix cancer, while lowest on the miRNA expression dataset of Lynch syndrome with cancer, rectal cancer 0.871 as shown in Fig. 6a–c

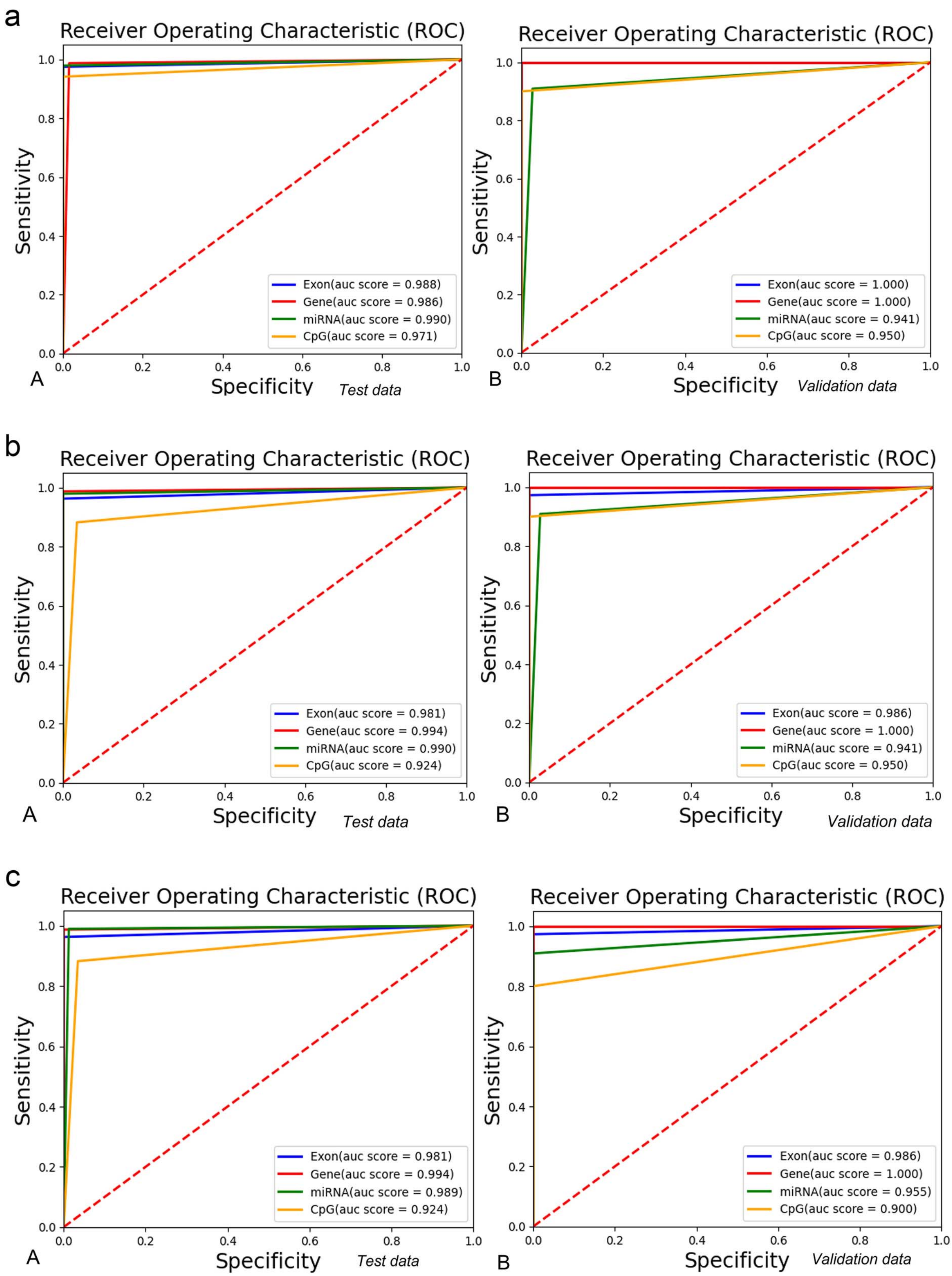

Figure 5. AUC score and ROC plot results of proposed (a) DOMSCNet, (b) Model2, and (c) Model3 respectively.

and Supplementary Fig. S1b. We observed that the performance with Geodataset is poor compared to the TCGA dataset due to the small sample size present in the dataset. Among all Geodataset, the cervix cancer DNA methylation dataset has the highest number of samples and the results of DOMSCNet achieved the highest. However, the DOMSCNet has not shown any overfitting and underfitting results due to its small size samples. The experimental results of DOMSCNet on external dataset signified its robustness for classification of cancer samples over normal using multi-layer omics complex dataset. The comparison

Table 7. The comparison results of DOMSCNet with the existing literature

Technique	Dataset	Results	References
Residual network (Resnet)	TCGA gastric cancer (mRNA, methylation, and copy number variation data)	AUC of 0.971	[50]
Graph neural network (GNN)	TCGA-STAD (mRNA expression, CNV, clinical information, and DNA methylation)	AUC: 0.976 ± 0.007	[51]
DeepKEGG	TCGA-BRCA, LIHC, PRAS, BLCA (mRNA-pathway, SNV-pathway, and miRNA-pathway)	0.876 (BRCA), 0.947 (LIHC), 0.799 (PRAD), 0.961 (BLCA)	[33]
DOMSCNet	TCGA-STAD and LIHC (mRNA, miRNA, Exon expression, and DNA methylation)	0.990 (STAD), 0.994 (LIHC)	Proposed study

results of the DOMSCNet with the existing model are depicted in Table 7. The comparison is based on TCGA-STAD and TCGA-LIHC datasets.

Bioinformatics analysis

Combining Exon, mRNA, and miRNA expression, and DNA methylation data allows a comprehensive understanding of the regulatory mechanisms influencing gene expression. These multi-layer profiles are intricately linked to gene expression regulation and play a pivotal role in identifying potential molecular signatures. Here, we performed different bioinformatics-based analyses of shortlisted top genes obtained from multi-layer omics profiles feature subsets and observed their functions in different cases.

To identify the top genes obtained from feature subsets, we cross-examined the common features between the (i) LMB and SML, (ii) LML and SMB, (iii) SMB, and LMB, and (iv) SML and LML HFS techniques, respectively. We selected 35 shared features from the Exon dataset using these techniques. Similarly, for the mRNA expression data, we identified 63 top genes from the combined feature sets of these four techniques. Additionally, 22 miRNAs were selected as top features, and 27 CpG sites were shortlisted for further analysis. The results of the Reactome pathway analysis, as shown in Fig. 7 and Supplementary Table S5, indicated that *HSP90AA1, COL4A2*, *COL12A1*, and *DYNC1H1* Exon's genes were highly significant and involved in the top ten pathways, including *DDX58/IFIH1*-mediated induction of interferon-alpha/beta, collagen chain trimerization, and autophagy. From the gene expression data, *ACADM and HADH* were found to be significantly associated with several key pathways shown in Fig. 7 and Supplementary Table S5. These associations were identified with a *P*-value of <.001. In contrast, no CpG methylated site genes were associated with the top pathways, as shown in Supplementary Table S5.

Additionally, we performed hub-gene identification for shortlisted top genes. To identify hub genes, we extracted the gene–protein network from the STRING database, with 700 confidences as shown in Fig. 8a and b. Figure 8c shows the results of the hub genes obtained from Exon's data. The network nodes are represented for genes, and edges are represented for connection between genes. The hub gene identification showed that *TTC8*, *COL12A1*, *COL4A2*, *HSP90AA1*, *DYNC1H1*, and *CAD* are the overlapped Exon genes with the top 35 genes of Exons as shown in Fig. 8c and marked in green contour. Figure 8c signified the results of hub gene analysis of mRNA expression for the top genes and there are seven genes found to be overlapped with identified 63 top genes. The overlapped mRNA genes are marked in yellow contour. The hub-gene network analysis of DNA methylation data is shown in Fig. 8c and overlapped genes namely *KCNMA1*, *FUT8*, and *MGAT5* between top genes and hub genes are marked in red contour. miRNA regulates gene expression posttranscriptionally by targeting specific mRNAs. The identified top miRNA interpretation is not directly available in bioinformatics tools. Therefore, we retrieved miRNA-targeted mRNA genes from the TargetScan human database and observed their functions in SC progression. The miRNA-targeted mRNA genes network is shown in Fig. 8c and only one gene overlapped in the gene–protein network. Similarly, Fig. 8c shows the hub gene network with overlapped CpG site genes in red color nodes. There are a total of 14 hub genes found in the network.

The overall survival analysis and pathological stage-wise expression analysis were performed for overlapped genes (marked in green, yellow, dark green, and red contour, respectively) extracted from the hub gene network. This interpretation was executed to see how these genes serve for progression in SC samples. The overall survival analysis of genes *COL4A2*, *COL12A1*, *TTC8*, *GPX3*, and *CMTM5* was significantly associated with SC, as shown in Fig. 9a. These four genes statistically impact patients' overall survival outcomes and are related to poor survival results in SC. The existing study [54] revealed the potential of the tumor-promoting function of hypoxic Cancer-associated fibroblasts (CAFs) in SC could be linked to the decreased expression of gene *COL4A2* in a hypoxic environment. Through multivariate Cox regression analysis, the study identified *COL4A2* as a signature and independent prognostic biomarker in SC patients, with its high expression correlating with a poorer prognosis in those treated by surgery [55]. Duan *et al.* identified *COL12A1* as a prognostic biomarker of SC due to its significantly poor prognostic factors [56]. The study [57] found that *GPX3* levels in SC were favorably connected with immune cell markers, immune cell filtration, and immune checkpoint expression. While gene *GPX3* may play a role in preventing the development of tumors during the malignant progression from gastritis to SC and later stages [58]. Another study's experimental results predicted lower tumor growth time in individuals older than 60 years and showed *GPX3* hypermethylation in SC [58]. Poorer overall survival was strongly associated with the gene *CMTM5*, suggesting that *CMTM5* expression may be a new prognostic factor for SC patients [59]. Till date, there were no studies have been reported linking the gene *TTC8* to SC.

Similarly, genes *HSP90AA1*, *COL4A2*, *TTC8*, *HADH*, *MGAT5*, and *KCNMA1* showed significant results of pathological stage-wise expression of SC as shown in Fig. 9b. These gene expressions are changes between stage 1 and stage 4 for individuals suffering from SC. However, there are no miRNA-targeted mRNA genes correlated with the progression and survival of SC patient samples. *HSP90AA1*, its potential as a detective marker and target for liver

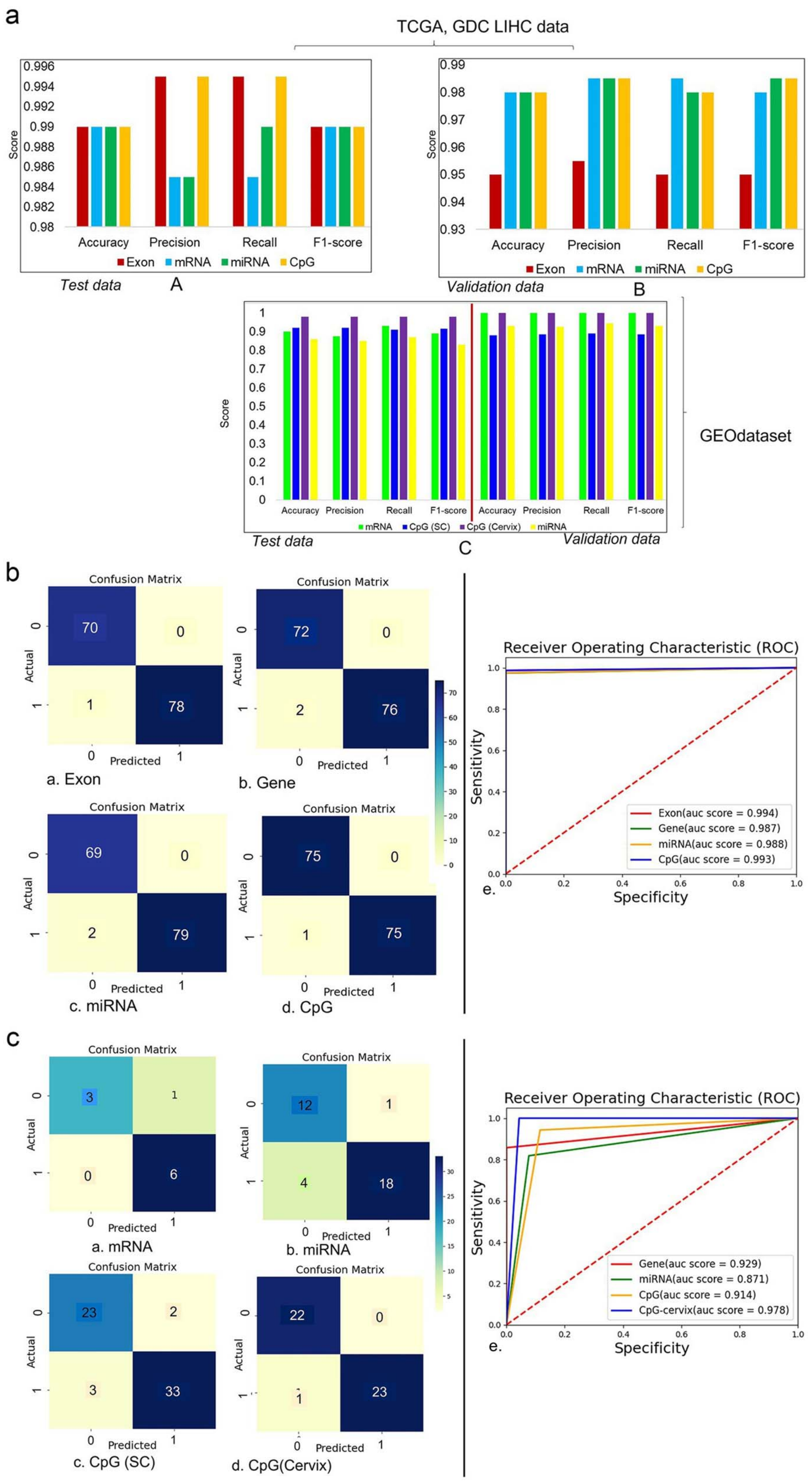

Figure 6. The accuracy, precision, recall, F1-score, confusion matrix results, AUC score and ROC plots of DOMSCNet on additional datasets.

metastasis of SC was highlighted by the significant upregulation of *HSP90AA1* in both transcriptional and translational levels in advanced SC associated with significant SC and insignificant SC related to typical mucosa [60]. The gene *HADH* expression is significantly transformed in SC, influencing tumor development and prognosis, with the study highlighting that monitoring its expression could offer valuable insights into the metabolic reprogramming associated with gastric tumors and their clinical

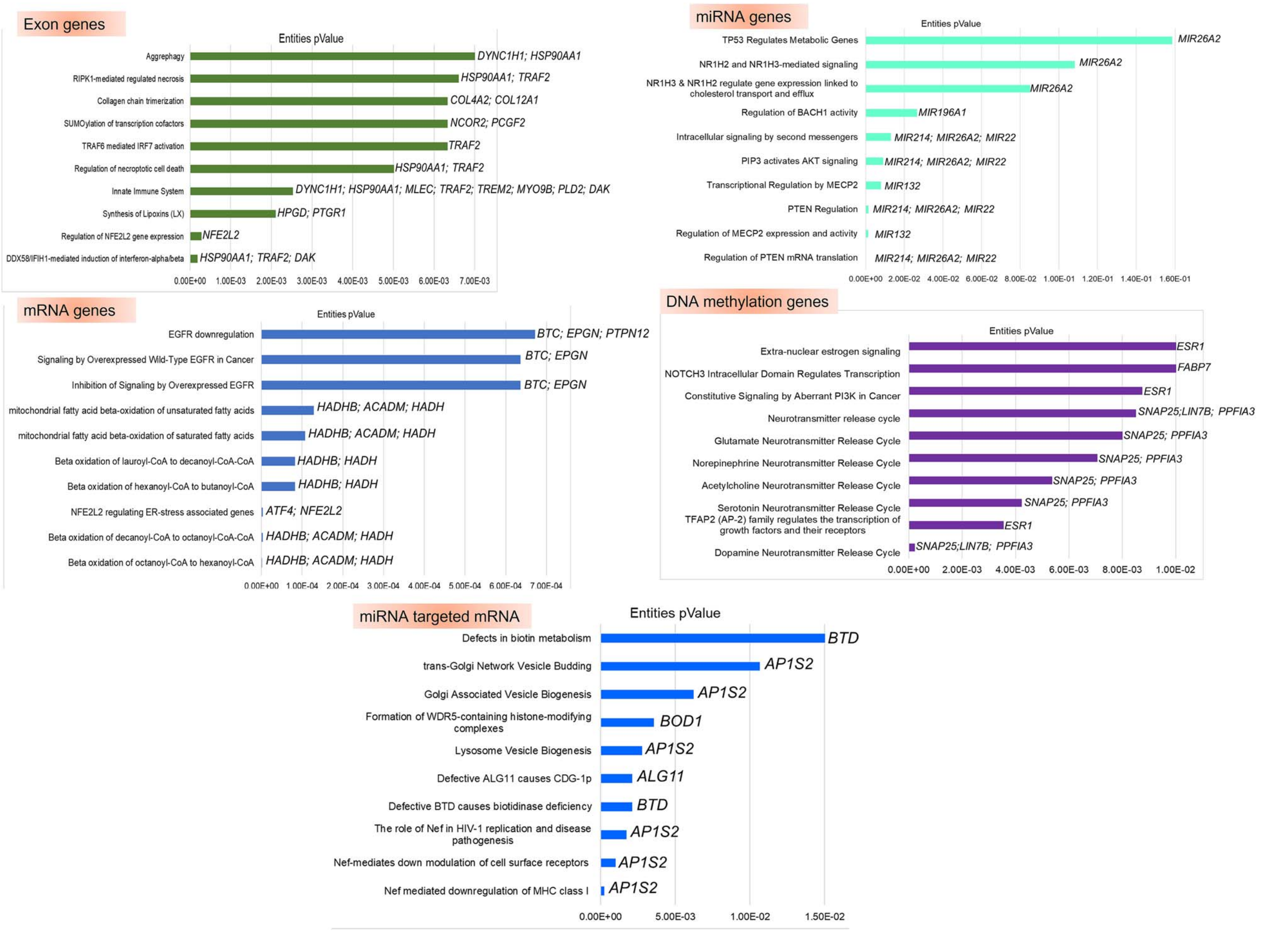

Figure 7. Top Reactome pathway mapping results of top genes with their *P*-values.

outcomes [61]. The study shows that by activating the Akt signaling pathway, the downregulation of *HADH* plays a critical role in the advancement of gastric cancer [62]. *MGAT5* was identified by Wang *et al.* as a potential SC target agent. They observed that the overexpression of the gene *MGAT5* in SC encourages the growth and spread of tumor cells [63]. The study examined the expression of the gene *KCNMA1* in relation to digestive malignancies, such as SC, indicating that it may serve as a marker for prognosis [64]. The gene *KCNMA1* hypermethylation influences cancer cell apoptosis and malignant behavior via regulating *PTK2* expression and is associated with poor survival results [64]. Our comprehensive bioinformatics analysis and existing literature survey revealed that certain genes: *HSP90AA1*, *COL4A2*, *COL12A1*, *TTC8*, *HADH*, *GPX3*, *CMTM5*, *MGAT5*, and *KCNMA1* are strongly linked to the progression of SC samples. The other top genes did not show significant results of stage-wise expression as well as month-wise overall survival outcome as shown in Fig. 9a and b. The sample-specific expression interpretation performed for genes showed significant results in stage-wise expression analysis and survival analysis. We observed in Fig. 9c, that the genes are differentially expressed between two different conditions of normal and healthy samples. The existing study (literature survey based on molecular laboratory and computational experiments) signified obtained top gene's molecular mechanism related to SC diagnosis and treatment. Some genes or markers are differently expressed and their expression leads to SC tumor growth which can be utilized for SC diagnosis. The biological or clinical traits of the identified molecular markers linked to the prognosis of a disease aid in forecasting how the disease will progress or how a patient will react to the therapy. The pathway analysis study (shown in Supplementary Table S5) revealed that identified molecular markers' downregulation and upregulation functions and other molecular mechanisms can play a critical role in the advancement of SC. Additionally, our proposed model can precisely identify SC samples with the aid of identified molecular marker expression across samples, which may be useful for SC patient's diagnosis. These nine key genes were found to be significant poor prognostic factors based on experimental results and literature survey. As a result, they could be potential molecular signatures for better diagnosis, more effective chemotherapy, and prognostic outcome prediction.

Strength and limitations of DOMSCNet

A single-omics dataset is inherently complex due to high dimensionality. The proposed model encounters challenges in capturing the complex multi-omics profiles across different layers to classify samples effectively. To address this, the integration of filter-wrapper-embedded techniques as HFS technique (SMB) was used to reduce dimensionality from the omics dataset and identify relevant features for model enhancement to classify samples. There is a challenge that occurs during the reduction of dimensionality from DNA methylation and Exon expression data, which contain millions of features relative to a small number of samples. This also affects the model's ability to identify subtypes due to the unavailability of a properly labeled dataset of SC. However, one limitation of the proposed model is that its accuracy decreases

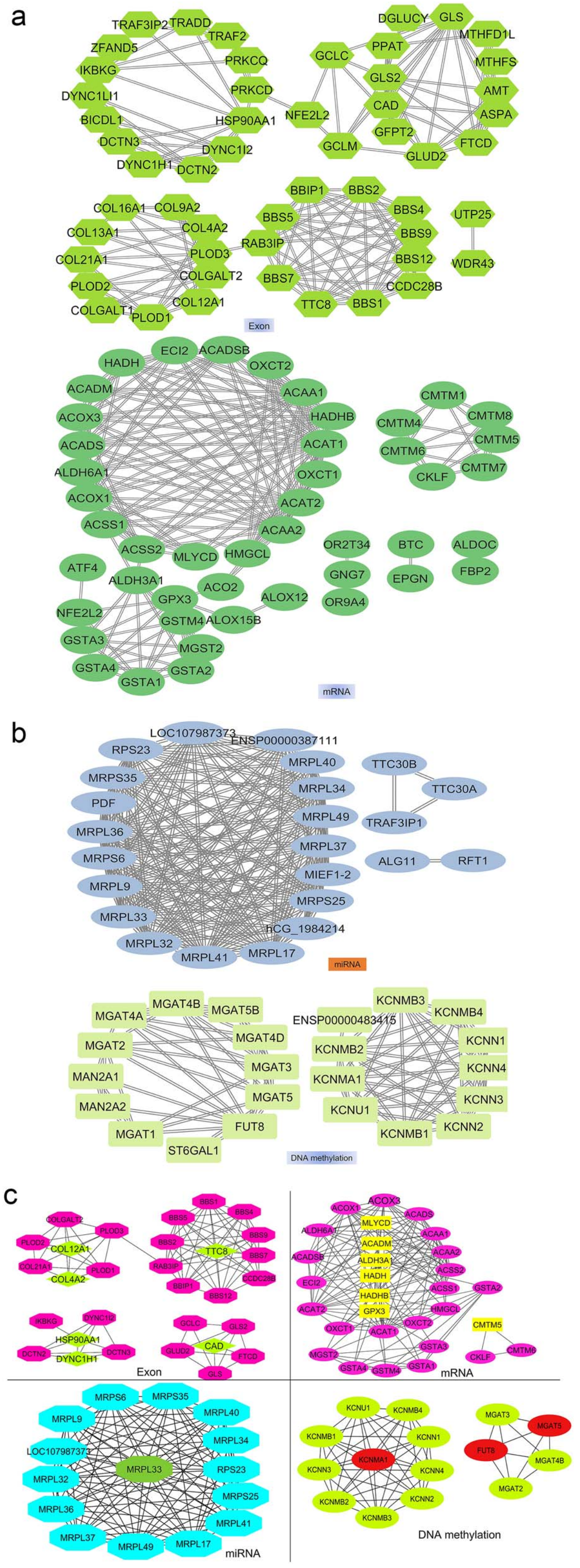

Figure 8. (a)-(b) The gene–protein interaction and (c) the hub gene interaction network of top genes.

when applied to datasets with a small sample size compared to larger datasets. SC is highly heterogeneous, with well-established molecular subtypes that play a promising role in determining the clinical outcomes of SC patients. Future studies should aim to identify SC subtypes using both bulk RNA-Seq and single-cell RNA-Seq (scRNA-Seq) data and robustness for small-sized datasets, which can pave the way for advancements in precision medicine studies.

Conclusion

The present study aims to identify a robust model for the classification of SC and normal tissue from the multi-layer omics data. The high-dimensional NGS sequencing data contained a small number of samples over thousands of features. These features produced noise, storage issues, and model overfitting, as well as underfitting problems. The proposed HFS techniques can tackle the issue by selecting the most relevant features and improving the proposed model's performance. The experimental results of Exon, mRNA, miRNA expression, and DNA Methylation data of SC demonstrate the efficiency of the proposed SMB HFS technique outperforming the other FS techniques. The proposed DOMSCNet outperforms Model2 and Model3 along with state-of-the-art existing models for all multi-layer omics datasets. From the experimental and MCDA results, we found that DOMSCNet efficiently classifies SC from a multi-layer omics profile with AUC scores of 0.988 (Exon), 0.986 (mRNA), 0.99 (miRNA), 0.971 (DNA methylation), and TOPSIS scores of 1, respectively. The proposed model DOMSCNet performance was validated with an external dataset of NCBI Geodataset (obtained highest AUC score of 0.978) and TCGA-LIHC multi-layer omics dataset (obtained highest AUC score of 0.994). The DOMSCNet performed better on TCGA-LIHC across all four datasets. From the bioinformatics analysis and existing literature survey validated that the top selected genes of Exon, mRNA, and DNA methylation profiles, the gene *HSP90AA1*, *COL4A2*, *COL12A1*, *TTC8*, *HADH*, *GPX3*, *CMTM5*, *MGAT5*, and *KCNMA1* are closely associated with SC poor prognosis, poor survival, and cancer pathway mapping signified that these nine genes could be a potential putative molecular marker of SC. Additionally, this bioinformatics analysis of top genes could be helpful for a better understanding of the molecular pathophysiology of SC. The experimental and literature survey results help clinicians and researchers to provide treatment against the survival of SC. However, a thorough clinical study is required to improve the understanding of the molecular mechanism of markers identified over SC.

Key Points

- Hybrid feature selection (HFS) techniques performed to reduce the high dimensionality of multi-layer omics datasets for uncovering optimal features.
- Integration of effective HFS algorithm with hybrid deep recurrent neural network model for classification of stomach cancer over normal samples of Exon, mRNA, miRNA expression, and DNA methylation dataset.
- Validation of external datasets to improve the robustness and generalizability of the proposed DOMSCNet model.
- Interpretation of cancer prognosis of identified potential molecular signatures through bioinformatics approach.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

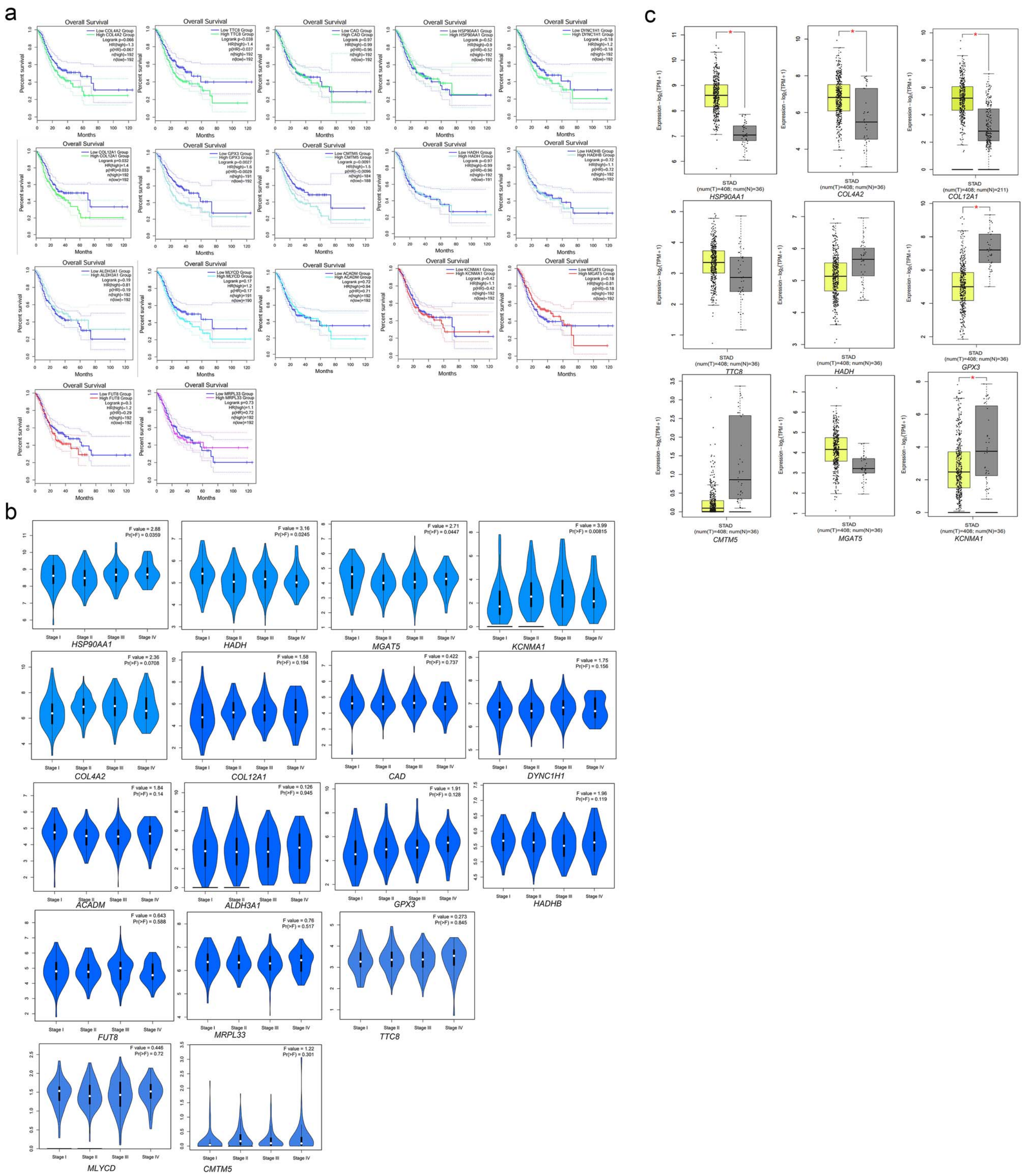

Figure 9. (a) Overall survival analysis, (b) pathological stage-wise expression, and (c) differential expression of top genes.

Author contributions

Conceptualization: K.B., H.S.D., S.M.; Data Curation: K.B.; Methodology: K.B., H.S.D.; Investigation: K.B.; Formal Analysis: K.B.; Software: K.B.; Validation: K.B., H.S.D; Visualization: K.B. Writing–Original Draft: K.B, H.S.D; Writing–Review & Editing: K.B., H.S.D, R.K.B., K.A., S.M.; Supervision: H.S.D., S.M.; Funding acquisition: R.K.B.

Conflict of interest: None declared.

Funding

This research was funded by Researchers Supporting Project Number (RSPD2025R947), King Saud University, Riyadh, Saudi Arabia.

Data availability

The mRNA expression and DNA methylation datasets are available at https://portal.gdc.cancer.gov/. The Exon expression and miRNA expression datasets are available at https://xena.ucsc.edu/. The external validation dataset is available at NCBI Geodataset with Accession id: GSE36968, GSE30601, GSE198834, and GSE30760.

Codes availability

The proposed study-related codes are available at https://github.com/kasmikaborah/HFS-DL-Classification.

References

1. Heo YJ, Hwa C, Lee G-H. *et al.* Integrative multi-omics approaches in cancer research: from biological networks to clinical subtypes. *Mol Cells* 2021;**44**:433–43. https://doi.org/10.14348/molcells.2021.0042.
2. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. Feature selection for high-dimensional data. *Prog Artif Intell* 2016;**5**: 65–75. https://doi.org/10.1007/s13748-015-0080-y.
3. Ferreira AJ, Figueiredo MA. Efficient feature selection filters for high-dimensional data. *Pattern Recogn Lett* 2012;**33**:1794–804. https://doi.org/10.1016/j.patrec.2012.05.019.
4. Mehmood A, Kaushik AC, Wei DQ. DDSBC: a stacking ensemble classifier-based approach for breast cancer drug-pair cell synergy prediction. *J Chem Inf Model* 2024;**64**:6421–31. https://doi.org/10.1021/acs.jcim.4c01101.
5. Borah K, Das HS, Seth S. *et al.* A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. *Funct Integr Genomics* 2024;**24**:139. https://doi.org/10.1007/s10142-024-01415-x.
6. Bhadra T, Mallik S, Hasan N. *et al.* Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer. *BMC Bioinformatics* 2022;**23**:153. https://doi.org/10.1186/s12859-022-04678-y.
7. Ahn T, Goo T, Lee C-ʜ. *et al.* Deep learning-based identification of cancer or normal tissue using gene expression data. In: Huiru Zheng, Xiaohua Hu, Zoraida Callejas et al. (eds.), *IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Bioinformatics and Biomedicine*. Madrid, Spain: IEEE, 3-6 Dec 2018, pp. 1748–52, 2018.
8. Guo L-Y, Wu A-H, Wang Y-X. *et al.* Deep learning-based ovarian cancer subtypes identification using multiomics data. *BioData Mining* 2020;**13**:1–12.
9. Dutta P, Patra AP, Saha S. Deeprog: a deep attention-based model for diseased gene prognosis by fusing multi-omics data. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**19**:2770–81.
10. Schmidt B, Hildebrandt A. Deep learning in next-generation sequencing. *Drug Discov Today* 2021;**26**:173–80. https://doi.org/10.1016/j.drudis.2020.10.002.
11. Ilic M, Ilic I. Epidemiology of stomach cancer. *World J Gastroenterol* 2022;**28**:1187–203. https://doi.org/10.3748/wjg.v28.i12.1187.
12. Machlowska J, Baj J, Sitarz M. *et al.* Gastric cancer: epidemiology, risk factors, classification, genomic characteristics and treatment strategies. *Int J Mol Sci* 2020;**21**:4012. https://doi.org/10.3390/ijms21114012.
13. Sohn BH, Hwang JE, Jang HJ. *et al.* Clinical significance of four molecular subtypes of gastric cancer identified by the cancer genome atlas project. *Clin Cancer Res* 2017;**23**:4441–9. https://doi.org/10.1158/1078-0432.CCR-16-2211.
14. Wang Q, Liu G, Hu C. Molecular classification of gastric adenocarcinoma. *Gastroenterol Res* 2019;**12**:275–82. https://doi.org/10.14740/gr1187.
15. Jeong YS, Eun YG, Lee SH. *et al.* Clinically conserved genomic subtypes of gastric adenocarcinoma. *Mol Cancer* 2023;**22**:147. https://doi.org/10.1186/s12943-023-01796-w.
16. Li B, Zhang F, Niu Q. *et al.* A molecular classification of gastric cancer associated with distinct clinical outcomes and validated by an XGBoost-based prediction model. *Mol Ther Nucl Acids* 2023;**31**:224–40. https://doi.org/10.1016/j.omtn.2022.12.014.
17. Bhonde SB, Wagh SK, Prasad JR. Identification of cancer types from gene expressions using learning techniques. *Comput Methods Biomech Biomed Eng* 2023;**26**:1951–65. https://doi.org/10.1080/10255842.2022.2160243.
18. Metipatil P, Bhuvaneshwari P, Basha SM. *et al.* An efficient framework for predicting cancer type based on microarray gene expressions using cnn-filstm technique. *SN Computer Sci* 2023;**4**:381.
19. Susmi SJ. An efficient gene expression data classification using optimized bidirectional long short-term memory with self-attention mechanism. *Multimed Tools Appl* 2024;**83**:74159–76. https://doi.org/10.1007/s11042-024-18387-6.
20. Babichev S, Liakh I, Kalinina I. Applying the deep learning techniques to solve classification tasks using gene expression data. *IEEE Access* 2024;**12**:28437–48. https://doi.org/10.1109/ACCESS.2024.3368070.
21. Mallick P, Sinha M, Poray J. *et al.* Recognition of altered gene-gene interaction using bilstm in different stages of lung adenocarcinoma. *Procedia Comput Sci* 2024;**235**:1213–21. https://doi.org/10.1016/j.procs.2024.04.115.
22. El-Manzalawy Y, Hsieh T-Y, Shivakumar M. *et al.* Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Med Genet* 2018;**11**:19–31. https://doi.org/10.1186/s12920-018-0388-0.
23. Sahu B. Multi-tier hybrid feature selection by combining filter and wrapper for subset feature selection in cancer classification. *Indian J Sci Technol* 2019;**12**:1–11. https://doi.org/10.17485/ijst/2019/v12i3/141010.
24. Tabakhi S, Lu H. Multi-agent feature selection for integrative multi-omics analysis. In: Riccardo Barbieri, Editor in Chief. *Proceedings of 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Glasgow, Scotland: Institute of Electrical and Electronics, Engineers (IEEE), 11-15 Jul 2022, pp. 1638–42. ISBN 9781728127835.
25. Li M, Guo H, Wang K. *et al.* Avbae-modfr: a novel deep learning framework of embedding and feature selection on multi-omics data for pan-cancer classification. *Comput Biol Med* 2024;**177**:108614. https://doi.org/10.1016/j.compbiomed.2024.108614.
26. Wang Y, Gao X, Ru X. *et al.* A hybrid feature selection algorithm and its application in bioinformatics. *PeerJ Comput Sci* 2022;**8**:e933. https://doi.org/10.7717/peerj-cs.933.
27. Mahto R, Ahmed SU, Rahman RU. *et al.* A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection. *BMC Bioinformatics* 2023;**24**:479. https://doi.org/10.1186/s12859-023-05605-5.
28. Zhou Y, Jia E, Shi H. *et al.* Prediction of Time-Series Transcriptomic Gene Expression Based on Long Short-Term Memory with Empirical Mode Decomposition. *Int J Mol Sci* 2022;**23**:7532. https://doi.org/10.3390/ijms23147532.
29. Chai H, Deng W, Wei J. *et al.* A contrastive-learning-based deep neural network for cancer subtyping by integrating multi-omics data. *Interdiscip Sci* 2024;**16**:966–75. https://doi.org/10.1007/s12539-024-00641-y.

30. Zhenfei W, Ali MM, Sahibzada KI. *et al.* Hybrid feature extraction for breast cancer classification using the ensemble residual VGG16 deep learning model. *Curr Bioinform* 2024;**20**:149–63. https://doi.org/10.2174/0115748936333380240816053223.
31. Huang Y, Zeng P, Zhong C. Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning. *BMC Bioinformatics* 2024;**25**:132. https://doi.org/10.1186/s12859-024-05749-y.
32. Ren Y, Gao Y, Du W. *et al.* Classifying breast cancer using multi-view graph neural network based on multi-omics data. *Front Genet* 2024;**15**:1363896. https://doi.org/10.3389/fgene.2024.1363896.
33. Mohamed TIA, Ezugwu AE-S. Enhancing lung cancer classification and prediction with deep learning and multi-omics data. *IEEE Access* 2024;**12**:59880–92. https://doi.org/10.1109/ACCESS.2024.3394030.
34. Lan W, Liao H, Chen Q. *et al.* DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. *Brief Bioinform* 2024;**25**:bbae185. https://doi.org/10.1093/bib/bbae185.
35. Divate M, Tyagi A, Richard DJ. *et al.* Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures. *Cancers (Basel)* 2022;**14**:1185. https://doi.org/10.3390/cancers14051185.
36. Yang S, Wang Z, Wang C. *et al.* Comparative evaluation of machine learning models for subtyping triple-negative breast cancer: a deep learning-based multi-omics data integration approach. *J Cancer* 2024;**15**:3943–57. https://doi.org/10.7150/jca.93215.
37. Goldman MJ, Craft B, Hastie M. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* 2020;**38**:675–8. https://doi.org/10.1038/s41587-020-0546-8.
38. Kang M, Tian J. Machine learning: data pre-processing. In: Michael G. Pecht, Myeongsu Kang (eds.), *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*. Hoboken, New Jersey, U.S.: Wiley, 2018, pp. 111–30.
39. Ritchie ME, Phipson B, Wu D. *et al.* Limma powers differential expression analyses for RNA sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47–7. https://doi.org/10.1093/nar/gkv007.
40. Mallik S, Bhadra T, Maulik U. Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based feature selection for multi-omics data. *IEEE Trans Nanobiosci* 2017;**16**:3–10. https://doi.org/10.1109/TNB.2017.2650217.
41. He X, Cai D, Shao Y. *et al.* Laplacian regularized Gaussian mixture model for data clustering. *IEEE Trans Knowl Data Eng* 2010;**23**:1406–18.
42. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 2005;**27**:1226–38. https://doi.org/10.1109/TPAMI.2005.159.
43. Kursa MB, Jankowski A, Rudnicki WR. Boruta—a system forfeature selection. *Fundamenta Informaticae* 2010;**101**:271–85. https://doi.org/10.3233/FI-2010-288.
44. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol* 1996;**58**:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
45. Yue Z, Sun C, Gao L. *et al.* Machine learning efficiently corrects LIBS spectrum variation due to change of laser fluence. *Opt Express* 2020;**28**:14 345–56. https://doi.org/10.1364/OE.392176.
46. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* MIT Press 1997;**9**:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
47. Cho K, van Merriënboer B, Gulcehre C. *et al.* Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Alessandro Moschitti, Bo Pang, Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics, pp. 1724–34.
48. Das HS, Roy P. A cnn-bilstm based hybrid model for Indian language identification. *Appl Acoust* 2021;**182**:108274. https://doi.org/10.1016/j.apacoust.2021.108274.
49. Zhang Y, Wang J, Zhang X. Ynu-hpcc at semeval-2018 task 1: Bilstm with attention-based sentiment analysis for affect in tweets. In: Marianna Apidianaki, Saif M. Mohammad, Jonathan May *et al.* (eds.), *Proceedings of the 12th International Workshop on Semantic Evaluation, June 5-6, New Orleans, Louisiana*. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics, pp. 273–8, 2018.
50. Griss J, Viteri G, Sidiropoulos K. *et al.* Reactomegsa-efficient multi-omics comparative pathway analysis. *Mol Cell Proteomics* 2020;**19**:2115–25. https://doi.org/10.1074/mcp.TIR120.002155.
51. Liu X, Yang B, Huang X. *et al.* Identifying lymph node metastasis-related factors in breast cancer using differential modular and mutational structural analysis. *Interdiscip Sci Comput Life Sci* 2023;**15**:525–41. https://doi.org/10.1007/s12539-023-00568-w.
52. Tang Z, Kang B, Li C. *et al.* Gepia2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucl Acids Res* 2019;**47**:W556–60. https://doi.org/10.1093/nar/gkz430.
53. Saikia S, Si T, Deb D. *et al.* Lesion detection in women breast's dynamic contrast-enhanced magnetic resonance imaging using deep learning. *Sci Rep* 2023;**13**:22555. https://doi.org/10.1038/s41598-023-48553-z.
54. Park SG, Ji MJ, Ham IH. *et al.* Secretome analysis reveals reduced expression of COL4A2 in hypoxic cancer-associated fibroblasts with a tumor-promoting function in gastric cancer. *J Cancer Res Clin Oncol* 2023;**149**:4477–87. https://doi.org/10.1007/s00432-022-04361-y.
55. Gao X, Zhong S, Tong Y. *et al.* Alteration and prognostic values of collagen gene expression in patients with gastric cancer under different treatments. *Pathol Res Pract* 2020;**216**:152831. https://doi.org/10.1016/j.prp.2020.152831.
56. Duan S, Gong B, Wang P. *et al.* Novel prognostic biomarkers of gastric cancer based on gene expression microarray: COL12A1, GSTA3, FGA and FGG. *Mol Med Rep* 2018;**18**:3727–36.
57. He Q, Chen N, Wang X. *et al.* Prognostic value and immunological roles of GPX3 in gastric cancer. *Int J Med Sci* 2023;**20**:1399–416. https://doi.org/10.7150/ijms.85253.
58. Zhou C, Pan R, Li B. *et al.* GPX3 hypermethylation in gastric cancer and its prognostic value in patients aged over 60. *Future Oncol* 2019;**15**:1279–89. https://doi.org/10.2217/fon-2018-0674.
59. Liang Z, Xie J, Huang L. *et al.* Comprehensive analysis of the prognostic value of the chemokine-like factor-like MARVEL transmembrane domain-containing family in gastric cancer. *J Gastrointest Oncol* 2021;**12**:388–406. https://doi.org/10.21037/jgo-21-78.
60. Chang W, Ma L, Lin L. *et al.* Identification of novel hub genes associated with liver metastasis of gastric cancer. *Int J Cancer* 2009;**125**:2844–53. https://doi.org/10.1002/ijc.24699.

61. Wang X, Song H, Liang J. *et al.* Abnormal expression of HADH, an enzyme of fatty acid oxidation, affects tumor development and prognosis. *Mol Med Rep* 2022;**26**:1–8. https://doi.org/10.3892/mmr.2022.12871.
62. Shen C, Song YH, Xie Y. *et al.* Downregulation of HADH promotes gastric cancer progression via Akt signaling pathway. *Oncotarget* 2017;**8**:76279–89. https://doi.org/10.18632/oncotarget.19348.
63. Wang Y, Tan Z, Li X. *et al.* RUNX2 promotes gastric cancer progression through the transcriptional activation of MGAT5 and MMP13. *Front Oncol* 2023;**13**:1133476. https://doi.org/10.3389/fonc.2023.1133476.
64. Pesce A, Fagone P, Nicoletti F. *et al.* The role of KCNMA1 expression in digestive cancers: a potential prognostic biomarker. *Recent Pat Anticancer Drug Discov* 2022;**17**:324–5. https://doi.org/10.2174/1574892817666220104094425.

Briefings in Bioinformatics, 2025, **26(2)**, bbaf115
https://doi.org/10.1093/bib/bbaf115
Problem Solving Protocol