Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Classification of Cancer Microarray Data Based on Deep Learning: A Review

Jawaher Abdulwahab Fadhil [1], Adnan Mohsin Abdulazeez [2]
[1]jawaher.fadhil@auas.edu.krd, [2]adnan.mohsin@dpu.edu.krd
[1] Department of Information Technology, Technical College of Informatics- Akre, Akre University for Applied Science, Duhok, Kurdistan Region, Iraq.
[2] Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

Article Information

Submitted : 22 Jan 2024
Reviewed: 3 Feb 2024
Accepted : 10 Feb 2024

Keywords

Cancer Classification, Gene Expression, Deep Learning, Microarray

Abstract

This review article delves into applying deep learning methodologies in conjunction with microarray data for cancer classification. The study provides a comprehensive overview of recent advancements in utilizing deep learning techniques to accurately categorize cancer types based on intricate patterns discerned from microarray datasets. Various aspects are covered, including integrating deep learning algorithms, exploring diverse cancer types, and analyzing microarray data to enhance classification accuracy. The review synthesizes findings from recent research, highlighting the efficacy of deep learning in uncovering subtle and complex relationships within microarray data that contribute to improved classification outcomes. Key insights into the strengths and limitations of employing deep learning in this context are discussed, offering a critical appraisal of the field's current state. This review aims to provide a valuable resource for researchers, clinicians, and practitioners interested in cutting-edge developments in cancer classification methodologies by exploring the intersection of deep learning and microarray technology. The synthesis of knowledge presented herein contributes to a deeper understanding of the potential and challenges associated with harnessing deep learning for enhanced classification accuracy in the realm of cancer research.

Introduction

Cancer, a group of diseases characterized by the uncontrolled growth of malignant cells resulting from genetic alterations, poses a significant threat to human health. These aberrant cells proliferate unchecked, infiltrating organs and, in many cases, leading to fatal outcomes. Globally, cancer stands as the second most prevalent cause of mortality, surpassed only by cardiovascular diseases[1] In recent times, gene expression analysis has emerged as a pivotal tool in tackling the intricate challenges associated with cancer diagnosis and drug discovery[2],[3]´ . This analytical approach not only sheds light on the intricate molecular landscape of cancer but also unravels the roles of various genes in its initiation and progression. Consequently, alterations in gene expression patterns serve as valuable indicators for the early detection of cancer and identification of potential targets for drug development. This transformative use of gene expression analysis opens avenues for healthcare that is not only more personalized but also proactive and predictive[4]. By leveraging the insights derived from gene expression, we can envision a future where healthcare strategies are tailored to individual profiles, emphasizing prevention and early intervention for enhanced patient outcomes.

1. Gene Expression

Gene expression analysis constitutes identifying transcripts within specific cells or tissues, aiming to estimate the levels of expressed genes. The scientific field dedicated to quantitatively examining the transcriptome is known as transcriptomics[5] . In the initial stages of computational transcriptomics, Sanger sequencing was the prevailing method for analyzing expressed sequence tag (EST) libraries. These libraries consist of concise mRNA fragments derived from a single sequencing procedure applied to randomly chosen clones originating from cDNA libraries. Essentially, a cDNA library is a compilation of DNA sequences that have been cloned, serving as complements to mRNA extracted from an organism or tissue. A substantial milestone in this field has been the production of over 45 million EST libraries, encompassing a diverse array of approximately 1400 distinct cellular species to date.

While EST (expressed sequence tag) libraries offer a foundational resolution profile of expressed gene sequences, it's important to note their limitation in not containing full-length gene sequences. Consequently, technologies relying on EST libraries were surpassed by chemical tag-based techniques, with Serial Analysis of Gene Expression (SAGE) emerging as a prominent method. SAGE enables quantitative and simultaneous analysis of numerous transcripts within a specific cell system, requiring no prior knowledge of the genes involved. This method relies on a theoretical calculation assuming a random nucleotide distribution across the genome. The evolution from Sanger sequencing of EST libraries and SAGE led to the adoption of more advanced technologies, including DNA (Deoxyribonucleic Acid) microarrays and Ribonucleic acid (RNA-Seq.) within Next-Generation Sequencing (NGS) methods, for a more comprehensive and precise estimation of gene expression levels.

2.1 Microarray

Microarray data is derived from a laboratory technique where a DNA sequence is embedded in a two-dimensional array, often referred to as chips or slides, comprising thousands of microscopic spots. Each spot is designated for a single DNA sequence or gene. The hybridization process facilitates binding DNA samples to the microarray slide, followed by color scanning of the areas to measure gene expression[5] . In microarray data, rows signify gene expression levels, while columns represent individual samples. Microarrays serve multiple purposes, capable of identifying DNA (as in comparative genomic hybridization) or RNA, often in the form of cDNA following reverse transcription. These data contribute to a comprehensive understanding of cellular processes, offering genome-wide expression profiles linked to specific conditions or diseases, such as cancer. Beyond diagnostics, microarray data plays a crucial role in pharmaceutical research, pharmacogenomics, and the development of effective therapeutic medications.[6].

A notable advantage of DNA microarrays lies in their capacity to measure the expression level of thousands of genes simultaneously. However, it is essential to acknowledge their limitations, including accuracy, precision, and specificity challenges. The experimental setup's high sensitivity to variations in hybridization temperature, genetic material purity, degradation rate, and amplification potentially influences the accurate quantification of gene expression.[7].

2.2 RNA-Seq.

RNA-Sequencing (RNA-Seq.), a part of Next-Generation Sequencing (NGS) (Hu, Chitnis et al. 2021)methods, is distinguished by its rapid profiling capabilities, enabling researchers to explore the transcriptome of any species to determine the presence and quantity of RNA at specific times [8]. This method generates millions of sequences from intricate RNA samples, serving various purposes such as measuring gene expression, investigating variations in gene expression over time or in response to therapies, annotating complete transcripts, exploring post-transcriptional modifications, and characterizing alternative splicing and polyadenylation.

The versatility of RNA-Seq. lies in its capacity to analyze all RNA molecules within a cell or tissue, encompassing protein-coding RNA (mRNA), non-coding regulatory RNA (miRNA, siRNA), or functional RNA (tRNA, rRNA), and concurrently measure their abundances. Noteworthy qualities include high resolution and a broad dynamic range, contributing to substantial data acquisition and significant progress in transcriptomics research. Given these advantages, RNA-Seq has progressively supplanted microarrays in gene expression analysis, as highlighted in the comparison presented in Table 1, which assesses factors like discovered gene range, different isoforms, resolution, background noise, cost, rare/new transcripts, and non-coding RNA [9].In conclusion, RNA-Seq. stands out for its numerous advantages over microarray data.

Table 1. Distinguishing microarray from RNA-Seq. data.

Distinctive Attributes	***Microarray Datasets***	***RNA-Seq. Datasets***
Gene Discovery	No	Yes
Different Isoform	No	Yes
High Resolution	No	Yes
Background Noise	Yes	No
High Cost	Yes	No
Rare/New Transcript	No	Yes
Noncoding RNA	No	Yes

2. Public Datasests

This section describes the commonly available Microarray datasets. Different repositories provide Microarray datasets; this review focused on two widely used by researchers to evaluate their proposed models. These resources are explained as follows:

2.1 Gene Expression Omnibus (GEO)

[10]GEO is a comprehensive global data repository for functional genomics, facilitating MIAME-compliant data submissions [10], It accommodates diverse datasets, including RNA-seq and Microarray data, setting it apart from platforms like GEO, which predominantly focuses on Microarray data. Noteworthy is the expansive collection of 3635328 disease-specific samples accessible through GEO, offering a valuable resource for researchers. The repository is freely accessible for experimental purposes, providing meticulously curated gene expression profiles to support scientific investigations.

2.2 The Cancer Genome Atlas (TCGA)

TCGA is a pioneering initiative in cancer genomics, offering an extensive collection of 84,031 samples spanning 33 different types of cancer. [11] Notably, TCGA presents datasets that encompass measurements from microarray and RNA-seq instruments. It is crucial to acknowledge that the predominant focus of these datasets lies in assessing gene expression levels across normal and cancerous tissues, predominantly utilizing RNA-seq technologies. This dual approach enhances the comprehensiveness of the available data, providing researchers with a nuanced understanding of the genomic landscape in diverse cancer types.

3. Deep Learning Approaches

Deep learning techniques utilize artificial neural networks (ANNs) featuring multiple strata of processing units to acquire insights into data patterns. These approaches excel at assimilating intricate representations within expansive datasets, conferring a distinctive edge over traditional machine learning (ML) methodologies[12]. As a result, contemporary cutting-edge approaches to gene expression analysis capitalize on the distinctive competencies offered by these techniques [13] Prevalent neural network architectures encompass fully

connected networks (multi-layer perceptron NN), convolutional networks (CNN), recurrent networks (RNN), graph networks (GNN), and transformer networks (TNN) [14].

4.1 Multi-layer perceptron (MLP) 1

MLP is a prominent type of feedforward neural network within pattern recognition, classification challenges, and prediction, primarily applied to solve supervised learning problems [15] Operating through the mapping of input to output in a unidirectional flow of data and calculations, MLP typically comprises three layers: an input layer, an output layer, and at least one intervening layer known as a hidden layer [16] These layers are fully connected, with the input layer

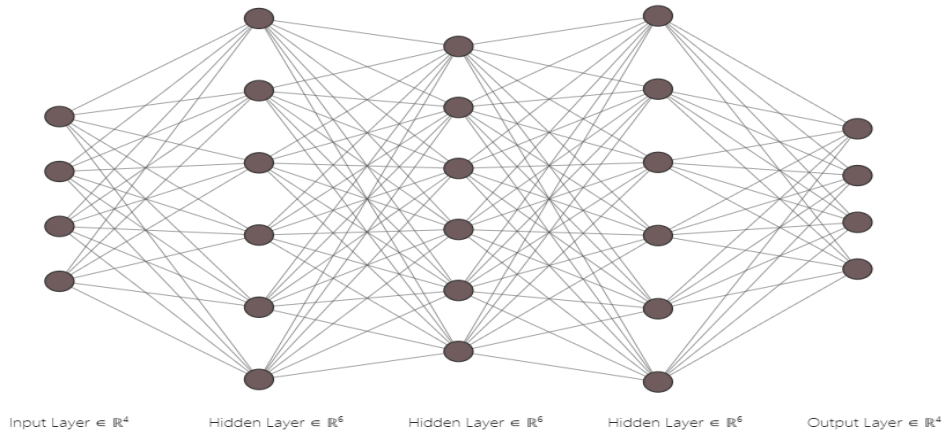

Figure 1.multi-layer perceptron structure

receiving signals from the external environment, hidden layers executing arithmetic operations from input to output, and the output layer making decisive predictions. Each layer features nodes or neurons, and the MLP workflow involves four key steps. First, the input data is propagated from the input layer to the output layer. Second, MLP learns through weight updates between neurons, employing a backpropagation algorithm after processing the input data for each node [14]. Third, errors are calculated by assessing the disparity between predicted and known classes, employing supervised learning to minimize these errors. Lastly, these steps iterate over multiple cycles to refine and perfect the weights in the learning process. MLP structure described in Figure 1.

4.2 Recurrent Neural Networks (RNN)

RNN, encompassing Feedforward Neural Networks, can transmit data across various time steps, as illustrated in Figure 2.. Unlike feedforward propagation, which allows information to flow in a singular direction, RNN employs recursion, creating a loop of information as depicted in Figure 2. This recursive approach involves scanning the entire data from left to right, with shared parameters for each time step [17]. Despite its merits, RNN has a limitation – it relies solely on information preceding a point in a sequence for predictions, neglecting any information occurring later in the sequence.

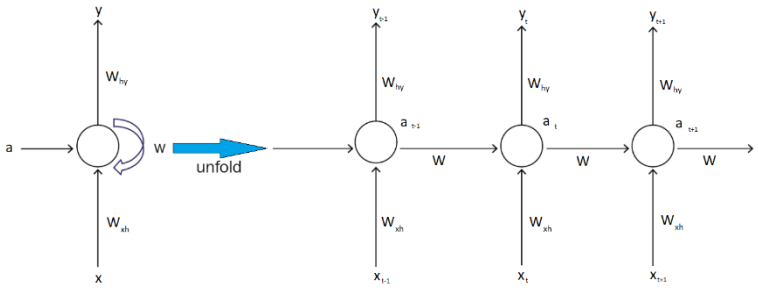

Figure 2. Recurrent Neural Network structure.

4.3 Convolutional Neural Networks (CNN)

Inspired by the visual processing in animals' brains, CNN is a sophisticated multi-layer neural network pioneered by LeCun et al. Its primary application domains encompass image processing and character recognition, as noted by [18] ,[19]. The architectural framework involves the initial layer discerning features, followed by intermediate layers that amalgamate these features to generate high-level input characteristics, culminating in a classification process. The accumulated characteristics undergo pooling to reduce dimensionality, and subsequent steps involve convolution and pooling, ultimately feeding into a fully connected multi-layer perceptron [20].

The final layer, the output layer, employs back-propagation techniques to recognize the image's distinctive features, as elucidated by[21]. CNN stands out due to its distinctive attributes, such as local connection and shared weights, contributing to heightened system accuracy and performance. It surpasses other

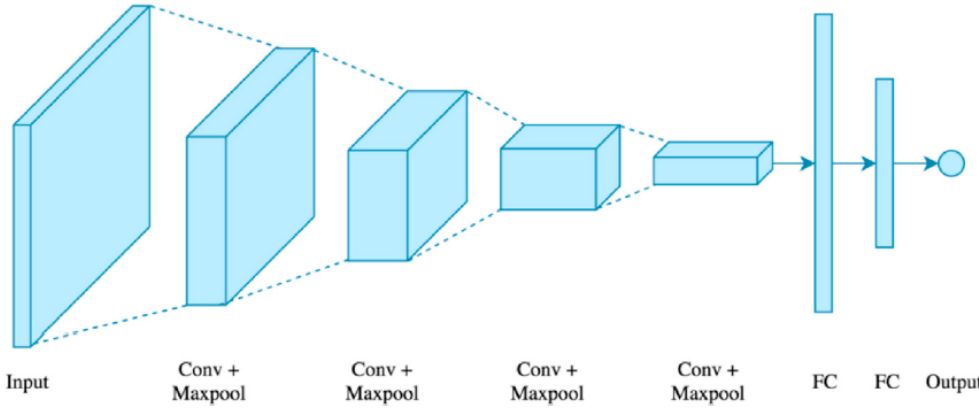

Figure 3. Convalution Neural Network structure.

deep learning techniques and stands as the most employed architecture. For a visual representation, refer to Figure 3.illustrating the structure of a convolutional neural network.

4.4 Long Short-Term Memory (LSTM)

LSTM networks, falling under the umbrella of recurrent neural networks (RNNs), exhibit a noteworthy proficiency in grasping long-term dependencies, as

exemplified in [22]. The architecture of an LSTM involves the intricate construction of a memory cell utilizing logistic and linear units with multiplicative interactions. This design facilitates a dynamic flow of information within the cell: information is admitted through the input gate, expelled when the forget gate is inactive, and accessed for reading by activating the output gate. Such nuanced operations empower LSTMs to effectively capture and retain information over extended sequences, showcasing their prowess in addressing scenarios with prolonged dependencies.

Utilizing Deep LSTM Recurrent Neural Networks presents a notable enhancement in speech recognition accuracy. This architecture is crafted by assembling a stack of LSTM layers, yet it can also be structured without stacking, resembling a Feedback Neural Network unrolled when each layer shares identical model parameters. Like the structure of Deep Neural Networks (DNNs), inputs may traverse one or multiple non-linear layers; however, the distinctive feature is that the information from a specific time instant undergoes processing by a singular non-linear layer before producing the result for that moment. As highlighted in[22], the depth in deep LSTMs holds a particular significance. The data traverses a sequence of LSTM layers within a specific time frame. The incorporation of deep layers in LSTM RNNs contributes to the network's ability to learn across various time scales, showcasing the effectiveness of this approach in capturing intricate temporal dependencies.

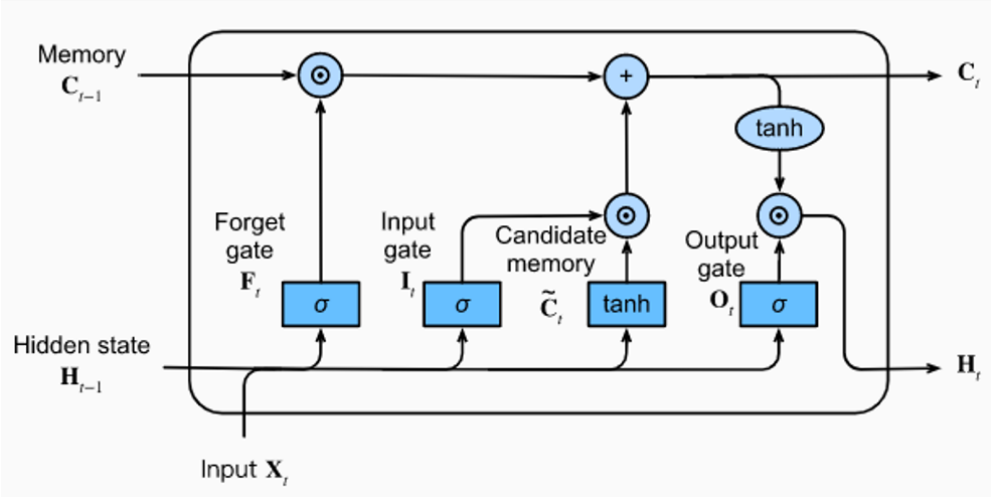

Figure 4. Long short-term memory (LSTM) structure.

4.5 Graph Neural Networks (GNN)

GNNs belong to the realm of deep learning algorithms tailored for the examination and interpretation of structured data encapsulated within graphs. Graphs, comprising interconnected nodes and edges, serve as versatile models to depict relationships and dynamics across diverse domains like social networks, biological systems, citation networks, and recommendation frameworks [23]. The primary objective of GNNs is to acquire nuanced representations of individual nodes within a graph[24]). This entails capturing not only the characteristics of a node's immediate surroundings but also discerning patterns in the broader structural context of the entire graph.

In the realm of GNNs, the process begins with the representation of a graph where nodes symbolize entities, and edges signify relationships between these entities. Each node is endowed with features, offering insights into the corresponding entity.

The journey continues with node embeddings, where initial embeddings are assigned to nodes based on their features. GNNs engage in iterative message-passing steps, allowing nodes to gather information from their neighbors and update their embeddings accordingly; this involves the exchange of neighborhood information and using learnable aggregation functions such as mean, sum, or attention mechanisms [23]. Stacked aggregation layers further refine node embeddings by assimilating information from increasingly expansive neighborhoods to encapsulate local and global graph structures; some employ graph pooling layers, contributing to hierarchical representation[24]. Ultimately, the process culminates in the output layer, where the final node embeddings derived from these intricate steps can be applied to diverse tasks such as node classification, link prediction, or graph classification.

4. Evaluation Performance

5.1 Accuracy

Accuracy (AC): AC is a metric used to evaluate the performance of a classification model. In machine learning, particularly in classification tasks, accuracy measures how well a model correctly predicts the labels of the instances in the dataset. It is described as in (1).

$$Accuracy \ \frac{Correct\ predictions}{All\ predictions} \qquad (1)$$

The number of correct predictions represents the count of instances for which the model's prediction matches the actual labels. The total number of Predictions refers to the sum of correct and incorrect predictions, representing the total number of instances in the dataset. Accuracy is usually expressed as a percentage, ranging from 0% to 100%. A higher accuracy indicates better performance, with 100% accuracy, meaning that the model made correct predictions for all instances[25].

5.2 Precision (Pre)

Pre is a metric in the realm of classification that quantifies the accuracy of positive predictions by assessing the ratio of true positives to the sum of true positives and false positives. Mathematically, precision is computed using the formula [26]:

$$Pre \ \frac{TP}{TP+FP} \qquad (2)$$

5.3 Recall (Rec)

Recall, also referred to as sensitivity, is a pivotal metric in classification evaluation, measuring the average probability of achieving comprehensive retrieval. This

metric gauges the model's ability to identify and capture all relevant instances within the positive class. The formula is below [26].

$$Rec = \frac{TP}{TP+FN} \quad (3)$$

5.4 F-1score (F1)

F1 serves as a harmonized measure, presenting a weighted average of both precision and recall. An ideal F1 score attains a value of 1, indicating perfect precision and recall synchronization, while the lowest achievable score is 0 [17].

$$F1 - score = 2 + \frac{Pre*Rec}{Pre+Rec} \quad (4)$$

In the context of classification metrics, the components of the precision formula are delineated as follows: True Positive (TP) signifies instances accurately predicted as belonging to the positive class by the model. Conversely, True Negative (TN) represents instances where the model precisely predicts cases as part of the negative class—a scenario, for instance, when non-cancerous cases are correctly identified as such. False Positive (FP) denotes instances erroneously predicted as part of the positive class, like when a patient is inaccurately identified as having cancer when they do not. Finally, False Negative (FN) characterizes instances where the model incorrectly predicts cases as part of the negative class, such as when a patient with cancer is not identified as such by the model.

5. Literature Review

Numerous recent studies have been made using deep learning approaches to classify gene expression data; this section presents a comprehensive review.

Researchers in [27] Present a novel approach called D-SVM (Deep support vector machine) integrating deep learning with the traditional SVM to predict breast cancer; they reported that their approaches outperformed the traditional classification, especially on small-sized datasets such as the breast cancer dataset, as their proposed approach reaches an accuracy of 69.8% while it reaches 69.6% and 59.4% with DNN and SVM, respectively on the same dataset.

In this work [28], Stacked denoising Autoencoder (SDAE) was employed to extract functional features from intricate high-dimensional gene expression profiles; the main goal of the SDAE model is to extract a mapping that possibly decodes the initial data as specifically as can be done without having an important loss of gene patterns. Subsequently, the efficacy of the extracted representation was scrutinized using supervised classification models, affirming the utility of the newly derived features in the context of cancer detection; the SDAE features were employed on three classification learning models ANN, SVM, SVM-RBF, yields an accuracy of 96.95%, 98.04%, 98.26%, respectively.

The analysis of tumor microarray data demands ample training models to construct a classifier with improved accuracy.as in [29] the scarcity of data Addressed by integrating diverse datasets encompassing multiple types of cancer. They employed the Multi-Task Deep Learning (MTDL) algorithm for microarray data analysis, effectively mitigating data scarcity issues. The application of MTDL

substantially boosted the accuracy of the classifier, demonstrating its effectiveness in accurately (98.5% Overall accuracy for 12 cancer datasets) identifying the type of cancer when tested across various cancer datasets.

T. Ahn *et al* Trained deep neural network (DNN) to differentiate between cancer and normal samples and then employ diverse gene selection strategies on TCGA and GEO datasets to yield an accuracy of 99.7%. The selection encompassed therapeutic target genes from commercial cancer panels and genes within NCI-curated cancer pathways. A systematic analysis method was proposed for interpreting the trained deep neural network. This approach was subsequently applied to identify the genes that predominantly contribute to classifying cancer in individual samples[30].

unsupervised feature learning framework that combines principal component analysis and an auto-encoder neural network to extract unique characteristics from gene expression profiles was introduced in [31]. These features are then used to construct an ensemble classifier (PCA-AE-Ada) based on the AdaBoost algorithm for predicting clinical outcomes in breast cancer. The proposed method is compared to a baseline classifier (PCA-Ada) using the same learning strategy but with different training inputs. Evaluation of multiple breast cancer datasets demonstrates that the deep learning approach outperforms other gene signature-based algorithms in predicting clinical outcomes. Their proposed classifier's best performance was on the GSE11121 dataset with an accuracy of 85%, and the lowest accuracy was on GSE2034 with 75% compared to the baseline classifier with an accuracy of 68% and 65%, respectively, on the same datasets.

a novel approach was proposed by [32] utilizing Convolutional Neural Networks (CNN) coupled with spectral clustering information processing for the classification of lung cancer. The method integrates protein interaction network data and gene expression profiles, demonstrating the effectiveness of this spectral-convolutional neural network.

the efficacy of a convolutional neural network (CNN) based deep learning algorithm for classifying microarray data Explored by researcher in [33]. This investigation included a comparative analysis with other established techniques, namely Vector Machine Recursive Feature Elimination and an enhanced Random Forest approach (mSVM-RFE-iRF and varSeIRF). The findings revealed that the performance of the CNN varied across different datasets, demonstrating that it does not universally outperform all other methods. Despite this variability, the experimental results on cancer datasets consistently highlighted the CNN's superiority in terms of accuracy (81.53% overall accuracy for the ten cancer datasets) and its ability to minimize gene-related features when classifying cancer, as compared to the hybrid mSVM-RFE-iRF approach.

Researchers in [34] Outline an innovative framework for supervised cancer classification termed Deep Cancer Subtype Classification (DeepCC). This approach leverages deep learning to analyze functional spectra, quantifying the activities of biological pathways for precise and effective cancer subtype classification.

D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez [35] Proposed a new approach that revolves around modeling enduring unit cells through Long Short-Term Memory (LSTM). LSTM, a subtype of Recurrent Neural Network (RNN) within the realm of Artificial Neural Networks (ANN), is a suitable framework. Employing

LSTM proves to be a pragmatic strategy for forecasting process durations, especially in scenarios where the classifier's learning from experience is uncertain and extended intervals between significant events are unpredictable; the model in this work was tested on Colon, lung, SRBCT, Lymphoma, Leukemia and prostate datasets from gene expression profile (GEP) datasets with accuracy of 89.6%, 88.3%, 85.3%, 84.7%, 77.6%, and 75.7%.

Investigation of classification methods has evaluated the accuracy of several potent deep learning algorithms, including the Deep Neural Network (DNN), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and an enhanced Deep Neural Network incorporating preprocessing techniques. Has been made by O. Ahmed and A. Brifcani, [36], to address model over-fitting, they apply a Dropout augmentation mechanism with DLBCL, Prostate, and Leukemia datasets on DNN, which effectively overcame the associated challenges, as the accuracy results came as follows: DLBCL =98.4%, Prostate =93.2%, Leukemia = 99%, and Colon = 91.4%.

Deep feedforward method to effectively classify microarray cancer data into distinct classes for subsequent diagnostic purposes employing a 7-layer deep neural network architecture with varied parameters tailored to microarray cancer datasets. Was developed by [37],A widely recognized dimensionality reduction technique, namely principal component analysis, was implemented to mitigate challenges related to small sample size and dimensionality. Feature values were standardized using the Min–Max approach, and the proposed methodology underwent validation on eight standard microarray cancer datasets. Binary cross-entropy was utilized for loss measurement, and optimization employed adaptive moment estimation.

a stochastic gradient descent-based (SGD-based) deep neural network was employed by researcher in [38] on the ten most common UCI (University of California Irvine) Cancer dataset yields an accuracy of 92%; this study utilizes deep learning techniques with a Softmax activation function. This approach is applied to the condensed features, specifically genes, to enhance the classification of diverse samples based on their gene expression levels.

Researcher in [39] , [40] used deep transfer learning with convolutional neural network on Lung Cancer, 11 Datasets Cancer types reaching an accuracy of 73.26% and 98%, while in [41]. 1D CNN model and a 2D CNN model was proposed; the two-dimensional CNN performed the best with an accuracy of 98.86%, while the one-dimensional CNN is only 80.36%.

the SVM-mRMRe model, was Introduced in [42] which is a novel approach to gene selection in cancer research from high-dimensional microarray data. SVM for feature ranking and mRMRe for improved selection are combined in this hybrid method. After being tested on eight different cancer datasets, SVM-mRMRe showed increased relevance and accuracy, consistent with known biological knowledge. Understanding cancer pathways may be improved by utilizing the model's emphasis on biologically relevant gene selection.

The PCC-DTCV model, a hybrid machine-learning technique designed to use complex microarray gene expression data for cancer classification. Was presented by researchers in [43] its main advantages was the Awareness of the difficulties of handling high-dimensional data, the model uses Pearson's correlation to facilitate

effective gene selection and optimizes a Decision Tree classifier using Grid Search CV. The PCC-DTCV model was tested on seven cancer datasets and showed impressive specificity and accuracy. It also successfully reduced the dimensionality of the data while identifying critical genes for cancer classification.

Researchers in [44] proposed a novel AIFSDL-PCD methodology for the identification and categorization of PCa. This innovative technique integrates distinct phases, including preprocessing, feature selection utilizing CIWO, classification through DNN, and hyperparameter optimization using RMSprop. The utilization of CIWO for feature selection contributes to a reduction in computational complexity and enhancement of classification accuracy, tested on the prostate dataset with an accuracy of 96%.

A novel feature selection approach, termed Intersection-Based Three Feature Selection Methods (ITFS), has been developed by [26] to strategically identify optimal features (genes) for classification while concurrently reducing the dimensionality of gene expression data. ITFS incorporates three distinct feature selection techniques, namely Mutual Information (MI), F-ClassIf, and Minimum Redundancy Maximum Relevance (mRMR). By employing the intersection concept, ITFS selectively identifies genes that are concurrently chosen by all three feature selection methods. These selected genes then serve as identifiers for training the classifier model.

Perceptron (MLP) compared to the standalone use of MLP. This underscores the enhanced performance achieved through the synergy of ITFS and MLP in the classification process. In the same year, another research published by [45] reached an accuracy of 98% using Fuzzy Gene Selection (FGS). This novel approach integrates Mutual Information, F-ClassIf, and Chi-squared feature selection techniques to rank genes based on their importance in cancer classification. Fuzzification and Defuzzification techniques are then applied to consolidate these rankings into a single optimal score for each gene. FGS is particularly effective in multi-class scenarios. A unique Fuzzy classifier is developed to address convergent decisions in classifiers, leveraging contributions from traditional deep classifiers at individual nodes. This combined approach enhances the robustness and accuracy of predictions in cancer classification.

pioneering optimization strategy, PSCS, in conjunction with deep learning for classifying brain tumors presented by A. A. Joshi and R. M. Aziz[46]. The PSCS enhances the classification process by refining Particle Swarm Optimization (PSO) by integrating the Cuckoo Search (CS) algorithm. Subsequently, deep learning is used to classify gene expression data associated with brain tumors, employing the PSCS optimization technique to identify distinct groups or classes relevant to specific tumor types. When combined with deep learning, the proposed optimization approach attains significantly enhanced classification accuracy (98.7%)compared to existing deep learning and machine learning models, as assessed through various evaluation metrics such as Recall, Precision, F1-Score, and the confusion matrix.

the development of a novel Multidimensional Fuzzy Deep Learning (MFDL) approach was introduced in[47], to meticulously identify a subset of crucial genes. This involved integrating fuzzy concepts seamlessly into filter and wrapper methods, enabling the selection of significant genes. Subsequently, these chosen

genes were employed to train the model, enhancing overall accuracy. The MFDL methodology further extended its impact by incorporating a fuzzy classifier, thereby refining cancer classification accuracy; extensive experimentation and validation were conducted on six distinct gene expression datasets, and the outcomes affirm the effectiveness of this methodology across diverse cancer datasets as it yields 98% accuracy.

This research [48] introduces an innovative hybrid methodology for gene selection in cancer classification, termed CSSMO (Cuckoo Search Spider Monkey Optimization). The fitness of the Spider Monkey Optimization (SMO) algorithm is tailored through integration with the Cuckoo Search Algorithm (CSA), resulting in CSSMO. This approach leverages the strengths of both metaheuristic algorithms to efficiently identify a subset of genes crucial for early-stage cancer prediction. To refine the CSSMO algorithm's accuracy, a cleaning process is implemented using the Minimum Redundancy Maximum Relevance (mRMR) technique. This process aims to reduce gene expression noise in cancer datasets, enhancing the robustness of the selected gene subset. Subsequently, deep learning (DL) is employed to classify these gene subsets. Eight microarray gene expression datasets were used.

Summary of the studies using deep learning approaches were applied to gene expression data are illustrated in Table2.

Table 2. Summary of studies used deep learning approaches to classify gene expression data.

Ref	***year***	***Dataset***	***model***	***Acc.%***	***Pros***	***Cons***
[27]	2017	Breast cancer	Deep-SVM	69.8	Direct clinical significance, utilization of large datasets (TCGA) Cancer Genome Atlas , and an complete evaluation through several performance indicators.	Potential complexity is brought about by combining SVM with deep learning, the lack of comparison analysis with alternative models, the possibility of overfitting when using deep learning on high-dimensional data, and issues with clinical interpretability.
[28]	2017	Breast cancer	Stacked Denoising Autoencoder (SDAE)	96.95 98.0498. 26	The promise for better cancer diagnosis through applying Stacked Denoising Autoencoders (SDAE) to extract significant features from high-dimensional gene expression profiles. Additionally, the technique identified a group of highly interacting genes as possible cancer biomarkers.	Obstacles include the necessity for huge datasets and additional confirmation of the discovered biomarkers. Furthermore, even though deep learning models may be scaled, they need a lot of processing power.
[29]	2017	12 Cancer datasets	multi-task deep learning (MTDL)	overall accuracy of 98.5.	In the face of sparse gene expression data, the paper introduces a novel Multi-task Deep Learning (MTDL) algorithm specifically for tumor classification. By effectively integrating data from many kinds of cancer, this method tackles data shortages head-on and produces more resilient models. Compared to standard approaches, MTDL greatly improves diagnostic accuracy	The hybrid model has drawbacks, including higher processing requirements and specialized knowledge requirements. To ensure it is resilient, more thorough testing using a variety of datasets and techniques would be beneficial. Furthermore, due to the inherent complexity of the model, there is a possibility of overfitting, which calls for cautious regularization. Additionally, there is still concern about the model's interpretability because it may need

					across various cancer types by collecting shared and unique traits among various cancer types.	help understanding or explaining the judgments and learned representations.
[30]	2018	TCGA, GEO	DNN	99.7%	The paper offers a novel deep-learning method for identifying cancer using extensive gene expression data. The (DNN) showed remarkable accuracy in cancer identification using data from reliable databases. The goal of the project is to create a universally applicable cancer classifier. Furthermore, utilizing large datasets, a proposed interpretation method sheds light on the functions performed by genes in cancer, which may lead to a more profound understanding and better treatments.	Even though DNNs are extremely powerful, they can be difficult to interpret and comprehend some genes. Disparities between data sources, like RNA-Seq. and microarray, might need fixing with consistency. The DNN's capacity to generalize new data raises some questions. Furthermore, even while the model is good at classifying data, it could not always provide a deep molecular understanding of cancer.
[31]	2018	Breast cancer	DNN	85, 75	Deep learning algorithms are integrated into the new method, which shows improved performance in biomedical applications. The PCA-AE-Ada technique outperforms other algorithms in terms of accuracy and other evaluation metrics across several datasets related to breast cancer.	The deep learning model's intricacy could make it more difficult to interpret. With more diverse datasets, the generalization of the method could be improved. Potential overfitting in deep learning models is a concern, particularly when data is lacking.
[32]	2019	Lung cancer	CNN	83.15.	The study surpasses conventional methods with an accuracy rate of 83.15% and blends spectral clustering with convolutional neural networks for enhanced biological data analysis. It properly maintains protein network topological links and tackles issues unique to processing omics data. The method is flexible enough to deal with different biological networks and can potentially integrate with multi-omics data.	Although the emphasis is on lung cancer particularly, adjustments may be needed to make it applicable to other cancers. Furthermore, deep learning is resource-intensive, putting computational demands on models that may need to be more transparent and easily interpreted. Furthermore, because of its intrinsic complexity, the model may be less able to generalize to new data due to overfitting.
[33]	2019	10 cancer datasets	CNN	81.53	With CNNs, the work leverages a state-of-the-art method designed for the sophisticated interpretation of intricate microarray data, demonstrating impressive precision, especially in certain cancer datasets. CNNs have the innate ability to handle large and complex datasets necessary for efficient microarray analysis. Furthermore, these networks automatically identify	CNNs are powerful, but they have drawbacks as well. For example, they need a lot of processing power, which makes large datasets difficult to use. It can be intimidating to deploy them, particularly for those not experienced with deep learning. Furthermore, the complex architecture of CNNs increases the likelihood of overfitting, which may jeopardize the model's generalizability. Moreover, their intrinsic "black box" nature may mask the biological

					and utilize pertinent features, reducing human participation requirements. Their proven adaptability to different cancer datasets highlights their potential for further field applications.	discoveries beneath. Interestingly, CNN performance varies with the cancer dataset, indicating differences in performance due to dataset complexity.
[34]	2019	Cancer subtypes	(DeepCC)	90	By utilizing the power of advanced deep learning, DeepCC improves cancer subtype classifications to unprecedented levels. Its architecture is notable for its ability to withstand common problems like platform variances, batch errors, and data gaps. Molecular subtyping accuracy increases the possibility of tailored cancer therapies. DeepCC performs remarkably better in tests than traditional techniques, demonstrating its applicability in real-world scenarios. Moreover, its quick analysis of single samples makes it an effective tool for expediting clinical choices.	Although DeepCC has sophisticated cancer subtype classification skills, its technological complexity and interpretability pose obstacles. Concerns regarding biases are raised by the model's reliance on certain gene expression datasets, and rigorous validation is necessary due to its propensity for overfitting. More widespread validation across a range of patient populations and cancer types is necessary for its clinical significance to be fully understood.
[35]	2019	six datasets from (GEP) datasets	RNN [LSTM-AIRS]	89.6, 88.3, 85.3, 84.7, 77.6, 75.7 99.3	The Artificial Immune Recognition System (AIRS) and Long Short-Term Memory (LSTM) networks are combined in this paper to present a unique bioinformatics approach for finding tumor-related genes. Promisingly, the suggested PAIRS2 algorithm outperformed other techniques on the Lymphoma dataset with a high accuracy of 99.3%.	Concerns have been raised regarding possible overfitting in the study because of its heavy focus on obtaining high accuracy. In addition, the evaluation's breadth and depth may be limited by employing linear classifiers. Moreover, the significance of the results might be limited to microarray datasets, which could restrict their applicability to other kinds of experimental data.
[36]	2019	DLBCL, Prostate, Leukemia and Colon	improved-DNN	98.4, 93.2, 99, 91.4	The research explores the use of advanced deep learning algorithms for bioinformatics classification, highlighting the high accuracy of an improved Deep Neural Network on various datasets bolstered by thorough methods and preprocessing strategies.	Although the work provides insights into bioinformatics, it emphasizes technical rather than biological elements, is unclear on dataset generalizability, and might use more visual aids for better comprehension.
[37]	2020	8 microarray cancer datasets	7-layer deep neural network architecture	90	The work presents a novel method based on multivariate beta mixtures that minimize the requirement for human data labeling due to its parameter-less design, hence providing efficiency. Empirical validation based on real-world data from widely-used community Q&A sites highlights the method's potential application on various online platforms.	Although it shows promise, a more thorough assessment against current techniques is required, particularly regarding its scalability and performance on larger or more diverse datasets.

[38]	2020	10 most common UCI Cancer Datasets	Elephant search optimization based deep learning approach	92	The paper offers a unique approach to gene expression selection that combines deep learning and optimization based on Elephant search. It performs a comprehensive analysis on multiple cancer microarray datasets and compares favorably with conventional techniques to demonstrate its efficacy. Additionally, the study uses stringent statistical testing, indicating its applicability for more general uses in bioinformatics and clinical cancer diagnosis.	Its use seems to be mainly restricted to the cancer datasets under study, which begs the question of its wider generalizability. Ten-fold cross-validation is used, but worries regarding overfitting and the model's capacity to adjust to fresh data still exist. More validation is necessary, particularly on bigger datasets or in clinical contexts.
[39]	2020	Lung Cancer	Deep Transfer Learning +CNN	73.26	The study uses CNNs to analyze gene-expression data, pre-training on a large Pan-Cancer dataset, and then fine-tuning for individual cancer types using a transfer learning approach. It better captures characteristics by converting RNA-Seq. samples into gene-expression pictures. In order to advise prospective customized treatments for lung cancer, The technique aimed to predict PFI for lung cancer and performed better than other machine learning methods with regards to AUC values.	Implementing and understanding the method can be difficult for individuals unfamiliar with the field. Its heavy reliance on large amounts of data for sufficient model training raises questions regarding its generalizability across various cancer types and therapeutic settings. Furthermore, the approach is computationally intensive and is dangerous to overfit, which could result in less-than-ideal performance on fresh data.
[40]	2020	11 Cancer Datasets	Deep Transfer Learning +CNN	98.9	GeneXNet exhibits a remarkable 98.9% accuracy over 33 cancer types from 26 organ sites. Most notably, it does away with the lengthy and conventionally required gene feature selection process. Moreover, a transfer learning feature makes it more flexible, enabling it to adapt to tumors with less information.	GeneXNet presents challenges with data constraints, potentially limiting its broad applicability. Concerns arise regarding overfitting and the model's ability to generalize without further validation. Moreover, its computational demands are significant, and the inherent complexity of the model makes its decisions less interpretable due to its "black box" nature.
[42]	2021	8 microarray datasets	SVM-mRMRe model	Average 99	A thorough approach to microarray data analysis is provided by SVM-mRMRe, which exhibits improved classification accuracy for cancer tissue in various datasets. Its clinical value is reinforced by the gene selections that align with existing biomedical knowledge. Notably, the model consistently performs well across various datasets and adeptly negotiates the challenges of high dimensionality and constrained sample sizes in microarray data.	Due to its dual-stage and ensemble design, SVM-mRMRe may require careful parameter tweaking and provide computational hurdles. Its complexity may make it difficult to comprehend the underlying biological processes. Additionally, even if SVM-mRMRe performs admirably, comparative assessments point to situations where alternative methods might perform better. Furthermore, the method's efficacy may differ based on the dataset, which could restrict its broad use.

[43]	2021	7 cancer datasets	PCC-DTCV model	96	For efficient gene selection, the model uses Pearson's correlation coefficient (PCC), which lowers the complexity of the data. It demonstrates adaptability and relevance when tested on seven different microarray cancer datasets, and metrics like specificity and accuracy bolster its effectiveness. Furthermore, interpretability is improved via decision trees (DT), and their efficiency is optimized using grid search and other optimization approaches.	Grid Search and multi-dataset processing make the model's implementation computationally demanding. Although accuracy and AUC are prioritized, false positive and false negative rates—crucial for medical applications—are noticeably neglected. Moreover, the multi-phase procedure, which involves preprocessing and feature selection utilizing Pearson's correlation coefficient (PCC), introduces intricacy and may exclude certain pertinent genes in particular situations.
[41]	2022	Liver cancer	Deep Transfer Learning +CNN	98.86 80.36	H By employing an optimized VGG16 model, the technique demonstrated remarkable 100% accuracy in distinguishing liver cancer from normal sequences, highlighting its effectiveness in DNA sequence analysis. The study provides a diverse and potentially comprehensive representation of genomic data by utilizing three distinct numerical mapping techniques and combining CNN models such as VGG16. Moreover, the research gains credibility and eases further confirmation through the reliable NCBI database.	here are doubts regarding the model's generalizability to other applications due to the study's reliance on a tiny dataset that only includes four genes—four cancerous and four healthy. The complex fusion of several mapping techniques and deep learning methodologies may present difficulties for those unfamiliar with the field. Furthermore, there isn't a thorough analysis of the research's shortcomings and possible dangers. Furthermore, achieving a 100% accuracy score raises concerns about how well-suited and consistent the model is for use with bigger or more diverse datasets.
[44]	2022	Prostate Cancer	AIFSDL-PCD	96.44	The efficacy of the AIFSDL-PCD approach was highlighted by its impressive 96.44% accuracy in detecting prostate cancer. A state-of-the-art method incorporates modern approaches like deep neural networks (DNN) and chaotic invasive weed optimization (CIWO). Furthermore, the CIWO technique suggests possible computing advantages, particularly optimizing feature selection. The model's validation on a dataset with 102 tissue examples further confirms its dependability.	Even if the model comes from a reliable source, its dependence on a tiny dataset can limit its broader applicability. Deep learning's intrinsic computational needs might make it difficult to implement. There are still unanswered questions regarding the model's possible overfitting to fresh data, and its deep learning components might make it harder to see how the machine makes decisions transparently, making it harder to understand how it works.
[45]	2023	14 cancer datasets	Fuzzy deep leaning	92.8% to 100%,	The FGS-FC model presents a novel strategy by combining a fuzzy classifier with fuzzy gene selection, and it achieves good accuracy in cancer classification. It efficiently manages complex gene expression data, reducing overfitting and exhibiting adaptability to various cancer	The study offers the FGS-FC cancer classification model, which has limitations but shows promise. It restricts larger applicability by targeting certain types of cancer. Despite being novel, its ambiguity and complexity may make it difficult to use and understand. Furthermore, the model's efficacy depends on the

					types and datasets. The paper also identifies avenues for future research to improve cancer classification methods, including combining multi-omics data and utilizing deep learning.	availability of high-quality gene expression data, which presents difficulties for regular application.
[26]	2023	6 cancer datasets	MLP	average 96	The research presented a method that creatively combined three feature selection strategies to obtain an excellent classification accuracy of approximately 96%. This method successfully addressed the issues caused by the high dimensionality of gene expression data and was proven in several datasets. Furthermore, the model showed improved performance when combined with Multilayer Perceptron.	The model's combination of several methodologies creates difficulties for execution and interpretation. Its effectiveness depends on how good and complete the dataset is, but a more thorough comparison with other approaches now in use is needed. Furthermore, it could be difficult to understand the model's predictions and biological significance if the results are difficult to comprehend.
[46]	2023	Brain tumor	CNN	98.7	An innovative approach to AI-driven diagnostics, the revolutionary integration of PSCS with deep learning has greatly improved the classification accuracy of brain tumors. The approach has been thoroughly evaluated using a variety of measures, highlighting its potential to improve patient outcomes by facilitating more accurate tumor classifications.	Analysis increases with the combination of several approaches. Furthermore, the method's effectiveness is inextricably linked to the caliber and volume of gene expression data that are readily available. Caution is also necessary because of ethical concerns about decision openness, accuracy, and data privacy. Detailed clinical validation is required for the strategy to demonstrate its practical value.
[47]	2023	Six cancer Datasets	multidimensional fuzzy deep learning (MFDL)	98	AI-driven diagnostics have taken a new turn with the combination of PSCS and deep learning, improving the accuracy of brain tumor categorization. A thorough assessment of the method's success has been made possible by applying various indicators. Ultimately, by permitting more accurate tumor classification, this novel technique holds the potential for bettering patient outcomes.	Implementation and interpretation issues arise when numerous methodologies are integrated. Furthermore, the number and quality of the gene expression data that is now accessible are integrally linked to the approach's success. Accuracy, decision openness, and data privacy are other ethical issues brought to light. As such, additional clinical validation is necessary to thoroughly validate the procedure and establish its efficacy in the real world.
[48]	2023	8 cancer datasets	CSSMO+CNN	99	The technique improves the accuracy of cancer categorization by using effective gene selection and a synergistic combination of the Cuckoo Search and Spider Monkey Optimization algorithms. This careful gene selection reduces the likelihood of overfitting. Furthermore, its adaptability points to uses other than cancer categorization.	The technique highlights potential difficulties and the significance of accurate parameter setups when working with high-dimensional datasets. Transparency is hampered by its difficulty interpreting nature, and its effectiveness varies with the kinds of biological data. Consequently, more optimization is required to improve its suitability for wider genomics applications.

6. Challenges

Volume versus Cohort Size: Many gene expression databases exist. Nevertheless, the small cohort sizes and the wide range of factors, such as gene expression levels, present a major obstacle. Both classical machine learning (ML) algorithms and deep learning (DL) algorithms face difficulties as a result of this complexity [49] Although curated public datasets are made available by platforms such as TCGA and GEO, combining these datasets necessitates thorough pre-processing and harmonization to guarantee data consistency.

Data Availability: Primary databases such as TCGA and GEO provide the majority of publicly available resources for cancer gene expression. Known for their voracious appetite for large amounts of data, deep learning models encounter difficulties creating correct models for newly available cancer datasets. Researchers have looked into dropout strategies, data augmentation, regularization techniques (such as ridge and lasso), and streamlining neural network topologies as potential answers. Still, a conclusive answer to this enduring problem is elusive.

The "curse of dimensionality" presents a substantial obstacle to the use of artificial intelligence in gene expression analysis [34]this phrase captures the challenges that come with working with high-dimensional data. Random effects that are not reliably reproducible across similar patient groups may result from the sheer number of dimensions[50].

7. Conclusion

This paper carefully examines the various ways that Deep Learning (DL) approaches are applied in the complex and varied field of cancer research on a wide range of cancer types that cover an extensive range of human anatomy and physiology; this in-depth examination explores the uses and effects of DL in many types of cancers such as lung, breast, kidney, liver, prostate, gallbladder, and central nervous system (CNS). It offers an integrative perspective by illuminating the revolutionary potential of deep learning for improving our understanding and diagnosis of these complicated cancers. The scope of these Deep Learning studies encompasses diverse objectives such as cancer identification, subtype classification, and gene biomarker identification. A comprehensive analysis identifies the prevailing tools utilized for gauging gene expression disparities between benign and malignant tissues. Noteworthy datasets commonly employed in evaluating Deep Learning (DL) models using gene expression data are spotlighted, shedding light on the standard practices in this domain. It's critical to acknowledge that, despite significant recent developments, the analysis of gene expression data in cancer research continues to be a challenging and dynamic field. There are still several obstacles to overcome, creating additional research and creativity opportunities.

Every obstacle offers a chance to learn more, improve methods, and uncover new perspectives that can fundamentally alter our molecular understanding of cancer.

Over the past few years, there has been a noticeable achievement in methodology; the most recent developments have been demonstrated to be notably more accurate and effective than the previous years, This impressive advancement may

be credited to the careful design and extensive testing of these new techniques on a wide range of datasets, improving their performance and showing their adaptability in handling complex issues.. In summary, the paper explores deep learning approaches by examining their revolutionary potential and demonstrating their uniqueness in identifying all of the constraints associated with traditional machine learning approaches. In cancer research, these sophisticated algorithms are employed to deduce insights from the challenging task of analyzing gene expression data. As a result, they have proven to be exceptionally effective and proficient in refining, improving, and elevating the analysis process—an almost perfect capability. And this has the alluring potential to not only address present-day problems but also to overcome them, resulting in a period defined by increased precision, deeper insights that are subtle throughout, and groundbreaking discoveries.

8. References

[1] K. D. Miller *et al.*, "Cancer statistics for the US Hispanic/Latino population, 2021," *CA. Cancer J. Clin.*, vol. 71, no. 6, pp. 466–487, 2021.

[2] G. Munkácsy, L. Santarpia, and B. Győrffy, "Gene Expression Profiling in Early Breast Cancer—Patient Stratification Based on Molecular and Tumor Microenvironment Features," *Biomedicines*, vol. 10, no. 2, p. 248, 2022.

[3] A. Brewczyński *et al.*, "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, no. 13, p. 3256, 2021.

[4] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, p. baaa010, 2020.

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

[6] N. N. Mohammed and A. M. Abdulazeez, "Evaluation of partitioning around medoids algorithm with various distances on microarray data," in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, IEEE, 2017, pp. 1011–1016.

[7] D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and D. A. Zebari, "Machine learning and region growing for breast cancer segmentation," in *2019 International Conference on Advanced Science and Engineering (ICOASE)*, IEEE, 2019, pp. 88–93.

[8] S. Jungjit, M. Michaelis, A. A. Freitas, and J. Cinatl, "Extending multi-label feature selection with KEGG pathway information for microarray data analysis," in *2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, IEEE, 2014, pp. 1–8.

[9] Y. Wang *et al.*, "Changing technologies of RNA sequencing and their applications in clinical oncology," *Front. Oncol.*, vol. 10, p. 447, 2020.

[10] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, 2002.

[11] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project,"

Nat. Genet., vol. 45, no. 10, pp. 1113–1120, 2013.
[12] A. Perdomo-Ortiz, M. Benedetti, J. Realpe-Gómez, and R. Biswas, "Opportunities and challenges for quantum-assisted machine learning in near-term quantum computers," *Quantum Sci. Technol.*, vol. 3, no. 3, p. 030502, 2018.
[13] B. Korbar *et al.*, "Deep learning for classification of colorectal polyps on whole-slide images," *J. Pathol. Inform.*, vol. 8, no. 1, p. 30, 2017.
[14] W. Zhu, L. Xie, J. Han, and X. Guo, "The application of deep learning in cancer prognosis prediction," *Cancers*, vol. 12, no. 3, p. 603, 2020.
[15] A. Maharjan, "Machine Learning Approach for Predicting Cancer Using Gene Expression," University of Nevada, Las Vegas, 2020.
[16] R. Xie, J. Wen, A. Quitadamo, J. Cheng, and X. Shi, "A deep auto-encoder model for gene expression prediction," *BMC Genomics*, vol. 18, pp. 39–49, 2017.
[17] S. Babichev, I. Liakh, and I. Kalinina, "Applying a Recurrent Neural Network-Based Deep Learning Model for Gene Expression Data Classification," *Appl. Sci.*, vol. 13, no. 21, p. 11823, 2023.
[18] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep learning for brain MRI segmentation: state of the art and future directions," *J. Digit. Imaging*, vol. 30, pp. 449–459, 2017.
[19] D. Zahras and Z. Rustam, "Cervical cancer risk classification based on deep convolutional neural network," in *2018 International Conference on Applied Information Technology and Innovation (ICAITI)*, IEEE, 2018, pp. 149–153.
[20] S. M. S. Abdullah and A. M. Abdulazeez, "Facial expression recognition based on deep learning convolution neural network: A review," *J. Soft Comput. Data Min.*, vol. 2, no. 1, pp. 53–65, 2021.
[21] S. Gupta and M. K. Gupta, "A comprehensive data-level investigation of cancer diagnosis on imbalanced data," *Comput. Intell.*, vol. 38, no. 1, pp. 156–186, 2022.
[22] M. Kaur and A. Mohta, "A review of deep learning with recurrent neural network," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, 2019, pp. 460–465.
[23] X. Jing, Y. Zhou, and M. Shi, "Dynamic Graph Neural Network Learning for Temporal Omics Data Prediction," *IEEE Access*, vol. 10, pp. 116241–116252, 2022.
[24] X.-M. Zhang, L. Liang, L. Liu, and M.-J. Tang, "Graph neural networks and their current applications in bioinformatics," *Front. Genet.*, vol. 12, p. 690049, 2021.
[25] D. Al-obidi and S. Kacmaz, "Facial Features Recognition Based on Their Shape and Color Using YOLOv8," in *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, IEEE, 2023, pp. 1–6.
[26] M. Khalsan, M. Mu, E. S. Al-Shamery, L. Machado, M. O. Agyeman, and S. Ajit, "Intersection Three Feature Selection and Machine Learning Approaches for Cancer Classification," in *2023 International Conference on System Science and Engineering (ICSSE)*, IEEE, 2023, pp. 427–433.
[27] D. Sun, M. Wang, H. Feng, and A. Li, "Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine: supervised feature extraction and classification for breast cancer prognosis prediction," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, IEEE, 2017, pp. 1–5.
[28] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for

cancer detection and relevant gene identification," in *Pacific symposium on biocomputing 2017*, World Scientific, 2017, pp. 219–229.
[29] Q. Liao, L. Jiang, X. Wang, C. Zhang, and Y. Ding, "Cancer classification with multi-task deep learning," in *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, IEEE, 2017, pp. 76–81.
[30] T. Ahn *et al.*, "Deep learning-based identification of cancer or normal tissue using gene expression data," in *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, IEEE, 2018, pp. 1748–1752.
[31] D. Zhang, L. Zou, X. Zhou, and F. He, "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.
[32] T. Matsubara, T. Ochiai, M. Hayashida, T. Akutsu, and J. C. Nacher, "Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles," *J. Bioinform. Comput. Biol.*, vol. 17, no. 03, p. 1940007, 2019.
[33] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, "Gene selection and classification of microarray data using convolutional neural network," in *2018 International Conference on Advanced Science and Engineering (ICOASE)*, IEEE, 2018, pp. 145–150.
[34] F. Gao *et al.*, "DeepCC: a novel deep learning-based framework for cancer molecular subtype classification," *Oncogenesis*, vol. 8, no. 9, p. 44, 2019.
[35] C. B. Şahín and B. Dírí, "Robust feature selection with LSTM recurrent neural networks for artificial immune recognition system," *IEEE Access*, vol. 7, pp. 24165–24178, 2019.
[36] O. Ahmed and A. Brifcani, "Gene expression classification based on deep learning," in *2019 4th Scientific International Conference Najaf (SICN)*, IEEE, 2019, pp. 145–149.
[37] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," *CAAI Trans. Intell. Technol.*, vol. 5, no. 1, pp. 22–33, 2020.
[38] M. Panda, "Elephant search optimization combined with deep neural network for microarray data analysis," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 8, pp. 940–948, 2020.
[39] G. Lopez-Garcia, J. M. Jerez, L. Franco, and F. J. Veredas, "Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data," *PloS One*, vol. 15, no. 3, p. e0230536, 2020.
[40] T. Khorshed, M. N. Moustafa, and A. Rafea, "Deep learning for multi-tissue cancer classification of gene expressions (GeneXNet)," *IEEE Access*, vol. 8, pp. 90615–90629, 2020.
[41] B. Das and S. Toraman, "Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences," *Biomed. Signal Process. Control*, vol. 72, p. 103317, 2022.
[42] P. El Kafrawy, H. Fathi, M. Qaraad, A. K. Kelany, and X. Chen, "An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data," *IEEE Access*, vol. 9, pp. 155353–155369, 2021.
[43] H. Fathi, H. AlSalman, A. Gumaei, I. I. Manhrawy, A. G. Hussien, and P. El-Kafrawy, "An efficient cancer classification model using microarray and high-dimensional data," *Comput. Intell. Neurosci.*, vol. 2021, 2021.

[44] A. M. Alshareef *et al.*, "Optimal deep learning enabled prostate cancer detection using microarray gene expression," *J. Healthc. Eng.*, vol. 2022, 2022.
[45] M. Khalsan, M. Mu, E. S. Al-Shamery, L. Machado, S. Ajit, and M. O. Agyeman, "Fuzzy Gene Selection and Cancer Classification Based on Deep Learning Model." arXiv, May 04, 2023. doi: 10.48550/arXiv.2305.04883.
[46] A. A. Joshi and R. M. Aziz, "Deep learning approach for brain tumor classification using metaheuristic optimization with gene expression data," *Int. J. Imaging Syst. Technol.*, p. e23007, 2023.
[47] M. Khalsan, M. Mu, E. S. Al-Shamery, S. Ajit, L. Machado, and M. O. Agyeman, "A Novel Fuzzy Classifier Model for Cancer Classification Using Gene Expression Data," *IEEE Access*, 2023.
[48] R. Mahto *et al.*, "A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection," *BMC Bioinformatics*, vol. 24, no. 1, p. 479, 2023.
[49] D. L. Barbour, "Precision medicine and the cursed dimensions," *NPJ Digit. Med.*, vol. 2, no. 1, p. 4, 2019.
[50] D. M. Abdulqader, A. M. Abdulazeez, and D. Q. Zeebaree, "Machine learning supervised algorithms of gene selection: A review," *Mach. Learn.*, vol. 62, no. 03, pp. 233–244, 2020.