[1]**Begum S.**

[1]**Sandipan Dey**

[2]**Chakraborty D.**

[1]**Hembrom T.**

[2]**Hazra S.**

[1]**Barman D.**

Utilizing Machine Learning Techniques for Categorizing Cancer Based on Gene Expression Data: A Review

Journal of Electrical Systems

Abstract: - Cancer is a group of diseases which share one common feature: the growth of abnormal cells, thus ranking as the second leading cause of death globally, after cardiovascular diseases in WHO's report. Evaluating gene expression is based on the fact that it is the genesis of early cancer detection, the concurrence of the molecular and genetic processes. By using DNA microarray techniques and RNA-sequencing approaches, researchers in computational genomics can give quantitative measures of gene expression levels providing very accurate input data for computational evaluation. The current paper is about machine learning technology, which identifies cancer subtypes according to patterns of gene expression. It embraces the two distinct methodologies which are traditional plus deep learning with high proficiency focused on the cancer-related gene. The outline includes the most popular deep neural network designs such as MLPs, CNN, RNN, GNN, and the recently emerged Transformer networks. The review describes common data collection methods used in this field and some of the essential datasets for supervised machine learning. In addition, the specific techniques developed to cover the complicated horizontal spread of gene expression data are also presented. The article explores theoretical possibilities for the promotion of machine learning-based gene expression analysis in cancer classification towards the end.

Keywords: Microarray data, machine learning, RNA-sequencing, deep neural network

I. INTRODUCTION

Cancer as a generic name of various diseases is a group of diseases manifested by genetic mutations that lead to uncontrollable cell proliferation within the body. The uncontrolled development resulting from this can get into various organs, and usually, it ends up fatal. Cancer is the second largest cause of death after cardiovascular diseases globally [1]. The newest hits point to gene expression analysis as a critical method for solving problems in cancer diagnosis and drug discovery [2,3]. Researchers can gather precious knowledge about the genes involved in cancer initiation and development through the investigation of gene expression patterns. Gene alterations can be considered potential signs for early cancer diagnosis and help find the appropriate therapeutic targets. These insights form the basis of the predictive, anticipatory and preventive approach to healthcare [4]. Gene expression is a dynamic link connecting information stored in DNA with the synthesis of proteins or other molecules. The intricate process commences with the transcribing of DNA into mRNA, which in turn is translated into proteins. Analysing gene expression enables us to understand the sequence of genetic changes taking place under various conditions and on a tissue level or cell-by-cell basis. It enables us to measure DNA transcript levels present in samples, the active genes, and the degree to which they are active. A significant part of the gene expression quantification is to map the sequenced reads to the known genome or transcriptome references. The correct measure of it is dependent on

[1]Department of Computer Science & Engineering, Govt. College of Engineering & Textile Technology, Berhampore, India

E-mail: shemim_begum@yahoo.com

E-mail: sandipandey2508@gmail.com

E-mail: www.tanmay8000@gmail.com

[2]Asansol Engineering. College, India

E-mail: debasis.cse@aecwb.edu.in

Email: simanta.hazra@gmail.com

*Corresponding Author: Begum S

E-mail: shemim_begum@yahoo.com

the paired unique read data that could create a link between the nucleic acid sequences and the corresponding genes by bioinformatics.

The most used methods for analysing gene expression data consist of DNA microarrays and next-generation sequencing (NGS). DNA microarrays employ the use of a matrix of two dimensions in which specific genes bind to known DNA bases through hybridization. Whereas NGS methods (i.e., RNA sequencing (RNA-Seq), to name a few), provide access to high-throughput analyses, scalability, and fast results [6, 7]. RNA-Seq converts RNA molecules to the cDNA (complementary DNA), and the DNA sequence determines the RNA expression. Unlike DNA microarrays, however, RNA-Seq [8, 9] is characterized by improved specificity, higher resolution, increased sensitivity to differential expression, and a broader dynamic range. Moreover, RNA-Seq does quantitative transcriptomics on multiple levels at specific time intervals.

Analysing gene expression involves employing computational techniques to decipher gene regulation and understand their roles within cells and tissues. Machine learning (ML) methodologies have become instrumental in uncovering how genetic variations and regulatory regions influence traits, well-being, and overall health [10,11]. Initially, traditional ML techniques like Decision Trees and Support Vector Machines dominated this domain. However, over the last decade, deep learning (DL) approaches have surged in popularity. These DL methods excel in predicting the functionality and structure of genomic elements, including promoters, enhancers, and specific gene sequences, offering enhanced insights into genetic mechanisms [12,13]. In the realm of gene expression analysis, feature engineering stands as a crucial computational technique, particularly given the challenge posed by the vast dimensionality of data juxtaposed with a limited sample size. This examination delineates feature engineering strategies into three categories: filter, wrapper, and embedded methods [14]. Filter techniques sift through data to eliminate irrelevant or redundant features, basing decisions on each feature's correlation with the target prediction. Wrapper methodologies harness classification algorithms to gauge feature significance, encapsulating the classifier within a search mechanism to pinpoint the optimal feature subset. Conversely, embedded strategies [15,16] integrate feature selection directly into the classifier's learning phase, spotlighting pivotal features to augment classification performance. While filter methods prioritize efficiency and computational simplicity, wrapper and embedded techniques excel in isolating pertinent features, thereby enhancing classification accuracy. In existing literature, a range of deep neural network (NN) architectures has been utilized for cancer classification using gene expression data. These include multi-layer perceptron (MLP), convolutional neural networks (CNN), recurrent neural networks (RNN), graph neural networks (GNN), and transformer neural networks (TNN). MLPs are characterized by connections linking every neuron to all preceding and succeeding layers. In gene expression data analysis, the MLP's input layer processes gene expression profiles, with individual probes corresponding to distinct neurons. The MLP's output layer then yields class probabilities for the gene expression sample [17].

CNNs, originally tailored for processing multidimensional arrays like images, utilize two-dimensional convolutional filters to learn hierarchical data representations. Some studies have restructured gene expression data into image-like two-dimensional arrays, leveraging this format as CNN inputs [18]. Given CNNs' adeptness at capturing local spatial relationships, they often outperform MLPs in gene expression analysis classification. Moreover, one-dimensional CNN variants feed each gene expression data row directly to the network, utilizing layers with one-dimensional convolutional filters. CNNs consistently emerge as leading deep learning models for gene expression studies. RNNs incorporate recurrent connections, specifically crafted for sequential data modelling. Utilizing a state vector, RNNs amalgamate current input information with previously stored data to generate outputs. This design makes RNNs apt for discerning correlations within sequential gene expression data, shedding light on underlying cancer developmental processes [19, 20]. However, RNNs can exhibit elevated computational demands and increased susceptibility to overfitting, especially with limited data, relative to CNNs. GNNs, on the other hand, harness architectures tailored for learning graph-based data representations through edges and nodes.

These models convert gene expression data into graph-based representations, leveraging gene expression topology to discern correlations among genes [21]. The capability of GNNs to comprehend graph-structured data positions them promisingly for future gene expression analyses, as evidenced by recent research. TNNs employ a network design incorporating the self-attention mechanism, facilitating the recognition of distant dependencies in sequential data. This feature equips TNNs to pinpoint correlations within gene expression datasets, leading to their adoption in prior research. Distinctively, TNNs enable the concurrent processing of input samples during model training, accelerating the analysis of extended sequences. Moreover, researchers have crafted hybrid architectures, like TNNs integrated with 1D convolutional layers, aimed at capturing shared genetic information across cancer types without necessitating feature selection [22]. Finally, the emerged transfer learning methods, which make use of knowledge possessed by a model with a large dataset to solve the problem of data scarcity among another model and model dimension, have become a major solution to the problem of data limitation and high dimensionality of gene expression data [23,24].

Top of Form

Notwithstanding many milestones in machine learning (ML) for cancer classification with gene expression data are being achieved, there are still a lot of obstacles because the role of gene expression is ambiguous. The variance in the data sets of gene expression can be surprisingly less despite a very vast dimensionality due to the large number of genes. Redundant data is normally removed by trying different kinds of the feature-engineering techniques, which allow to select the meaningful points for classification. Although old ML implementations predominantly rely on expert feature engineering and especially careful data preprocessing, integrated feature engineering alleviates the need for such a high level of expertise. Moreover, Transfer Learning has become a tool that facilitates overcoming data volume issues, and this is again with limited data. Traditionally, DL methods have shown superior performance to ML algorithms as reflected in the metrics that measure model accuracy, therefore a future that relies on DL for building gene expression analysis models. At present, the model ensemble employing techniques such as MLP and CNN and the best practice of transfer learning and feature engineering sets the classification performance of achieving the classification accuracy of 90% and above. Nevertheless, these approaches have significant scope for error, when considering elaborate parameters and necessary improvement for their more inclusive applications. To add to this, most current methods are complicated by interpretability difficulties and have a narrow range of uses for diverse data types and modalities. In the sentence above, the author emphasizes the need for the incorporation of advanced tools that would address the shortcomings of current techniques. Such negative consequences include interpretability difficulties, a limited range of uses for diverse data types, and a lack of compatibility with diverse modalities.

Although there are a lot of review articles in different scholarly literature that have closely inspected the advancement of computational techniques for expression gene analysis, this article has focused on innovative approaches. Table 1 samples the most recent review papers which are focusing on this study of traditional “Machine learning (ML) methodologies, feature engineering methodologies, and Deep Learning (DL) techniques towards solving the same issue that we pursue. Notably, the data type is being analysed as well. Some of the review papers choose to focus on the traditional ML techniques used in the assessment of gene expression, while some manuscripts explore the feature-engineering techniques or specific mediums used in gene expression analysis. The earlier genome-wide computational analysis research was mostly concerned with DNA microarray analysis of gene expression profiles. Quite interestingly, while there is a considerable overlap between subjects of reviews done before 2019 in Table 1 and what this article discusses, the current contribution looks at certain issues not previously looked at in other publications. The core focus of this survey is its in-depth examination of both the traditional machine learning and the contemporary deep learning approaches in mRNA analysis. The earlier reviews refer to the implementation of DL models like MLP, CNN and RNN without discussing the involvement of GNN and TNN which, despite their emergence as models for the retrospective analysis phase, is yet to be discussed thoroughly [25,26]. Additionally, this review discusses the relationship between RNA-Seq and gene expression modelling, an area of research that is currently highly visible. Moreover, the area of interest is complemented by the examination of associated feature-engineering methods and datasets that are essential in ML driven gene expression analyses, and which are indeed lacking articles in different literature reviews available.

Table 1. List of earlier review papers for gene expression work using ML and DL approaches

Reference	ML Approaches	DL Approaches	Microarray Data	RNA-Seq Data
Sathe et al., 2019 [29]	No	CNN & RNN	Yes	Yes
Koumakis et al., 2020 [30]	No	CNN & RNN	Yes	Yes
Zhu et al., 2020 [31]	No	CNN, RNN, MLP & NN	Yes	Yes
Gunavathi et al., 2020 [32]	No	CNN	Yes	Yes

Tabares et al., 2020 [33]	Yes	CNN & MLP	Yes	No
Mazlan et al., 2021 [34]	Yes	CNN	Yes	Yes
Karim et al., 2021 [35]	Yes	CNN, RNN, MLP & NN	Yes	Yes
Thakur et al., 2021 [36]	Yes	CNN	Yes	Yes
Montesinos-López et al., 2021 [37]	No	CNN, RNN, MLP & NN	Yes	Yes
Bhandari et al., 2022 [38]	Yes	CNN, RNN, MLP & NN	Yes	No
Khalsan et al., 2022 [39]	Yes	CNN, RNN, MLP & NN	Yes	Yes

II. GENE EXPRESSION DATA

Gene expression study means the investigation of the number of transcripts produced by particular cells or tissues providing insights into the levels of the gene activity. Transcriptomics, a related field of technology, measures the size of the transcriptome. In the early era of computational methods of transcriptomics, Sanger sequencing was used to do 'expressed sequence tag' libraries.

The DNA of cDNA is an organism or tissue sample built from only mRNA. Up to this point, approximately 171 billion DNA libraries from about 1,400 cellular organisms have been created. Although EST libraries provide sequences of expressed genes, they often fail to furnish optimally complete sequences of these genes. As a result, paint tag-centric techniques like Serial Analysis of Gene Expression (SAGE) progressed and enabled the quantitative analysis of various transcripts within cellular systems without prior knowledge of the gene. SAGE works from a theoretical background, whereby the same nucleotide distribution is assumed to be prevalent throughout the genome. Over time, methods like Sanger sequencing of EST libraries and SAGE were devised as the premonition of further methods like DNA microarrays or most impressively, RNA-Seq, thus raising the accuracy and the range of estimation of gene expression.

2.1 MICROARRAY DATA

Microarray data are an outcome of a laboratory method where the researchers use a 2D device, usually referred to as chips or slides, coated in a lot of micro spots. There are many lights lit on this slide, each one representing a specific DNA sequence or a gene. Using a hybridization technique, responding DNA samples attach to these spots. The second stage through analysis reveals the colour of these marks and measures the expression level of the specific genes. In the data model, each row represents the expression of a gene while columns correspond to each biological sample.

Microarrays assist in various tasks, such as the separation of DNA like RNA or comparative genomic hybridization in the majority of occasions as cDNA after retrotranscription. Such analyses offer views into genome-wide expression profiles related to specific conditions, such as cancer to support research and drug development, pharmacogenomics, and therapeutic strategies.

Despite its usefulness in contrarily measuring the expression of many genes at the same time, DNA microarrays have their limitations. They are mostly associated with inexactness, faultiness, imprecision, and uncertainty. Moreover, the experimental design is sensitive to factors like the degradation rates of genetic materials and amplification processes that can be responsible for imprecision in quantifying gene expression.

2.2 RNA-SEQE DATA

Next-Generation Sequencing (NGS) method RNA sequencing (RNA-Seq) tends to be the most utilized tool [41] because it allows for profiling at a higher speed than its counterparts. This approach allows researchers to trace RNA of any organism with a fantastic degree of precision, as well as to accurately identify its presence and quantities at specific epoch of time [42]. By creating millions of RNA profiles based on complex RNA samples, RNA-Seq covers several different purposes. Specifically, techniques that involve, the determination of gene expression, keeping a record of the expression patterns over time or response to treatments, transcript annotation, tackling post-transcriptional modifications, and uncovering alternative splicing and polyadenylation are taken into account. RNA-Seq's versatility plants it well for the simultaneous analysis of different types of RNA molecules within cells or tissues – coding messenger mRNA, non-coding regulatory RNA (miRNA, siRNA) and functional RNA (tRNA, rRNA) – altogether quantifying their levels. Primarily, RNA-Seq gives high resolution and impressive measurability setting ahead of transcriptome studies [43]. Thanks to these benefits RNA-Seq is undermining more and more microarrays in gene expression research.

Table 2 shows the comparison between microarray data and RNA-Seq data covering genes raged, various isoforms, resolution, background noise, costs, rare and new transcript detection, and non-coding RNA identification. To sum up, the RNA-Seq data outperforms the microarray in versatile aspects.

For instance, single-cell RNA sequencing (scRNA-Seq) ways [44], generate extensive transcriptome profiles at the individual cell level. This technique allows for more in-depth evaluations at a quicker rate, thus creating larger datasets compared to the traditional method of RNA-Seq. It gives researchers a necessary and accurate way to discover genes expressed in a diverse sample down to the single-cell level, measuring their expression across thousands of cells.

2.3 RNA-SEQ DATA COLLECTION

To generate RNA-seq data for identification, the cDNA is obtained by the transcription of RNA. This leads to the creation of the cDNA libraries that are sequenced and analysed through the NGS platform. The activity of sequencing is to purify and isolate mRNA molecules. mRNA transcribes to cDNA through the process of reverse transcription that occurs due to the presence of the retrovirus enzyme. Then, cDNA libraries are synthesized from cDNA fragments which are repeatedly copied and amplified to create them. These types of libraries are commonly subjected to NGS for the analysis of the relative number of transcripts of various genes. The efficacy is dominated by the quantity of biological and technical replicates, the sequencing depth, and the universality of the target transcriptome.

Some of those experimental decisions may not be even noticeable in the quality of the RNA-Seq, but meticulous testing is extremely important. The design applies the strategy of achieving quality objectives and at the same time overcoming the confines of time and cost.

2.4 GENE EXPRESSION DATASETS

The scientific community devoted years of work to collecting, organizing, and integrating various gene expression datasets. Table 3 shows RNA-Seq and microarray platforms gene expression datasets, focussing human tissue samples. These datasets are in the public domain, can easily be obtained, and are widely applied for cancer classification and similar research issues.

Table 2. Gene Expression Datasets

Reference	No. of Classes	Types of Cancer	Sample Size	Datasets
Mohammed et al., 2021 [45]	Multiclass Classification	Breast Cancer, Colon Adenocarcinoma, Lung Adenocarcinoma, Ovarian, and Thyroid cancer	2166	RNA-Seq
Li et al., 2022 [46]	Two	Renal carcinoma	945	RNA-Seq
Zhang et al., 2022 [47]	Two	Liver r Carcinoma	424	RNA-Seq
Coleto-Alcudia et al., 2022 [48]	Two	Breast Cancer	1178	RNA-Seq
Abdelwahab et al., 2022 [49]	Two	Lung Adenocarcinoma	549	RNA-Seq
Houssein et al., 2021 [50]	Two	Leukemia	72	Microarray
Hira et al., 2021 [51]	Multiclass Classification	Different types of Cancer	2096	Microarray
Vaiyapuri et al., 2022 [52]	Binary classification	Ovarian Cancer	253	Microarray

Lin Ke et al., 2022 [53]	Binary classification	Lung Cancer	181	Microarray
Deng et al., 2022 [54]	Binary classification	Myeloma	173	Microarray

III. FEATURE ENGINEERING

Feature engineering converts raw data into handcrafted features or descriptors which can highlight some critical information or fixing some data analysis constraints for machine learning models. The purpose of this transformation is either feature construction or feature refinement, working in favour of both supervised and unsupervised learning, causing simplification of data transformation and increases in prediction accuracy to be achieved. Feature extraction techniques have proved immensely crucial in eliminating the marker genes that can be associated with the increased discriminating capacity of machine learning in the identified dimensions [55-57], while ascertaining which genetic attributes are most important. As far as RNA-Seq data goes, which is a large-gene collection relative to the number of samples, feature selection has a specific role of zeroing in on a small subset of genes that most accurately capture the patterns of a dataset in the simplest space achievable which increases the signal-to-noise ratio.

Feature engineering algorithms for gene expression data generally fall into three categories: shifter, filter, and enclosed method. Figure 1 depicts the action of these three primary feature engineering methodologies from scrutinising the data to the establishment of the final models.

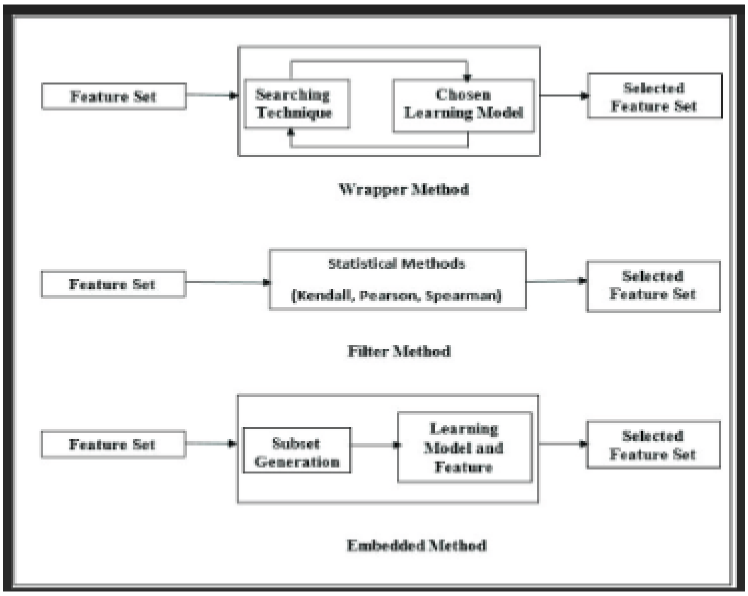

Figure 1. Three Types of Feature selection techniques methods a) Wrapper b) Filter c) Embedded. Method [58]

3.1 FILTER METHODS

Filtration methods in feature engineering designate such methods generated to delete the features of data that are considered non-essential hence the models established to predict gene expression are improved [59]. In most cases, filters are used as first processes in the preprocessing. The selection process is done by applying a specific filter to determine the genes’ relevance and assigning scores to these genes, then using a threshold to select only the relevant genes [60]. These techniques distribute weights according to the properties of the data, assessing the disassembling of every feature's contribution. Hence, the reference is made only to significant aspects of the data while the rest of the least influential part is filtered out. We demonstrate an evolutionary algorithm approach (Grouping Genetic Algorithm or GGA) in classifying imbalanced RNA-Seq samples into different cancer classes that have different initial dataset sizes. Filter techniques are flexible in that they are computationally efficient, fast and cost-effective, and can thus be used in data that are big due to RNA-Seq. Table 4 summarizes some of the filter methods used in RNA-Seq feature engineering and provides the results of model prediction performance that takes advantage of the same respective filter methods.

Table.2. Different Feature selection methods in gene Expression Data

Reference	Feature Selection Method	Feature Selection Algorithm	Dataset Types	Accuracy (%)
Park et al., 2019 [61]	Filter Methods	Artificial Neural Network	RNA-Seq	90.71
Wu and Hicks, 2021 [62]	Filter Methods	Support vector machine Naïve Bayes Decision Tree K-nearest Neighbour	RNA-Seq	85 85 86 90
Liu and Yao, 2022 [63]	Filter Methods	Deep Neural Network	RNA-Seq	99
Mahin et al., 2022 [64]	Filter Methods	K-nearest neighbour	RNA-Seq	100
Liu et al., 2022 [65]	Wrapper Methods	Random Forest	RNA-Seq	99.68
Al Abir et al., 2022 [67]	Wrapper Methods	Support vector machine. SVM-RFE	RNA-Seq	99.93
Kong and Yu, 2018 [67]	Embedded Method	Graph Embedded Deep Feedforward Network	BRCA- RNA-Seq	94.50
Zhang and Liu, 2021 [68]	Embedded Method	Robust Biomarker Discovery	RNA-Seq	97-98-99
Abdelwahab et al., 2022 [49]	Embedded Method	Support vector machine. SVM-RFE	RNA-Seq	94
Coleto-Alcudia et al., 2022 [48]	Embedded Method	Filtering+ SVM	RNA-Seq	93

3.2 WRAPPER METHODS

Wrapper functions evaluate the role of features along with their reputation by using a classification method when doing so. Such approaches happen directly with the classifier and are used to discover a related group of characteristics that have the potential to be good predictors. Firstly, deep profiling of the gene subsets is performed by a selected classifier which is then used for retraining a classifier using the top significant genes only. The main type of feature subset selection methods used within wrapper methods is based on how well the learning algorithm of the model performs. In general, it can be said that wrapper methods are like containers in the sense that they package the learning algorithm into a search mechanism, whose goal is to find the most beneficial feature subset. In this approach, the "black box" learning algorithm works as an optimization metric, with its efficiency as the optimization metric. Wrapper methods can be further divided into deterministic or randomized approach groups. Well-known wrapper methods, for instance, k-nearest neighbours [69], random forests [70,71], support vector machine [72,73], and so on are different algorithms that have been integrated into classifiers for feature selection. Wrapper methods involve training and evaluations of numerous classifiers which cover different feature breakdowns, thus making them more complex and demanding more time as well as computer resources in contrast with the filter methods. Nevertheless, such outputs are most likely to result in greater performance.

3.3 ENSEMBLE METHODS

The embedded methods, designed as feature engineering algorithms, aim to optimize subset composition effectively by combining with a specific classifier. These methods are designed to utilize the strong sides of both types interchangeably, so that they may fit the subtleties of the given learning vector. In embedded learning methods usually do better than filter and wrapper methodologies, which is mainly caused by the capability to address the problem of feature interaction. Such factors occur due to certain subsets of genes interacting with other genes, which could take the feature selection process to locally optimal subsets and undermine better computational performance as a whole [74].

Top of Form

3.4 HYBRID METHODS

A variety of papers in the literature have analyzed hybrid methods, a collection of the individual strengths of these approaches. For example, first, filter methods can be used so that the feature set gets reduced before moving on with the wrapper approach. Finally, the feature importance will be generated by a classifier model, and it will determine the optimal gene subset among all [75]. This class of algorithms are a trade-off of computational requirements and end-game results. Similarly, ensemble methods such as Bagging, Boosting Ensembles, and Random Forests have since proven very flexible and enduring solutions for tackling feature interactions involved in high-dimensional, complex tasks. Through the application of an ensemble of weak classifiers that may work on a subset of the training data or feature inputs, the ensemble methods reduce overfitting tendencies and help to produce a more accurate predictive outcome in gene expression analysis [76].

3.5 BENEFITS & DRAWBACKS OF FEATURE ENGINEERING

The presentation of Table 5 provides a summary of feature-engineering techniques' benefits and drawbacks related to gene expression analysis. The line is drawn between the two types of filters: univariate which assesses each feature independently and multivariate which considers all the relationships between different features. The pre-existing knowledge of feature distributions along with wrapper methods when categorized, give us the advantages and drawbacks respectively. For gene expression analysis, deterministic models, which help represent the variability of features in the context of the predictive model, are used while the randomized methods operate without presuming any specific data distribution or factoring in feature fluctuations.

Table 3. Varying natures of the feature selection methods across different categories

Feature Selection	Filter Method	Wrapper Method	Embedded Method
Pros	**Univariate** Scalable to large datasets. Independent of the classifier.	**Deterministic** Interacts with the classifier simply. It takes less time to compute than randomised methods.	Interact with the classifier in a complex way. Feature Dependant. Less computational complexity Less prone to the feature than the wrapper method
	Multivariate Feature Dependencies. Independent of the classifier. Less computational complexity is than the wrapper method.	**Randomized** Interact with the classifier. Feature Dependencies. Less prone to the feature interaction problem	
Cons	**Univariate** Feature independent. Independent of the classifier.	**Deterministic** Highly prone to overfitting. Classifier dependant.	Dependant on Classifiers.

IV. METHODS FOR GENE EXPRESSION

Various ML methods have been applied in gene expression analysis to recognise potential cancers and incorporate insights into potential treatment methods.

4.1 TRADITIONAL MACHINE LEARNING MODEL

Different traditional machine learning algorithms, k-nearest Neighbour (kNN), Support Vector Machines (SVM), Random Forest (RF) and Naïve Bayes (NB) have been extensively adopted in research projects focused on early cancer detection [77, 78]. Like Segal and colleagues developed a genome centric SVM method especially for title discrimination [79]. Students applied the t-test to compress the genetic set to 256 genes which were then classified through a linear SVM classifier achieving the astonishing 75 correct in 76 cases of leave-one-out cross-validation.

Furthermore, it is often seen in research that ML methods with feature selection have been used effectively for prediction. Zhang et al. [87] used an SVM that found the best features in an RFE framework and used a constrained search to find the best parameters of the SVM, which was named SVM–RFE–PO. This method utilized grid search feature selection, Partial Swarm Optimization, and a genetic algorithm in a hybrid fashion choosing parameters during the feature selection step, resulting in an SVM method for storage cancer classification.

Moreover, a random forest ensemble is used to pick out the 237 genes and to this end, they were insusceptible to attack the model's predictive accuracy. Additionally, according to Hijazi et al., to target cancer subtype genes, they introduced a two-stage feature selection tool for attribute estimation and Genetic Algorithm base [81].

The model showed a high level of accuracy of 99.89% and 99.40% precision on certain specified cancer-type data coming from five separate datasets, yet its performance wasn’t kept constant when dealing with other types of cancer. The Evolutionary Programming-trained Support Vector Machine (EP-SVM) approach [81] aims to create a probabilistic SVM framework with separate classification features, and the probabilistic predictor outputs reflect the corresponding binary classifier probabilities. Here, the summary of past literature concerning machine learning methods used for gene expression profiling is given as Table 7.

Table.6. **Conventional ML approaches for gene expression analysis**

Reference	Dataset	Algorithm	Type of Dataset	Classification Accuracy
Segal et. al., 2003 [79]	Cancer	SVM	Gene Expression	98.5
Hijazi et. al., 2013 [81]	Mixed -Lineage Leukemia	SVM Linear	Gene Expression Data	99.89
Ram et al., 2017 [80]	Colon cancer	RF	Microarray Data	87.39
Zhang et al., 2018 [70]	Breast Cancer	SVM-RFE-PSO	Gene Expression Data	81.54
Yuan et al., 2020 [82]	Tumour-educated platelets	SVM	Gene Expression Data	95.93
Abdulqader et al., 2020 [83]	Lymphoma	KNN & NB	Microarray Data	94.7 & 74.83

Broadly speaking, machine learning (ML) algorithms excel at uncovering intricate patterns within intricate and multi-dimensional datasets across a variety of fields. Consequently, they've been particularly effective in analysing and categorizing gene expression data [82]. Nonetheless, the efficacy of traditional ML algorithms is largely contingent upon the calibre of the input features. As such, their success hinges on the effectiveness of the integrated feature selection techniques.

4.2 DEEP LEARNING TECHNIQUE

Deep learning techniques implement artificial NN with several stages of processing elements which enable obtaining data representations instead of input directly. This category of learning excels in capturing the hierarchy relationships observed within large inputs and therefore, is a major advantage over traditional ML approaches. With such distinctive capabilities, then they are opening door to futuristic methods for gene expression analysis [85]. Applying the most popular NN architectures, which include fully connected NN (multi-layer perceptron NN), convolutional NN (CNN), recurrent NN (RNN), graph NN (GNN), and transformer NN (TNN)), is an example demonstration [31].

4.2.1 Multi-Layer Perceptron (MLP) Neural Networks

The Multilayer Perceptron (MLP) represents a neural network structure characterized by fully connected layers, ensuring each neuron within a hidden layer connects to every other neuron in adjacent layers. In the realm of gene expression analysis, MLP classifiers have been at the forefront of cancer classification efforts.

For example, Lai et al. [86] devised an MLP model amalgamating various gene expression and clinical data sources to forecast the overall survival rates of non-small cell lung cancer (NSCLC) patients. By integrating 15 biomarkers with clinical insights, they crafted an integrative MLP classifier via bimodal learning, achieving notable metrics like an AUC of 0.8163 and an accuracy of 75.44%.

In another study, Zhang et al. [87] introduced an unsupervised feature learning paradigm that amalgamated principal component analysis (PCA) with an autoencoder-based MLP model to discern diverse attributes from gene expression profiles. Utilizing an ensemble classifier dubbed PCA-AE-Ada, they predicted clinical outcomes in breast cancer cases.

Gao et al. [88] innovated with the Deep Cancer Subtype Classification (DeepCC) method, emphasizing supervised cancer categorization grounded in functional spectra analysis, indicative of biological pathway activities. By employing multilayer neural networks instead of manually curated features and leveraging enrichment analysis, they achieved over 90% balanced accuracy in classifying breast and colorectal cancers.

Similarly, Chandrasekar et al. [89] leveraged an MLP-centric classification strategy, prioritizing accuracy in predicting cancer severity and identifying the illness using a compact set of gene subsets. Their focus was on optimizing predictions concerning the disease's severity.

Lastly, Laplante et al. [90] tailored an MLP framework to distinguish cancers across 20 distinct anatomical regions by leveraging miRNA stem-loop data sets. With an initial layer boasting 1,046 input neurons corresponding to individual miRNAs and a concluding layer identifying 27 cancer types, their model realized an impressive average accuracy of 96.9%.

4.2.2 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) stand out within the neural network category due to their inherent recurrent connections among neuron units, endowing them with memory capabilities. This intrinsic memory enables RNNs to leverage past observations to make sense of current or even forecast future observations within a sequence. Such properties equip RNNs with dynamic sequential processing abilities, ideal for analysing sequential data and discerning intricate relationships and trends.

Sahin et al. innovated with an RNN framework that integrated a Long Short-Term Memory (LSTM) network with the Artificial Immune Recognition System (AIRS) to devise a stability mechanism for robust microarray dataset feature selection, securing an accuracy of 89.6% [91]. Aher et al. introduced the RCO-RNN, harnessing the rider chicken optimization (RCO) technique to extract pertinent genes from gene expression data. This method showcased an impressive 95% accuracy across datasets like the Leukemia database, Small Blue Round Cell Tumor (SBRCT), and Lung Cancer Dataset.

Majji et al [92]. put forth the JayaALO-based DeepRNN, a cutting-edge technique for automated cancer prediction that melded the Jaya ant lion optimization (ALO) with an RNN structure [93]. Their method exhibited peak accuracy, reaching 95.97% across varied datasets such as AP Colon Kidney, AP Breast Ovary, and others.

In related research domains, Suresh et al. crafted an innovative strategy to interpret genome sequencing by fusing the bat sonar algorithm with the LSTM model for disease detection [94]. RNNs, particularly LSTM recurrent structures, have been recurrently deployed in various studies. They've been instrumental in identifying genes associated with tumour diagnosis, pinpointing breast cancer cells, distinguishing between cancerous and healthy cells, and recognizing biological entities [95-98]. Furthermore, Zhao et al. pioneered an RNN-centric model targeting transcriptional target factor identification [99]. Other techniques, such as the memetic approach, have been utilized to fine-tune RNN parameters, while methodologies like LASSO-RNN have been instrumental in reconstructing gene regulatory networks (GRNs) [100]. A consolidated overview of recent RNN-centric studies is encapsulated in Table 8.

RNN architectures offer several benefits for gene expression analysis, notably enhancing efficiency by capturing and preserving sequential feature details [101]. Furthermore, these networks can flexibly adapt to evolving dynamics in uncertain systems, like scenarios where the importance of genetic data might shift over time. However, it's worth noting some limitations of RNNs in this context. They tend to have extended processing durations compared to methods like CNNs, leading to more prolonged and intricate training processes. Moreover, when dealing with extended genomic sequences, RNNs may exhibit challenges in capturing dependencies as effectively as techniques such as GNN and TNN [102, 103].

4.2.3 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) were initially crafted for image analysis, utilizing convolutional filters to autonomously discern spatial features within input data. These architectures integrate stacked convolutional layers with pooling layers, often supplemented with regularization layers like Batch Normalization or Dropout [104]. In the realm of gene expression analysis, a CNN-centric ensemble method is introduced, achieving a remarkable 98% precision across three distinct cancer RNA-Seq datasets, including Lung Adenocarcinoma, Stomach Adenocarcinoma, and Breast Invasive Carcinoma [105]. Other studies [18,106] leveraged CNNs to classify tumour categories by translating high-dimensional RNA-Seq data into 2D image representations [18, 106]. For instance, Elbashir et al. [18] proposed a streamlined CNN model for breast cancer classification, yielding an accuracy of 98.76%. Additionally, an investigation employed three distinct CNN models on a vast dataset spanning 33 cancer types, showcasing the 1D-CNN model's proficiency in predicting breast cancer subtypes with an 88.42% precision [107].

CNN architectures, spanning 1D to 3D convolutions, have been tailored for gene expression data analysis. One-dimensional convolutions adeptly handle genomic sequences, capturing sequential patterns. Meanwhile, 2D convolutions are adept at processing gene expression data by transmuting them into image formats. However, one challenge is the potential loss of information when discretizing continuous gene expression values into colour palettes for image creation. An alternative method bifurcates the image creation process: first converting a biological functional hierarchy into an image template, and subsequently mapping gene expressions onto this template, thereby retaining the continuous expression values without the pitfalls of discrete colour mapping [108]. Overall, CNNs have demonstrated superior efficacy in gene expression analysis compared to RNNs (referenced in Table 9), credited to their prowess in swiftly and accurately processing vast genetic datasets, and adeptly extracting pivotal information from both local and global gene expression features [36,109].

4.2.4 Graph Neural Networks (GNN)

Graph Neural Networks (GNNs) are deep learning frameworks tailored for analyzing data represented in graph formats, characterized by vertices (nodes) and connections (edges) [110]. These networks disseminate feature information across nodes, facilitating the acquisition of context-specific features by examining relationships between objects and entities within the graph. The core principle involves iteratively aggregating and transforming neighbouring node features, thereby generating updated node embeddings. In biological contexts, nodes typically represent genes, transcripts, or proteins, while edges symbolize experimentally determined functional or similarity connections. For instance, in generating network graphs from gene expression data, correlation coefficients, such as Pearson's, assess the similarity between gene expression profiles to establish graph edges [111].

GNNs are instrumental in multi-omics pan-cancer data analysis, encompassing gene expression patterns, DNA methylation, gene mutations, and clinical metrics, with a focus on cancer prediction [25]. Pfeifer et al. pioneered an interpretable GNN framework targeting cancer subnetwork identification by leveraging patient-specific protein-protein interaction (PPI) network topologies enriched with multi-omics data [111]. This approach enhances subnetwork detection within disease contexts. Concurrently, other research endeavours to harness GNNs to predict cancer types and identify oncological markers using multi-omics and PPI networks [113]. Zhou et al. utilized gene interaction networks to predict cancer across multi-dimensional omics datasets via graph convolutional networks (GCNs) [113]. Their method integrated contour-aware information aggregation and gated graph attention mechanisms [26], improving semantic understanding of gene-molecular function relationships and achieving high diagnostic accuracy [115].

GNNs offer distinct advantages in gene expression data analysis, leveraging inherent capabilities to propagate and aggregate attributes, and capturing intricate cellular relationships evident in RNA-Seq data graphs. These networks adeptly aggregate cell-cell relationships, leveraging domain expertise in gene regulation to address data gaps [116] By amalgamating topological neighbour propagation, GNNs facilitate the construction of gene regulatory networks (GRNs), augmenting predictive prowess [117]. However, GNNs can be susceptible to data noise during graph structure formulation [118].

4.2.5 Transformer Neural Networks (TNN)

Temporal Neural Networks (TNNs) utilize network structures grounded in the multi-head self-attention mechanism, enabling them to discern long-distance dependencies within sequences [119]. This methodology excels in handling sequential data types like genomic sequences, time series, acoustic signals, and natural language data. Within gene expression analysis, TNN stands out due to its capability to concurrently process information from diverse representation facets across genomic sequences. For instance, the Gene Transformer model [27] harnesses multi-head self-attention components to navigate the intricacies of high-dimensional gene expression, identifying pertinent biomarkers across diverse cancer subtypes. Similarly, the multi-omic transformer, inspired by Osseni et al.'s architecture [28], adeptly discerns intricate phenotypes (specific cancer types) by amalgamating transcriptomic, epigenomic, copy number variation, and proteomic data types. Additionally, Lv et al. [10] devised a transformer-centric fusion network, PG-TFNet, merging pathological imagery with genomic data for nuanced cancer survival predictions. This transformer-centric approach facilitates insightful intra-modality relationship exploration across varied pathological slide perspectives.

In the realm of performance metrics, TNN models have showcased heightened resilience compared to both CNN and RNN counterparts, delivering commendable outcomes across diverse data schemas. The inherent self-attention mechanism empowers TNNs to harness contextual cues from any sequence segment, effectively capturing extensive dependencies—a feat less efficiently achieved by CNNs and enhancing parallel processing relative to RNNs [121]. However, TNNs come with their set of challenges, notably their appetite for expansive datasets; consequently, their efficacy might wane when handling genetic datasets with constrained sample sizes, potentially trailing behind other neural network paradigms [122].

4.3 TRANSFER LEARNING

Transfer learning, as outlined in, seeks to enhance the performance of downstream models by leveraging information from distinct but related source domains [130]. Kakati et al. applied transfer learning to the DEGnext CNN model, utilizing knowledge representation from feature maps to predict significant up-regulated (UR) and down-regulated (DR) genes in untrained cancer datasets obtained from The Cancer Genome Atlas database [122]. Similarly, Das et al. employed transfer learning with 2D CNN models on spectrogram images of digital DNA sequences for automated recognition of liver cancer genes [24].

In a different application, Zhang et al. fine-tuned a Convolutional Long Short-Term Memory network (CLSTM) through transfer learning to model temporal genetic information in dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) of cancer genes [124]. A Deep Transfer Learning (DTL) framework that utilized individual cell information without employing profiling or reduction methods, resulting in a 30% acceleration of the process and improved performance. The characteristics of various transfer learning approaches are detailed in Table.5.

Table.5. Transfer Learning based on DL methods for gene expression data analysis.

Reference	Type of Cancer	Algorithm	Dataset Type	Accuracy (%)
Lopez-Garcia et al., 2020 [108]	Lung Cancer	Transfer Learning with CNN	Gene expression Data	73.26
Zhang et al., 2021 [124]	Breast Cancer	Transfer Learning with CNN & CLSTM	Breast Cancer	With CNN 90 With CLSTM 93
Kakati et al., 2022 [123]	Uterine and breast cancer	Transfer Learning with CNN	Gene expression Data	AUC:99
Das et al., 2022 [24]	Liver Cancer	Deep Transfer Learning	Gene Sequence	98.86

4.4 PATHWAY ANALYSIS

Pathway analysis serves as a pivotal method for gleaning biological insights from extensive gene expression datasets. The primary objective of these methodologies is to discern which specific pathways might be disrupted owing to distinct gene expression patterns [125]. For example, the adipocytokine signalling pathway has proven instrumental in distinguishing between breast cancer and tumours of the colon and stomach [126]. Such techniques are instrumental in pioneering bioinformatics tools, empowering scientists to unravel the genetic and pathway alterations intrinsic to various cancer types and identify potential therapeutic avenues.

Over the preceding decades, a plethora of pathway analysis techniques have emerged, categorized into three distinct generations based on their temporal evolution and methodological approach. The initial two epochs encompass over-representation analysis (ORA) and functional class scoring (FCS), both of which treat pathways as collections of genes. Contrarily, the third generation, known as topology-based (TB) pathway analysis, elevates the analytical approach by integrating pathway topology for enhanced accuracy and performance [127, 128, & 129]. Within the realm of TB pathway analysis, the methodology amalgamates two pivotal data facets: genes differentially expressed within a pathway and supplemental biological insights concerning the magnitude and positioning of alterations across all differentially expressed genes, their interrelationships, and interaction dynamics. Esteemed biological pathway repositories like KEGG and Reactome leverage meticulously curated insights spanning years to delineate the spatial arrangements and interplay among genes within specific pathways [130, 131].

V. FUTURE DIRECTIONS

This section discusses future directions that may potentially advance the research on ML-based gene expression analysis.

In the realm of future research, expanding the types of input features in conjunction with established learning algorithms is paramount, as gene expression's complete impact transcends merely genetic sequences. For example, DNA methylations and mutations present viable feature types for refining cancer classification. These DNA methylations, observed at CpG dinucleotides and non-CpG sites, play crucial roles in both normal developmental processes and pathological alterations, including gene silencing of tumour suppressors and DNA repair genes. Integrating such methylations and mutations with RNA-Seq data can furnish valuable features enhancing tumour classification accuracy.

Simultaneously, the meticulous design of computational algorithms stands pivotal. Researchers can pioneer techniques tailored for optimal performance on benchmark datasets like the unique molecular identifier (UMI), underpinned by experimentally validated reference genes. These endeavours can foster comparative analyses of single-cell methods and algorithmic efficacy assessments across protocols like SMART-Seq, Cel-Seqs, and droplets.

A prospective research avenue entails pinpointing cancer-centric biomarkers, leveraging methodologies such as IntPath for functional pathway analyses of relevant genes. Deep learning (DL) techniques, applied to 2D images, hold promise for discerning cancer-specific biomarkers. Graph neural networks (GNN) can facilitate the integration of single-cell multi-omics data through heterogeneous graph structures, encompassing technologies like Droplet scRNA-Seq and Smart-Seq2, elucidating cell-type-specific regulatory mechanisms and T-cell ancestries.

A notable emphasis must be on architecting interpretable ML models, elucidating their decision-making rationales and potential pitfalls. Models elucidating both local and global ML properties based on counterfactuals or feature attributions warrant heightened exploration. Recent endeavours, paralleling genomic assessments of pre-malignant lesions within The Cancer Genome Atlas (TCGA) project, underscore the integration of diverse modalities—from imaging to proteomic and epigenetic—to delineate surrogate cancer gene prevention biomarkers. This comprehensive approach can revolutionize cancer diagnosis, treatment, and patient monitoring.

Analysing multidomain genomic data emerges as another critical research trajectory, harnessing multimodal and multitask ML methods, such as early and late fusion strategies, to potentially outperform extant methodologies. Addressing the nuances between clinically akin cancers necessitates a granular focus on genomic and transcriptomic variations. Techniques like optical genome mapping and structural variant analysis can be pivotal across diverse cancer datasets, refining prognosis and therapeutic interventions.

Lastly, delving deeper into circRNA dynamics—its localization, transportation, degradation, and a comprehensive interactome—alongside single-cell profiling, promises insights pivotal for refining cancer gene prediction methodologies [1].

Table.6. The prospective avenues for the application of machine learning-based methods in gene expression data

New Features	**Innovation in computational algorithm**	**Finding Biomarkers**	**Incorporating graph networks to integrate single-cell multi-omics data.**	**Create approaches that are interpretable and provide clear explanations.**	**Develop approaches that involve multi-modal and multi-task learning.**
Enhancing the discriminative performance of established learning algorithms can be achieved by incorporating supplementary input features, such as DNA methylations and mutations.	Advancing gene expression analysis relies on the crucial development of innovative computational algorithms and benchmarking approaches.	Explore techniques for pinpointing biomarkers unique to each type of cancer.	Graph Neural Network (GNN) structures can facilitate the fusion of single-cell multi-omics data through the utilization of heterogeneous graphs.	Highlight the importance of embracing interpretable machine learning models that aid in comprehending the decision-making process and offer explanations for instances when these models encounter failures.	The utilization of multimodal and multitask machine learning methods, employing both early and late fusion strategies, holds the promise of enhancing classification performance.

VI. CONCLUSIONS

Recent progress in deep learning offers significant promise for the analysis of intricate, high-dimensional datasets, especially in the realm of multi-omics data analytics. This review delves into the advancements made in applying both conventional machine learning and deep learning methodologies to analyse gene expression patterns, utilizing RNA-sequencing and DNA microarray data for detecting cancer. We provide an overview of prevalent data acquisition techniques in gene expression studies and highlight widely recognized datasets crucial for supervised machine learning tasks. Furthermore, the manuscript delineates a taxonomy encompassing techniques pertinent to feature engineering and data preprocessing—essential facets of gene expression analysis.

The discussion transitions to machine learning-centric methodologies for gene expression evaluation, spotlighting the merits of deep learning techniques given their pronounced efficacy in this domain. We delve into prior endeavours utilizing neural network paradigms, including multi-layer perceptron, convolutional, recurrent, graph, and transformer networks. Notably, the employment of deep learning in categorizing cancers using RNA-Seq data has exhibited commendable accuracy, with numerous research endeavours achieving notable success rates across diverse cancer types. Anticipating the trajectory of this field, it becomes imperative to address existing hurdles like result generalizability, resilience, and interpretability, thereby catalysing advancements in cancer diagnosis and patient care.

Significantly, this study stands out by offering an exhaustive review encapsulating recent endeavours in cancer classification via gene expression analysis. It bridges the understanding of feature engineering methodologies, pivotal datasets in this domain, and the application spectrum of both conventional and cutting-edge machine learning techniques. A distinctive facet of this review lies in spotlighting contemporary neural network architectures—specifically graph and transformer networks—that remain underrepresented in extant literature. Additionally, we underscore the predominant role of RNA-Seq methodologies, marking them as the prevailing data modality in contemporary research endeavours.

REFERENCES

[1] Miller K.D, Ortiz A.P, Pinheiro P.S, Bandi P, Minihan A, Fuchs H.E, Martinez Tyson D, Tortolero-Luna G, Fedewa S.A. and Jemal A.Metal, "Cancer Statistics for the US Hispanic/Latino Population", *CA A Cancer J. Clin,* Vol. 71, pp. 466–487,2021.

[2] Munkácsy G, Santarpia L. and Gy″orffy B, "Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Tumor Microenvironment Features*", Biomedicines,* Vol. 10, pp. 248, 2022.

[3] Brewczy ´nski A, Jabło ´nska B, Mazurek A.M, Mrochem-Kwarciak J, Mrowiec S, Snietura M, Kentnowski M, Kołosza Z, ´ Składowski K. and Rutkowski T, "Comparison of Selected Immune and Hematological Parameters and Their Impact on Survival in Patients with HPV-Related and HPV-Unrelated Oropharyngeal Cancer". *Cancers,* Vol. 13, pp. 3256, 2021.

[4] Ahmed Z, Mohamed K, Zeeshan S. and Dong X, "Artificial Intelligence with Multi-Functional Machine Learning Platform Development for Better Healthcare and Precision Medicine." Database 2020, 2020, baaa010.

[5] Anna A. and Monika G, "Splicing Mutations in Human Genetic Disorders Examples, Detection, and Confirmation" *J. Appl. Genet* ,Vol. 59,pp. 253–268, 2018.

[6] Slatko B.E , Gardner A.F. and Ausubel F.M ,"Overview of Next-Generation Sequencing Technologies". *Curr. Protoc. Mol. Biol*,Vol. 122, cpmb.59, 2018.

[7] Briglia N, Petrozza A, Hoeberichts F.A, Verhoef . N. and Povero. G, "Investigating the Impact of Biostimulants on the Row Crops Corn and Soybean Using High-Efficiency Phenotyping and Next Generation Sequencing", *Agronomy* ,Vol.9, pp. 761,2019.

[8] Phan T, Fay E.J, Lee Z, Aron S, Hu W.S. and Langlois R.A, "Segment-Specific Kinetics of MRNA, CRNA, and VRNA Accumulation during Influenza Virus Infection". *J. Virol*. Vol 95, pp. e02102-20,2021.

[9] Monaco G, Lee B, Xu W, Mustafah S, Hwang Y.Y, Carré C, Burdin N, Visan L, Ceccarelli M. and Poidinger Metal, "RNA-Seq Signatures Normalized by MRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types*", Cell Rep*. Vol. 26, pp.1627–1640.e7, 2019.

[10] Lunshof J.E, Bobe J, Aach. J, Angrist M, Thakuria J.V, Vorhaus D.B, Hoehe M.R. and Church G.M, "Personal Genomes in Progress: From the Human Genome Project to the Personal Genome Project", *Dialogues Clin. Neurosci,*Vol.12,pp. 47–60,2010.

[11] Khan M.F, Ghazal T.M, Said R.A, Fatima A, Abbas S, Khan M.A, Issa G.F, Ahmad M. and Khan M.A, "An IoMT-Enabled Smart Healthcare Model to Monitor Elderly People Using Machine Learning Technique". *Comput, Intell, Neurosci,* Vol,pp. 2487759,2021.

[12] Bhonde S.B. and Prasad J.R, "Deep Learning Techniques in Cancer Prediction Using Genomic Profiles. In Proceedings of the 2021 6th International Conference for Convergence in Technology (I2CT)", *Maharashtra, India,*Vol. 2–4, pp. 1–9, April 2021.

[13] Celesti F, Celesti A, Wan J. and Villari M," Why Deep Learning Is Changing the Way to Approach NGS Data Processing" *A Review. IEEE Rev. Biomed. Eng,* Vol. 11, pp.68–76,2018.

[14] Alomari O.A, Khader A.T, Al-Betar M.A. and Alkareem Alyasseri Z.A, "A Hybrid Filter-Wrapper Gene Selection Method for Cancer Classification. In Proceedings of the 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS)", *Kuching, Malaysia,* Vol. 113–118, pp. 24–26,2018.

[15] Del Amor R, Colomer A, Monteagudo C. and Naranjo V, "A Deep Embedded Refined Clustering Approach for Breast Cancer Distinction Based on DNA Methylation". *Neural Comput. Applic*. Vol. 34, pp.10243–10255,2022.

[16] Zhou J.R, You Z.H, Cheng L. and Ji B.Y, "Prediction of LncRNA-Disease Associations via an Embedding Learning HOPE in Heterogeneous Information Networks". *Mol. Ther. Nucleic Acids* , Vol. 23,pp. 277–285,2021.

[17] Ravindran U. and Gunavathi C, "A Survey on Gene Expression Data Analysis Using Deep Learning Methods for Cancer Diagnosis". *Prog. Biophys. Mol. Biol.* S0079610722000803, 2022.

[18] Elbashir M.K, Ezz M, Mohammed M. and Saloum S.S, "Lightweight Convolutional Neural Network for Breast Cancer Classification Using RNA-Seq Gene Expression Data*". IEEE Access* ,Vol.7, pp.185338–185348, 2019.

[19] Monti M, Fiorentino J, Milanetti E, Gosti G. and Tartaglia G.G, "Prediction of Time Series Gene Expression and Structural Analysis of Gene Regulatory Networks Using Recurrent Neural Networks". *Entropy* , Vol.24, pp.141,2022.

[20] Bar Joseph Z, Gitter A. and Simon I, "Studying and Modelling Dynamic Biological Processes Using Time-Series Gene Expression Data". *Nat. Rev. Genet.* 2012, Vol.13, pp. 552–564, 2012.

[21] Lee H.J, Chung Y, Chung K.Y, Kim Y.K, Lee J.H, Koh Y.J. and Lee S.H, "Use of a Graph Neural Network to the Weighted Gene Co-Expression Network Analysis of Korean Native Cattle". *Sci. Rep,* Vol.12, pp. 9854, 2022.

[22] Lee D, Yang J. and Kim S, "Learning the Histone Codes with Large Genomic Windows and Three-Dimensional Chromatin Interactions Using Transformer". *Nat. Commun.* , Vol.13, pp. 6678, 2022.

[23] Kim H.E, Cosa-Linan A, Santhanam N, Jannesari M, Maros M.E. and Ganslandt T, "Transfer Learning for Medical Image Classification: A Literature Review. BMC Med". *Imaging,* Vol.22, pp. 69, 2022 .

[24] Das B. and Toraman S, "Deep Transfer Learning for Automated Liver Cancer Gene Recognition Using Spectrogram Images of Digitized DNA Sequences". *Biomed. Signal Process. Control* , Vol. 72, pp. 103317, 2022.

[25] Chereda H, Bleckmann A, Menck K, Perera Bel J, Stegmaier P, Auer F, Kramer F, Leha A. and Beißbarth T, "Explaining Decisions of Graph Convolutional Neural Networks: Patient-Specific Molecular Subnetworks Responsible for Metastasis Prediction in Breast Cancer". *Genome Med*, Vol.13, pp. 42,2021.

[26] Qiu L, Li H, Wang M. and Wang X," Gated Graph Attention Network for Cancer Prediction". *Sensors* , Vol. 21, pp. 1938,2021.

[27] Zhang T.H, Hasib M.M, Chiu Y.C, Han Z.F, Jin Y.F, Flores M, Chen Y. and Huang Y. "Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions". *Cancers* , Vol.14, pp. 4763, 2022.

[28] Osseni M.A, Tossou P, Laviolette F. and Corbeil J, "MOT: A Multi-Omics Transformer for Multiclass Classification Tumour Types Predictions". *BioRxiv* 2022.

[29] Sathe S, Aggarwal S. and Tang. J, "Gene Expression and Protein Function: A Survey of Deep Learning Methods". *SIGKDD Explor. Newsl*, Vol.21, pp. 23–38, 2019.

[30] Koumakis L, "Deep Learning Models in Genomics; Are We There Yet? Comput. Struct". *Biotechnol. J*, Vol.18, pp. 1466–1473, 2020.

[31] Zhu W, Xie L, Han J. and Guo X, "The Application of Deep Learning in Cancer Prognosis Prediction". *Cancers* , Vol.12, pp. 603,2020.

[32] Gunavathi C, Sivasubramanian K, Keerthika P. and Paramasivam C, "A Review on Convolutional Neural Network Based Deep Learning Methods in Gene Expression Data for Disease Diagnosis". *Mater. Today Proc*, Vol.45, pp. 2282–2285,2021.

[33] Tabares Soto R, Orozco Arias S, Romero Cano V, Segovia Bucheli V, Rodríguez Sotelo J.L. and Jiménez Varón C.F, "A Comparative Study of Machine Learning and Deep Learning Algorithms to Classify Cancer Types Based on Microarray Gene Expression Data". *PeerJ Comput. Sci*, Vol.6, e270,2020.

[34] Mazlan A.U, sahabudin N.A, Remli M.A, Ismail N.S.N, Mohamad M.S, Nies H.W. and Abd Warif N.B, "A Review on Recent Progress in Machine Learning and Deep Learning Methods for Cancer Classification on Gene Expression Data." *Processes* , Vol.9, pp. 1466, 2021.

[35] Karim M.R, Beyan O, Zappa A, Costa I.G, Rebholz Schuhmann D, Cochez M. and Decker S, "Deep Learning-Based Clustering Approaches for Bioinformatics. Brief ". *Bioinform*, Vol.22, pp. 393–415,2021.

[36] Thakur T, Batra I, Luthra M, Vimal S, Dhiman G, Malik A. and Shabaz M, "Gene Expression-Assisted Cancer Prediction Techniques". *J. Healthc. Eng*, Vol. 2021, pp. 643–648,2021.

[37] Montesinos López O.A, Montesinos López A, Pérez Rodríguez P, Barrón López J.A, Martini J.W.R, Fajardo Flores S.B, Gaytan Lugo L.S, Santana Mancilla P.C. and Crossa J, "A Review of Deep Learning Applications for Genomic Selection". *BMC Genom*, Vol.22, pp. 19,2021.

[38] Bhandari N, Walamb R, Kotecha K. and Khare S.P, "A Comprehensive Survey on Computational Learning Methods for Analysis of Gene Expression Data". *Front. Mol. Biosci*, Vol.9, pp. 907150, 2022.

[39] Khalsan M, Machado L.R, Al Shamery E.S, Ajit S, Anthony K, Mu M. and Agyeman M.O, "A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction". *IEEE Access* , Vol.10, 27522–27534, 2022.

[40] Alhenawi E, Al Sayyed R, Hudaib A. and Mirjalili S, "Feature Selection Methods on Gene Expression Microarray Data for Cancer Classification: A Systematic Review". *Comput. Biol. Med*, Vol.140, pp. 105051, 2022.

[41] Hu T, Chitnis N, Monos D. and Dinh A, "Next-Generation Sequencing Technologies: An Overview*". Hum. Immunol*, Vol.82, pp. 801–811, 2021.

[42] Jungjit S, Michaelis M, Freitas A.A. and Cinatl J, "Extending Multi-Label Feature Selection with KEGG Pathway Information for Microarray Data Analysis. In Proceedings of the 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology", *Honolulu, HI, USA*, Vol.21–24, pp. 1–8, May 2014.

[43] Wang Y, Mashock M, Tong Z. Mu X, Chen H, Zhou X, Zhang H, Zhao G, Liu B. and Li X, "Changing Technologies of RNA Sequencing and Their Applications in Clinical Oncology". *Front. Oncol*, Vol.10, pp. 447,2022.

[44] Das S, Rai A, Merchant M.L, Cave M.C. and Rai S.N, "A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies". *Genes* , Vol.12, pp. 1947, 2021.

[45] Mohammed M, Mwambi H, Mboya I.B, Elbashir M.K. and Omolo B, "A Stacking Ensemble Deep Learning Approach to Cancer Type Classification Based on TCGA Data.", *Sci. Rep*, Vol.11, pp. 15626, 2021.

[46] Li S, Xu X, Zhang R. and Huang Y, " Identification of Co-Expression Hub Genes for Ferroptosis in Kidney Renal Clear Cell Carcinoma Based on Weighted Gene Co-Expression Network Analysis and The Cancer Genome Atlas Clinical Data." *Sci. Rep.*, Vol.12, pp. 4821, 2022.

[47] Zhang G, Peng Z, Yan C, Wang J, Luo J. and Luo H, "A Novel Liver Cancer Diagnosis Method Based on Patient Similarity Network and DenseGCN*". Sci. Rep,* Vol.12, pp. 6797, 2022.

[48] Coleto Alcudia V. and Vega Rodríguez M.A, "A Multi-Objective Optimization Approach for the Identification of Cancer Biomarkers from RNA-Seq Data". *Expert Syst. Appl*, Vol.193, pp. 116480, 2022.

[49] Abdelwahab O, Awad N, Elserafy M. and Badr E, "A Feature Selection-Based Framework to Identify Biomarkers for Cancer Diagnosis: A Focus on Lung Adenocarcinoma". *PLoS ONE* , Vol.17, e0269126, 2022.

[50] Houssein E.H, Abdelminaam D.S, Hassan H.N, Al Sayed M.M. and Nabil E, “A Hybrid Barnacles Mating Optimizer Algorithm With Support Vector Machines for Gene Selection of Microarray Cancer Classification”. *IEEE Access* , Vol.9, pp. 64895–64905, 2021.

[51] Hira S. and Bai A, “A Novel Map Reduced Based Parallel Feature Selection and Extreme Learning for Micro Array Cancer Data Classification”. *Wirel. Pers. Commun*, Vol.123, pp. 1483–1505,2022.

[52] Vaiyapuri T, Liyakathunisa Alaskar H, Aljohani E, idevi S. and Hussain A, “Red Fox Optimizer with Data-Science-Enabled Microarray Gene Expression Classification Model”. *Appl. Sci*, Vol.12, pp. 4172, 2022. [CrossRef]

[53] Ke L, Li M, Wang L, Deng S, Ye J. and Yu X, “Improved Swarm-Optimization-Based Filter-Wrapper Gene Selection from Microarray Data for Gene Expression Tumor Classification”. *Pattern Anal. Applic*. 2022. [CrossRef]

[54] Deng X, Li M, Deng S. and Wang L, “Hybrid Gene Selection Approach Using XGBoost and Multi-Objective Genetic Algorithm for Cancer Classification”. *Med. Biol. Eng. Comput*, Vol.60, pp. 663–681, 2022.

[55] Hira Z.M. and Gillies D.F , “A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data”. *Adv. Bioinform.* 2015, Vol, pp. 198363, 2015.

[56] Swarna Priya R.M, Maddikunta P.K.R, Panimala M, Koppu S, Gadekallu T.R, Chowdhary C.L. and Alazab M,” An Effective Feature Engineering for DNN Using Hybrid PCA-GWO for Intrusion Detection in IoMT Architecture”. *Comput. Commun*, Vol.160, pp. 139–149, 2020.

[57] Chumerin N. and Van Hulle M, “Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information. In Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing”, *Maynooth, Ireland,* Vol.6–8 September 2006; pp. 343–348, .

[58] Xie W, Fang Y, Yu K, Min X. and Li W, “MFRAG: Multi-Fitness RankAggreg Genetic Algorithm for Biomarker Selection from Microarray Data*”. Chemom. Intell*. Lab. Syst, Vol.226, pp. 104573, 2022.

[59] Khalid S, Khalil T. and Nasreen S, “A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning.” In Proceedings of the 2014 Science and Information Conference, London, UK, 27-29 August 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 372–378.

[60] Abd Elnaby M, Alfonse M. and Roushdy M, “Classification of Breast Cancer Using Microarray Gene Expression Data: A Survey”. *J. Biomed. Inform.*, Vol.117, pp. 103764, 2021.

[61] Park S, Shin B, Sang Shim W, Choi Y, Kang K. and Kang K. Wx, “A Neural Network-Based Feature Selection Algorithm for Transcriptomic Data”. Sci. Rep, Vol.9, pp. 10500, 2019.

[62] Wu J. and Hicks C, “Breast Cancer Type Classification Using Machine Learning”. *JPM,* Vol.11, pp. 61, 2021.

[63] Liu S. and Yao W, “Prediction of Lung Cancer Using Gene Expression and Deep Learning with KL Divergence Gene Selection”. *BMC Bioinform*, Vol.23, pp. 175, 2022.

[64] Mahin K.F, Robiuddin M, Islam M , Ashraf S, Yeasmin F. and Shatabda S, “PanClassif: Improving Pan Cancer Classification of Single Cell RNA-Seq Gene Expression Data Using Machine Learning”. *Genomics* , Vol.114, pp. 110264,2022.

[65] Liu H.P, Wang D. and Lai H.M, “Can We Infer Tumor Presence of Single Cell Transcriptomes and Their Tumor of Origin from Bulk Transcriptomes by Machine Learning*”? Comput. Struct. Biotechnol. J,* Vol.20, pp. 2672–2679, 2022.

[66] Al Abir F, Shovan S.M, Hasan M.A.M, Sayeed A. and Shin J, “Biomarker Identification by Reversing the Learning Mechanism of an Autoencoder and Recursive Feature Elimination”. *Mol. Omics* , Vol. 18, pp. 652–661, 2022.

[67] Kong Y. and Yu T, “A Graph-Embedded Deep Feedforward Network for Disease Outcome Classification and Feature Selection Using Gene Expression Data”. *Bioinformatics*, Vol.34, pp. 3727–3737, 2018.

[68] Zhang Z. and Liu Z.P, “Robust Biomarker Discovery for Hepatocellular Carcinoma from High-Throughput Data by Multiple Feature Selection Methods”. *BMC Med. Genom*, Vol.14, pp. 112, 2021.

[69] Li Y, Kang K, Krahn J.M, Croutwater N, Lee K, Umbach D.M. and Li L, “A Comprehensive Genomic Pan-Cancer Classification Using the Cancer Genome Atlas Gene Expression Data*”. BMC Genom* , Vol.18, pp. 508, 2017.

[70] Zhang Y, Deng Q, Liang W. and Zou X, “An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data”. *BioMed Res. Int*, Vol. 2018, pp. 7538204,2018.

[71] Al Obeidat F, Rocha Á, Akram M, Razzaq S. and Maqbool F, “(CDRGI)-Cancer Detection through Relevant Genes Identification”. *Neural Comput Applic* , Vol.34, pp. 8447–8454, 2022.

[72] Perera, H., & Costa, L. (2023, July 28), “Personality Classification of Text Through Machine Learning And Deep Learning: A REVIEW” (2023). *International Journal for Research in Advanced Computer Science and Engineering*, *9*(4), 6–12. https://doi.org/10.53555/cse.v9i4.2266

[73] Al-Haddad , K. M. ., Ba-Break, M., Al-Jamrah , K. ., & Al Amad, M. . (2021), “Knowledge, Attitude and Practice of Gynecologists at Public Teaching Hospitals in Sana’a City Towards Cervical Cancer Screening, -Yemen, 2020.” *International Journal For Research In Health Sciences And Nursing*, *7*(5), 01–10. https://doi.org/10.53555/hsn.v7i5.1591

[74] Arowolo M.O, Adebiyi M.O, Aremu C. and Adebiyi A.A, “A Survey of Dimension Reduction and Classification Methods for RNA-Seq Data on Malaria” *Vector J. Big Data*, Vol.8, pp. 50, 2022.

[75] Liu S, Xu C, Zhang Y, Liu J, Yu B, Liu X. and Dehmer M, “Feature Selection of Gene Expression Data for Cancer Classification Using Double RBF-Kernels”. *BMC Bioinform*, Vol.19, pp. 396, 2018.

[76] Garrido Castro A.C, Lin N.U. and Polyak K, “Insights into Molecular Classifications of Triple-Negative Breast Cancer: Improving Patient Selection for Treatment”. *Cancer Discov*., Vol.9, pp. 176–198, 2019.

[77] Chabon J.J, Hamilton E.G, Kurtz D.M, Esfahani M.S, Moding E.J, Stehr H, Schroers Martin J, Nabet B.Y, Chen B. and Chaudhuri A.A metal, "Integrating Genomic Features for Non-Invasive Early Lung Cancer Detection". *Nature*, Vol.580, pp. 245–251, 2020.

[78] Crosby D, Bhatia S, Brindle K.M, Coussens L.M, Dive C, Emberton M, Esener S, Fitzgerald R.C, Gambhir S.S. and Kuhn P metal, "Early Detection of Cancer". *Science*, Vol.375, eaay9040, 2022.

[79] Segal N.H, Pavlidis P, Noble W.S, Antonescu C.R, Viale A, Wesley U.V, Busam K, Gallardo H, DeSantis D. and Brennan M.F metal, "Classification of Clear-Cell Sarcoma as a Subtype of Melanoma by Genomic Profiling". *JCO*, Vol.21, pp. 1775–1781, 2003.

[80] Ram M, Najafi A. and Shakeri M.T, "Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest". *Iran J. Pathol*, Vol.12, pp. 339–347,2017.

[81] Hijazi H. and Chan C, "A Classification Framework Applied to Cancer Gene Expression Profiles". *J. Healthc. Eng*, Vol.4, pp. 255–284,2013.

[82] Yuan L, Sun Y. and Huang G, "Using Class-Specific Feature Selection for Cancer Detection with Gene Expression Profile Data of Platelets". *Sensors*, Vol.20, pp. 1528,2020.

[83] Abdulqader D.M, Abdulazeez A.M. and Zeebaree D.Q, "Machine Learning Supervised Algorithms of Gene Selection" *A Review. Mach. Learn*, Vol.62, pp. 233–244, 2020.

[84] Perdomo Ortiz A, Benedetti M, Realpe-Gómez J. and Biswas R, "Opportunities and Challenges for Quantum-Assisted Machine Learning in near-Term Quantum Computers". *Quantum Sci. Technol*, Vol.3, pp. 030502, 2018.

[85] Korbar B, Olofson A.M, Miraflor A.P, Nicka C.M, Suriawinata M.A, Torresani L, Suriawinata A.A. and Hassanpour S, "Deep Learning for Classification of Colorectal Polyps on Whole-Slide Images". *J. Pathol. Inform*, Vol.8, pp. 30, 2017.

[86] Lai Y.H, Chen W,N, Hsu T.C, Lin C, Tsao Y. and Wu S, "Overall Survival Prediction of Non-Small Cell Lung Cancer by Integrating Microarray and Clinical Data with Deep Learning". *Sci. Rep*, Vol.10, pp. 4679, 2020.

[87] Zhang D, Zou L, Zhou X. and He F, "Integrating Feature Selection and Feature Extraction Methods With Deep Learning to Predict Clinical Outcome of Breast Cancer*." IEEE Access* , Vol.6, pp. 28936–28944, 2018.

[88] Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, Vermeulen L. and Wang X, "DeepCC: A Novel Deep Learning-Based Framework for Cancer Molecular Subtype Classification". *Oncogenesis* , Vol.8, pp. 44, 2019.

[89] Chandrasekar V, Sureshkumar V, Kumar T.S. and Shanmugapriya S, "Disease Prediction Based on Micro Array Classification Using Deep Learning Techniques. Microprocess". *Microsyst*, Vol.77, pp. 103189,2020.

[90] Laplante J.F. and Akhloufi M.A, "Predicting Cancer Types from MiRNA Stem-Loops Using Deep Learning. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)", *Montreal, QC, Canad*a, Vol.20–24, July 2020; pp. 5312–5315.

[91] Sahin C.B. and Diri B, "Robust Feature Selection with LSTM Recurrent Neural Networks for Artificial Immune Recognition System". *IEEE Access*, Vol.7, pp. 24165–24178, 2019.

[92] Aher C.N. and Jena A.K, "Rider-Chicken Optimization Dependent Recurrent Neural Network for Cancer Detection and Classification Using Gene Expression Data". *Comput. Methods Biomech. Biomed. Eng. Imaging Vis*, Vol.9, pp. 174–191, 2021.

[93] Suresh A, Nair R.R, Neeba E.A. and Kumar S.A.P, "RETRACTED ARTICLE: Recurrent Neural Network for Genome Sequencing for Personalized Cancer Treatment in Precision Healthcare". *Neural Process Lett*. 2021.

[94] Batur¸ Sahi N.C. and DiRi B, "Sequential Feature Maps with LSTM Recurrent Neural Networks for Robust Tumor Classification". *Balk. J. Electr. Comput. Eng.*, Vol.9, pp. 23–32, 2020.

[95] Siddalingappa R. and Sekar K, "Bi-Directional Long Short-Term Memory Using Recurrent Neural Network for Biological Entity Recognition". *IJ-AI* , Vol.11, pp. 89,2022.

[96] Jiang L, Sun X, Mercaldo F. and Santone A, "DECAB-LSTM: Deep Contextualized Attentional Bidirectional LSTM for Cancer Hallmark Classification". *Knowl. -Based Syst*, Vol.210, pp. 106486,2020.

[97] Zhao Y, Joshi P. and Shin D.G, "Recurrent Neural Network for Gene Regulation Network Construction on Time Series Expression Data. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)", San Diego, CA, USA, 18–21 November 2019, pp. 610–615.

[98] Liu L. and Liu J , "Reconstructing Gene Regulatory Networks via Memetic Algorithm and LASSO Based on Recurrent Neural Networks". *Soft. Comput*, Vol. 24, pp. 4205–4221, 2020.

[99] Chowdhury S, Dong X. and Li X, "Recurrent Neural Network Based Feature Selection for High Dimensional and Low Sample Size Micro-Array Data. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)", Los Angeles, CA, USA, pp. 4823–4828, 9–12 December 2019.

[100] Altschul S, Gapped BLAST and PSI-BLAST, "A New Generation of Protein Database Search Programs", *Nucleic Acids Res*, Vol. 25, pp. 3389–3402,1997.

[101] Ghorbani M, Jonckheere E.A. and Bogdan P, "Gene Expression Is Not Random: Scaling, Long-Range Cross-Dependence, and Fractal Characteristics of Gene Regulatory Networks", *Front. Physiol.*, Vol. 9, pp. 1446, 2018.

[102] Nguyen P.T, Nguyen T.T, Nguyen N.C. and Le T.T, "Multiclass Breast Cancer Classification Using Convolutional Neural Network. In Proceedings of the 2019 International Symposium on Electrical and Electronics Engineering (ISEE)", Ho Chi Minh, Vietnam, pp. 130–134, 10–12 October 2019.

[103] Majji R, Nalinipriya G, Vidyadhari C. and Cristin R, "Jaya Ant Lion Optimization-Driven Deep Recurrent Neural Network for Cancer Classification Using Gene Expression Data", *Med. Biol. Eng. Comput.,* Vol. 59, pp. 1005–1021, 2021.

[104] Lyu B. and Haque A, "Deep Learning Based Tumor Type Classification Using Gene Expression Data. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics", Washington DC USA, pp. 89–96, 29 August–1 September 2018.

[105] Karimi Jafarbigloo S and Danyali H, "Nuclear Atypia Grading in Breast Cancer Histopathological Images Based on CNN Feature Extraction and LSTM Classification", CAAI Trans Intel Tech, Vol. 6, pp. 426–439, 2021.

[106] López-García G, Jerez J.M, Franco L. and Veredas F.J, "Transfer Learning with Convolutional Neural Networks for Cancer Survival Prediction Using Gene-Expression Data", *PLoS ONE* , Vol.15, e0230536, 2020.

[107] Wang S, Chen Y, Chen S, Zhong Q. and Zhang K, "Hierarchical Dynamic Convolutional Neural Network for Laryngeal Disease Classification", *Sci. Rep.* , Vol. 12, 13914, 2022.

[108] Wang Y, Wang Y.G, Hu C, Li M, Fan Y, Otter N, Sam I, Gou H, Hu Y, Kwok T. and et al. "Cell Graph Neural Networks Enable the Digital Staging of Tumor Microenvironment and Precise Prediction of Patient Survival in Gastric Cancer", *MedRxiv* 2021.

[109] Azadifar S, Rostami M, Berahmand K, Moradi P. and Oussalah M, "Graph-Based Relevancy-Redundancy Gene Selection Method for Cancer Diagnosis", *Comput. Biol. Me.,* Vol. 147, 105766, 2022.

[110] Pfeifer B, Saranti A. and Holzinger A, "GNN-SubNet: Disease Subnetwork Detection with Explainable Graph Neural Networks", *Bioinformatics,* Vol. 38, pp. ii120–ii126, 2022.

[111] Ramirez R, Chiu Y.-C, Zhang S, Ramirez J, Chen Y, Huang Y. and Jin Y.-F, "Prediction and Interpretation of Cancer Survival Using Graph Convolution Neural Networks.", *Methods,* Vol. 192, pp.120–130,2021.

[112] Zhou Y, Graham S, Alemi Koohbanani N, Shaban M, Heng P.-A. and Rajpoot N, "CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)", Seoul, South Korea, pp. 388–398, 27–28 October 2019.

[113] Rajasekharan H, Chivilkar S, Bramhankar N, Sharma T. and Daruwala R, "EEG-Based Mental Workload Assessment Using a Graph Attention Network. In Proceedings of the 2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)", Banff, AB, Canada, pp. 78–84, 29–31 October 2021.

[114] Xiang M, Hou J, Luo W, Tao W. and Wang D, "Impute Gene Expression Missing Values via Biological Networks: Optimal Fusion of Data and Knowledge. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)", Shenzhen, China, pp. 1–8, 18–22 July 2021.

[115] Wang J, Ma A, Ma Q, Xu D. and Joshi T, "Inductive Inference of Gene Regulatory Network Using Supervised and Semi-Supervised Graph Neural Networks", *Comput. Struct. Biotechnol. J.,* Vol. 18, pp. 3335–3343, 2020.

[116] Zhang X.-M, Liang L, Liu L. and Tang M.-J, "Graph Neural Networks and Their Current Applications in Bioinformatics", *Front. Genet,* Vol. 12, 690049,2021.

[117] Xu Q, Zhu L, Dai T. and Yan C, "Aspect-Based Sentiment Classification with Multi-Attention Network", *Neurocomputing,* Vol. 388, pp. 135–143, 2020.

[118] Lv Z, Lin Y, Yan R, Yang Z, Wang Y. and Zhang F, "PG-TFNet: Transformer-Based Fusion Network Integrating Pathological Images and Genomic Data for Cancer Survival Analysis. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)", Houston, TX, USA, pp. 491–496, 9–12 December 2021.

[119] Wensel J, Ullah H. and Munir A, "ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos", ArXiv , arXiv:2208.07929, 2022.

[120] Dirgová Luptáková I, Kubovˇcík M. and Pospíchal J, "Wearable Sensor-Based Human Activity Recognition with Transformer Model", *Sensors,* Vol. 22, pp. 1911, 2022.

[121] Kakati T, Bhattacharyya D.K, Kalita J.K. and Norden-Krichmar, "T.M. DEGnext: Classification of Differentially Expressed Genes from RNA-Seq Data Using a Convolutional Neural Network with Transfer Learning", *BMC Bioinform,* Vol. 23,pp. 17, 2022.

[122] Zhang Y, Chen J.-H, Lin Y, Chan S, Zhou J, Chow D, Chang P, Kwong T, Yeh D.-C. and Wang, X. et al. "Prediction of Breast Cancer Molecular Subtypes on DCE-MRI Using Convolutional Neural Network with Transfer Learning between Two Centers." *Eur Radiol* , Vol. 31,pp 2559–2567,2021.

[123] Maudsley S, Chadwick W, Wang L, Zhou Y, Martin B. and Park S.-S, "Bioinformatic Approaches to Metabolic Pathways Analysis. In Signal Transduction Protocols; Luttrell, L.M., Ferguson, S.S.G., Eds.; Methods in Molecular Biology", *Humana Press: Totowa,* NJ, USA, Vol. 756, pp. 99–130, ISBN 978-1-61779-159-8, 2011.

[124] Dalamaga M, "Obesity, Insulin Resistance, Adipocytokines and Breast Cancer: New Biomarkers and Attractive Therapeutic Targets." WJEM 2013, 3, 34.

[125] Ho C.-H, Huang Y.-J, Lai Y.-J, Mukherjee R. and Hsiao C.K, "The Misuse of Distributional Assumptions in Functional Class Scoring Gene-Set and Pathway Analysis." *G3 Genes Genomes Genet,* Vol. 12, jkab365, 2022.

[126] Joshi P, Basso B, Wang H, Hong S.-H, Giardina C. and Shin D.-G, "RPAC: Route Based Pathway Analysis for Cohorts of Gene Expression Data Sets.", *Methods,* Vol. 198, pp. 76–87,2022.

[127] Ma J, Shojaie A. and Michailidis G, "A Comparative Study of Topology-Based Pathway Enrichment Analysis Methods." *BMC Bioinform,* Vol. 20, pp. 546,2019.

[128] Bauer-Mehren A, Furlong L.I. and Sanz F, "Pathway Databases and Tools for Their Exploitation: Benefits, Current Limitations and Challenges." *Mol. Syst. Biol*., Vol. 5, pp. 290,2009.

[129] Joshi-Tope G, "Reactome: A Knowledgebase of Biological Pathways.", *Nucleic Acids Res*, Vol. 33,pp. D428–D432,2004.

[130] Zhou H, Jin J, Zhang H, Yi B, Wozniak M. and Wong L, "IntPath–an Integrated Pathway Gene Relationship Database for Model Organisms and Important Pathogens." *BMC Syst. Biol*., Vol. 6, S2,2012.