IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 21, 2021, accepted January 5, 2022, date of publication January 25, 2022, date of current version March 16, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3146312

A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction

MAHMOOD KHALSAN[1,2], LEE R. MACHADO[3], EMAN SALIH AL-SHAMERY[2], SURAJ AJIT[1], KAREN ANTHONY[3], MU MU[1], (Member, IEEE), AND MICHAEL OPOKU AGYEMAN[1], (Senior Member, IEEE)

[1]Centre for Advanced and Smart Technologies, Faculty of Arts, Science and Technology, University of Northampton, Northampton NN1 5PH, U.K.
[2]Computer Science Department, College of Information Technology, University of Babylon, Babil 51002, Iraq
[3]Centre for Physical Activity and Life Science, Faculty of Arts, Science and Technology, University of Northampton, Northampton NN1 5PH, U.K.

Corresponding author: Mahmood Khalsan (mahmood.khalsan@northampton.ac.uk)

ABSTRACT Machine learning approaches are powerful techniques commonly employed for developing cancer prediction models using associated gene expression and mutation data. This manuscript provides a comprehensive review of recent cancer studies that have employed gene expression data from several cancer types (breast, lung, kidney, ovarian, liver, central nervous system and gallbladder) for survival prediction, tumor identification and stratification. We also provide an overview of biomarker studies that are associated with these cancer types. The survey captures multiple aspects of machine learning associated cancer studies, including cancer classification, cancer prediction, identification of biomarker genes, microarray, and RNA-Seq data. We discuss the technical issues with current cancer prediction models and the corresponding measurement tools for determining the activity levels of gene expression between cancerous tissues and noncancerous tissues. Additionally, we investigate how identifying putative biomarker gene expression patterns can aid in predicting future risk of cancer and inform the provision of personalized treatment.

INDEX TERMS Biomarker, cancer prediction, deep learning, feature selection, machine learning, microarray, RNA-Seq.

I. INTRODUCTION

CANCER is a global health concern that causes death worldwide, according to the World Health Organization. Cancer has been defined as a group of cells that arise from specific areas of the human body, that often readily spreads to distant metastatic sites [1]. Abnormal growth of cells occurs because of the complex interaction between genes (deregulated due to mutation and epigenetic modifications) and the environment (i.e. carcinogens) [2]. Consequently, whole-genome expression analysis has become an important tool to identify relevant genes pathways that are deregulated and drive abnormal cellular proliferation and metastatic spread. Whole-genome expression (transcriptomic) analysis has the potential to provide early cancer prediction, for diagnosis, determining clinical outcomes, and the potential for disseminated disease. In addition, molecular cancer classification can enhance the success of personalized treatments including immune-checkpoint inhibitors including anti-PD1 and anti-CTLA-4 [2]. Measuring gene expression differences of thousands of genes between healthy and unhealthy tissue using transcriptomic approaches such as microarrays and more recently RNA-Seq has required investigators to develop bioinformatic pipelines that include mathematical and statistical methods to analyze these large novel datasets. Biomarker genes typically represent the identification of a subset of genes that are associated with a specific disease or subset of diseases. This paper explores the biomarker genes and gene signatures that have been implicated in cancer studies. Identification of new gene sigatures will help in early cancer prediction and potentially identify patient subgroups that may be responsible or non-responsive for specific therapies. In this sense, biomarkers have the potential to play a key role in determining treatment strategies. The recent public availability of these datasets now allows researchers to apply deep learning methods that may speed up data analysis and greatly

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

improve the accuracy of cancer diagnosis, prognosis, and likely response to therapy. The rest of the paper is organized as follows: Section II explores recent studies analyzing gene expression data either by using classical ML and DL for seven types of cancer (including their subtypes). Section II is focused on existing studies to identify putative biomarker genes that associated with cancer and the methods that have been developed for this purpose. Section III discussed the challenges in analyzing gene expression data using novel ML and DL approaches. Summary and conclusion are provided at the end of the survey.

II. BACKGROUND & LITERATURE REVIEW

This section presents a comprehensive review of the challenges encountered in studies that have analyzed gene expression data in cancer. This includes using different mathematical and statistical methods to identify genes or gene pathways that are associated with cancer pathogenesis and using these gene signatures to improve cancer prediction.

A. MICROARRAY & RNA-SEQ TECHNOLOGY

DNA microarrays are a technology that allow biologists to monitor the level of gene activity in an organism [13]. This is achieved by measuring the expression levels of each gene between healthy tissues and abnormal tissues. Microarray associated tools have opened the door for researchers to use mathematical and statistical methods to calculate differentially expressed levels for each gene between cancerous and non-cancerous samples. This allows identification of the top differentially expressed genes that might be associated with a particular disease. Perhaps more importantly pathway and gene ontology enrichment analysis can be used to identify putative biological pathways involved in the disease based on prior experimental work that has already been done with these genes.

Technically, Microarrays measure the intensity of fluorescence (fluorescently labelled cDNA molecules), where the intensity of fluorescence reflects the corresponding gene expression levels. The ability of microarrays to measure the expression of thousands of genes simultaneously depends on slides (known as DNA Microarrays) pre-spotted with thousands of probes complementary in sequence to the fluorescently labelled cDNA molecules that are added to the array (usually referred to as DNA/Gene chips). The known position of the probes on the chip allows assignation of specific gene expression patterns to individual genes [14].

To implement a microarray experiment, both a reference sample (e.g. from normal tissue) and an experimental sample (i.e. cancer tissue) are collected, the mRNA is extracted and converted to fluorescently labelled cDNA typically with one sample labelled with a green fluorescent marker and the other a red fluorescent marker. Then the two samples are combined and hybridized to the microarray slide. The slide is then washed to remove non-specific cDNA molecules and scanned to measure the gene expression of every gene sequence hybridized to the slide. A specific spot on the microarray will appear as red if the expression of a certain gene in the experimental sample is higher than in the reference sample. Conversely, the spot appears green if the expression of a certain gene in the experimental sample is lower than in the reference sample. The spot appears yellow if the expression of a certain gene is equal in both samples. No color means the gene is not expressed in either sample. The information collected through Microarrays can be used to generate gene expression profiles which display concurrent changes in the expression of multiple genes compared with a reference sample. This may represent a specific treatment or condition. RNA-seq is a measurement tool employing next generation sequencing technology (e.g. Illumina HiSeq) which can be used to determine gene expression and sequence differences between different types of biological samples and has largely superseded microarray technology. Microarray methods only profile predefined mRNA transcripts by hybridizing complementary DNA (cDNA) to pre-spotted oligonucleotide probes on an array. In contrast, RNA-Seq measures read counts as a proxy for relative abundance measures of gene expression levels [15]. RNA-Seq also provides single base pair resolution, distinguishing allelic expression of genes, identification of novel genes and altered splice forms and has a larger dynamic range, a good signal to noise ratio and is more accurate in measuring gene expression levels [16], [18]. RNA-Seq requires mapping of processed sequence reads to a reference genome (or transcriptome) and therefore is dependent on the accuracy of these references assemblies rather than being limited by a predetermined choice of probes (on an array). This allows an increased ability to identify new gene disease associations that may have been missed with microarray approaches [16]. RNA-seq able to detect expression at the gene, exon, transcript, and coding DNA sequence (CDS) levels, whereas Microarray can detect expression in a gene, exon level only [17].

B. PROBLEM STATEMENT

Researchers have used different mathematical and statistical methods to analyze gene expression data for several purposes, including the identification of informative gene associated pathways, improved disease classification, disease prediction, drug discovery, and personalized therapy. Various methods have been developed to address these goals. The challenge in all approaches concerns the complexity and high dimensionality of gene expression data. In addition, the number of individual cancers is vast including at least >100 molecularly distinct types of cancer. Moreover, tools employed for calculating gene expression genome-wide continue to evolve, which leads to improved accuracy in gene expression values. Novel gene expression patterns (novel transcripts) can be readily identified using RNA-Seq instead of using the DNA microarrays. As a result, technology development requires new mathematical and statistical methods to analyze such heterogeneous data. Another challenge is additional interacting factors (e.g. environmental factors) such as smoking, asbestos and dietary factors that are likely

to interact with and influence genes associated with specific cancers.

C. THE STUDIES IN THIS FIELD

Recent studies employing machine and deep learning approaches for cancer prediction and biomarker gene identification will be evaluated. Useful resources for downloading cancer gene expression data will be highlighted which contributes to researchers being able to test and evaluate their proposed analytical techniques.

1) CLASSICAL MACHINE LEARNING & FEATURE SELECTION METHODS

Machine learning is a sub-field of artificial intelligence that allows computers to learn without being explicitly programmed [3]. There are many different techniques of classical ML that have been developed, including, K-Nearest-Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Bayes, etc. ML is distributed across three major classes which are supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is applied after the data used is labelled because it is mapping between input and output data. In supervised learning the ML task is to learn a function that maps an input to an output using example input-output pairs. It deduces a function from labeled training data consisting of a set of training examples. In unsupervised Learning, the model works independently to discover patterns and previously undetected information. Therefore, unsupervised Learning does not involve data labelling. Reinforcement learning (RL) can be applied for both types of data (labelled and unlabeled). RL makes the decision sequentially, e.g., the output depends on the current input, and the next input relies on the output of the previous input. In contrast, for supervised learning the decision will be made based on the initial input. RL models aim to maximize the notion of cumulative reward (Figure 1).

Feature Selection(FS) is defined as a statistical method that is used for reducing the dimensionality of the data by selecting the informative features and ignoring uninformative ones in the dataset [9]. FS techniques have valuable benefits when used as they reduce the time training of a model, are less complicated and are easy to interpret [10]. More importantly, when the FS method employs optimal features, it will help enhance the accuracy of a model and reduce possible overfitting [11]. In general, FS approaches are divided into three fundamental types: filter methods, wrapper methods, and embedded methods [12].

To discuss and analyze a study that uses machine learning (traditional ML and DL approaches) it is beneficial to understand the evaluation parameters that were employed to demonstrate that this proposed study outperformed the previous studies. The evaluation parameters are as a follows: Accuracy (AC) is a metric that is used for evaluating classification models in AI. AC is calculated by dividing the number of correctly classified instances by the total number of instances in the dataset. Mathematically, it is calculated as follows [8]:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

where TP is True Positive, TN is True Negative, FP is False0 Positive and FN is False Negative. A TP is the result that is accurately predicted by the model as a positive class. TN is a result where there is accurate prediction by the model of a case as a negative class. For example, non-cancerous cases are classified and correctly called as noncancerous by the model. FP is the result that is inaccurately predicted a positive class (e.g. a patient has cancer when they do not). FN is where the model inaccurately predicted by the model as negative class (e.g. a patient with cancer is not called as such).

Precision is the average probability of relevant retrieval as illustrated in [8]

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is defined as the average probability of complete retrieval. It is also known as sensitivity. The recall formula is described as [8]:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1 score is a weighted average of the precision and recall, where a perfect F1 score has a value of 1 and worst score at 0 [8]:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

Extreme Learning Machines (ELM) have been developed as classifier techniques with Correlation Coefficient as feature selection method to reduce the number of features (genes) and increase the performance of this kind of sophisticated model. In a particular study, accuracy was accomplished with 79% [52]. The study used 60 central nervous system tumors (in addition to other tumors). A caveat to this study is that it utilized a small dataset. In addition, it also used only one gene expression dataset for evaluating the performance of the proposed model, and that may only show good results for this specific dataset. It may not be broadly applicable.

Suleyman *et al.* [21] used five machine learning techniques, namely random forest, support vector machine, Naïve Bayes, C4.5, and K-nearest Neighbor. In this study, somatic mutation data, from 358 patients, was obtained from TCGA and was used to predict breast cancer. The highest accuracy accomplished with this study was 0.70 when using Random forest, while the other machine learning techniques used in the study were less accurate. This study failed to achieve the desired results (Table 1).

Naïve Bayes and two feature selection methods (ReliefF/Limma) were applied to classify different lung cancer subtypes [34]. This study used the gene expression dataset GDS3257, and DNA methylation data from TCGA. 19 genes were associated with adenocarcinoma, squamous cell carcinoma and carcinoma of the lung. The selected genes were

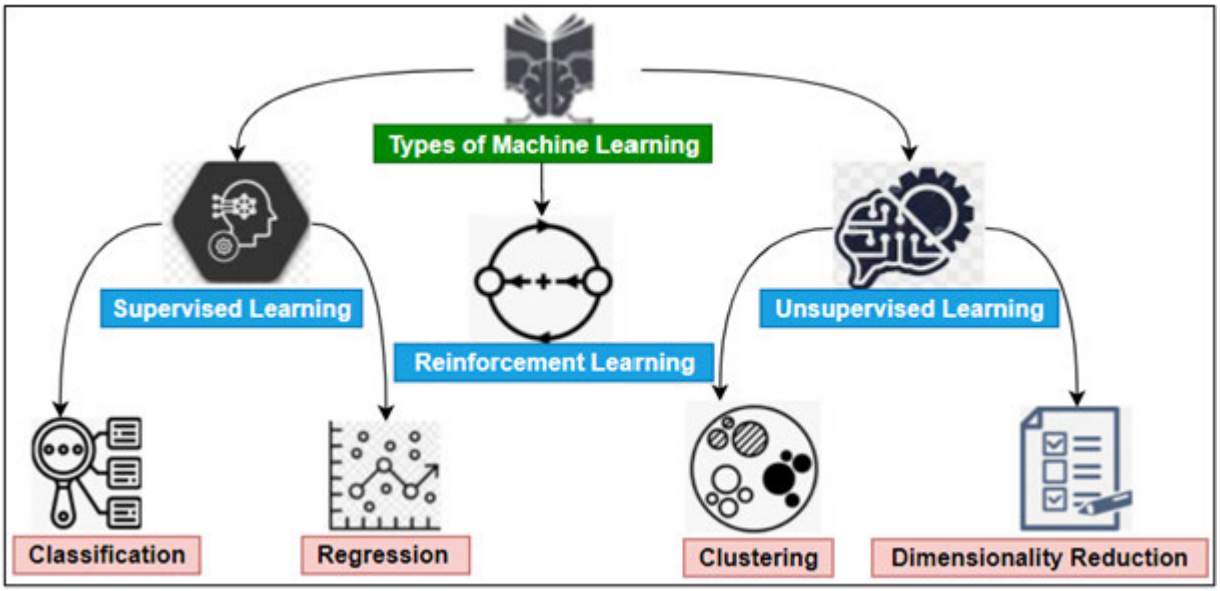

FIGURE 1. Types of machine learning.

selected employed as a gene signature for predicting lung cancer subtypes. The performance of the proposed model was 0.89 for both data (GDS3257 and TCGA DNA methylation). This study is limited as it lacked evaluation of the model performance with prospective data.

TABLE 1. Comparing the performance of ML approaches for analyzing breast cancer dataset.

Methods	Accuracy	TP	FP	TN	FN	F-measure
RF	70%	58%	19%	81%	42%	59 %
SVM	69%	49%	16%	84%	51%	53 %
C4.5	60%	47%	26%	74%	53%	47 %
NB	57%	45%	29%	71%	55%	44 %
KNN	49%	25%	16%	84	75%	31 %

Pineda *et al.* used gene expression (TCGA and GEO) and DNA methylation data (TCGA) to develop a classifier that effectively discriminated lung adenocarcinoma from lung squamous cell carcinoma cases [61]. The dataset for both lung cancer subtypes were collected from [62]–[64]. They applied a feature selector (ReliefF/Limma) to select 30 top relevant scoring variables associated with these lung cancer subtypes. This was reduced from 27,578 DNA methylation variables and 17,814 genes from microarray gene expression. The study achieved an AUC classification performance of 0.89 by applying a Naïve Bayes classifier and gene functional analysis using Ingenuity Pathway Analysis tool (IPA) and identified 19 genes of which 4 were specifically associated with lung cancer subtypes (AKR1B10, AQP10, CXCR2, TP73).

Four machine learning techniques were applied to four different sources of gene expression data for predicting lung cancer [33]. The datasets were from Harvard Medical School consisting of 203 samples probing expression of 12,600 genes; the University of Michigan (192 samples and 7129 genes); the University of Toronto (63 samples and 2,880 genes) and Brigham and Women's Hospital (181 samples and 12,533 genes). The highest accuracy achieved was between 0.88-0.94 when the Brigham and Women's Hospital dataset was used, and SVM applied. In addition, 0.83 was the highest accuracy accomplished when applying the C4.5 decision tree on the University of Toronto dataset.

Yuan *et al.* [38], used RF and SVM to classify two subtypes of lung cancer: (Adenocarcinomas (AC) and Squamous Cell Carcinomas (SCC)). They also applied Monte-Carlo (MCSF) and incremental feature selection methods to identify informative genes. Affymetrix U133 arrays (probing 20,502 genes) were used to generate data from 77 lung AC and 73 lung SCC samples from Gene Expression Omnibus (GEO GSE43580). The study showed that when 1100 optimal features (genes) were selected for classification using an SVM classifier, higher accuracy was achieved compared with using 43 informative features (genes) obtained using a MCSF method. Accuracy decreased from 0.96 to 0.86 using SVM and 0.93 to 0.88 with RF (Table 2).

TABLE 2. The performance of RF and SVM to analyze lung cancer subtypes.

Methods	Number of g nes	Accuracy	Precisio n	Recall	F1 score
SVM	1100	96%	93%	100%	96%
SVM	43	86%	80%	98%	88%
RF	260	93%	89%	98%	93%
RF	43	88%	82%	97%	89%

Tarek *et al.* [25], proposed a method for classifying three types of cancer; Leukemia, Colon, and breast cancer by applying a KNN algorithm and three feature selection methods to reduce the dimensionality and to select significant

features to improve cancer prediction. The dataset used for testing the system again was obtained from TCGA. The dataset had 6,500 colon, 24,481 breast and 3,571 leukemia samples. KNN was applied with three feature selection (Singular Value Decomposition Entropy (SVDE), Extreme Value Distribution (EVD), and Backward Elimination Hilbert-Schmidt Independence Criterion (BASIC)). The accuracy achieved with the system was 0.80 for colon cancer, 0.91 for breast cancer, and 0.92 for leukemia.

A deep learning approach with a stacked denoising autoencoder (SDAE) algorithm as a feature selection method was proposed to select informative genes that distinguish breast cancer samples from normal breast tissue samples. RNA-Seq gene expression data was analyzed [19]. This approach was applied to 1097 cancer samples and 113 healthy controls. The data was downloaded from The Cancer Genome Atlas (TCGA) [20] for both sample types. To evaluate the performance of the proposed algorithm, three classifier techniques which included artificial neural network (ANN), Support Vector Machine (SVM) and SVM-radial basis function (SVM-RBF)) were applied. The system achieved an accuracy of 0.91, 0.91 and 0.94 for ANN, SVM, and SVM-RBF, respectively.

Yeganeh *et al.* [46], used different machine learning techniques with multiple GEO ovarian cancer datasets (GSE12172, GSE14407, GSE9899, GSE37648, GSE18521, GSE38666, and GSE10971) for predicting ovarian cancer (n=530 cases). They used a 26-gene set panel for training different machine learning predictive models. The study achieved the highest accuracy of 0.89 when using a Random Forest pipeline. The drawbacks of the study used imbalanced dataset, and the achievement results require improvements.

In a study of microarray gene expression data from Lung adenocarcinoma samples (86 tumor samples and 10 non-tumor samples collected from Kent Ridge Bio-Medical Dataset Repository available from [60] with 7129 genes) an Info gain feature selection technique was applied to identify genes strongly associated with cancer samples using 70% of samples as a training set and 30% of samples as a test set. This study applied three classifier techniques to discriminate tumor and non-tumor samples after choosing the candidate genes that had known relevance to lung cancer [2]. Several selected genes were evaluated for biological relevance in lung cancer pathology. The system tested on the dataset provided an output of six genes with high Info Gain scores that might be linked with lung cancer (FABP4, FHL1, CLEC3B, Monoamine Oxidase-A, Platelet endothelial cell adhesion molecule-1, and Selenoprotein P). Additionally, the system employed these as biomarker genes to classify lung cancer by applying three classifier techniques (Multilayer Perceptron (MLP), Random subspace (RSS) and (Sequential Minimal Optimization (SMO)). The accuracy achieved was 0.87, 0.68, and 0.92 MLP, RSS, and SMO, respectively (Table 3).

Bhalla *et al.* [48], applied SVM and RF based models on 523 clear cell renal cell carcinomas (ccRCC)). The main aim of this study was to identify a minimum number of biomarker genes that would effectively discriminate early from late stage ccRCC and therefore would be effective in cancer staging. The models used identified an eight-gene signature (4 upregulated and 4 downregulated) that achieved a ROC of 0.77 with accuracy of 70.19%. The authors attempted to develop gender specific models as well to improve performance prediction and whilst some evidence of increased specificity was highlighted further work is required.

TABLE 3. Comparing the performance of MLP, RSS and SMO to classify lung cancer data.

Methods	Accuracy	Precision	Recall
MLP	86.6%	87%	83%
RSS	68.3%	64%	60%
SMO	91%	91%	90%

A novel multi-view feature selection method was used to analyze gene expression (RNA-Seq) data in combination with copy number alteration and protein array data to predict renal clear cell carcinoma (KIRC) survival [49]. eXtreme Gradient Boosting (XGB) was applied for training and testing the genes selected by the multi-view feature algorithm and rely on canonical correlation analysis (CCA). The study achieved 0.76 accuracy. One of the notable limitations of the proposed feature selection methods was that it was performed using unsupervised CCA that may lead to reduced accuracy. Additionally, the study had a quite low accuracy score so it requires some enhancement which might be achieved using supervised variants of the CCA method [50], [51]. The study again did not used additional sources of data for evaluating the proposed model efficiency.

Xu *et al.* [55], proposed a Multi-Grained Cascade Forest (gcForest) and dependent feature selection strategy for predicting four subtypes of breast cancer (Basal, Her2, Luminal A, and Luminal B). Again, TCGA RNA-Seq data was used and feature selection was developed for selecting 30 informative genes used for improving classification accuracy and reducing training time. The study compared the gcForest classifier with three different machine learning approaches (KNN, SVM and MLP). gcForest showed higher accuracy scores compared with the other classifiers. 0.92 accuracy was accomplished in this study. Although the research yielded valuable results. However, it has some caveats. The gcForest classifier works under the decision tree principle so it is poorly suited for processing continuous gene expression data and must perform discretization of the data which leads to information loss. Additionally, the study did not use external data for evaluating the proposed model.

Additional studies of conventional machine learning approaches that have been employed using cancer gene expression data are presented (Table 4) and includes the datasets and techniques that have been developed, as well as accomplished results, and references.

TABLE 4. Summary of traditional ML approaches applied to gene expression data.

Dataset	Techniques	Accuracy	Reference
Kent Ridge Bio-Medical Dataset Repository.	Info gain,SMO,RSS,MLP	86.6%,68.3% ,91%, MLP,RSS,SMO respectively	[2]
Breast Cancer from TCGA 1097 breast cancer samples and 113 healthy samples.	Stacked denoising Autoencoders (SDAE),SVM,SVM-RBF,ANN	91.74%, 91.74%, and 94.78%, ANN, SVM, and SVM-RBF, respectively	[19]
Breast cancer	RF,SVM,NB,C4.5, K-NN	RF=70%, SVM=69% NB=57%, C4.5=60% KNN=49%	[21]
leukemia, colon and breast cancer	K-NN and three feature selection techniques BAHSIC, EVD, SVD	92%,80% ,91% leukemia, colon, breast cancer respectively.	[25]
GEO(leukemia cancer, prostate and colon cancer data	signal to noise ratio, Fisher,ReliefF,T-statistics SVM,K-NN	Between 85% and 100%.	[27]
TCGA(Lung cancer	ReliefF,RF	83%	[30]
GSE4922, GSE2034, GSE6532, GSE7390 GSE11121 Breast cancer	To identify informative genes are used: 1.Principal component analysis algorithm 2.Autoencoder neural network. AdaBoost algorithm (PCA-AE-Ada) was constructed to predict clinical outcomes in breast cancer	The accuracy of this study was between 75%,72%,77%,75% and 85% GSE4922, GSE2034, GSE6532, GSE7390 and GSE11121 respectively.	[31]
Breast cancer	SVM, KNN, MLP, DT, RF, Logistic Regression (LR), Ada (Adaboost), GBM (Gradient Boosting Machines, Recursive Feature Elimination (RFB), Randomized Logistic Regression (RLR)	88.8% was the highest accuracy when RFE and SVM applied. While 87% the highest accuracy when RLR and SVM applied.	[32]
Harvard Medical School, University of Michigan , University of Toronto , Women's Hospital Breast cancer	SVM, C4.5	88%-94% when the Brigham and Women's Hospital dataset was use 83% with C4.5 .	[33]
GDS3257,DNA methylation from TCGA	NB, ReliefF,Limma	89%.	[34]
GEO,TCGA(Three lung cancer subtypes)	mRMR,Multi-category Receiver Operating Characteristic (Multi-ROC) Incremental Feature Selection (IFS) Random Forests, multiclass support vector	86.5% .	[36]
GEO(GSE43580)	RF, SVM	86%-96% with SVM 88%-93% with RF.	[38]
GEO(GSE161741)	Binary classification algorithm	97for positive(has cancer) 73% for negative (normal).	[39]
GEO(GSE6044, GSE2109) Harvard	Structural Binary Classification (SBC)	93%.	[41]
TCGA(LUAD and LUSC)	DGE ,PCA,mRMR Lasso,Xgboost,Overlapping,RF	92.9%	[42]
GEO(Lung,Ovarian and Colon)	Mutual Information(MI) Genetic Algorithm()GA,SVM	80% -98%	[43]
GEO(GSE12172, GSE14407, GSE9899, GSE37648, GSE18521, GSE38666, and GSE10971)	RF	89%.	[46]
clear cell renal cell carcinomas (ccRCC) from TCGA	SVM,RF	70%.	[48]
Renal clear cell carcinoma (KIRC) from TCGA	multi-view feature selection eXtreme Gradient Boosting (XGB)	76%.	[49]
Central Nervous System tumors	Extreme Learning Machines (ELM) Correlation Coefficient	79%.	[52]
Breast Cancer subtypes from TCGA	MLP,SVM,KNN and Multi-Grained Cascade Forest (gcForest)	92% was highest accuracy with gcForest .	[55]
GEO,TCGA	NB	89%.	[61]
GEO Colon cancer, Acute leukemia , Prostate tumor, High-grade Glioma Lung cancer II, Leukemia2 data	,Indipendent Component Analysis (ICA) Artificial Bee Colony (ABC) ,NB	Between 92% to 98%.	[72]
GEO Colon cancer, Acute leukemia , Prostate tumor, High-grade Glioma Lung cancer II, Leukemia2 data	,Indipendent Component Analysis (ICA) Artificial Bee Colony (ABC) ,NB,SVM	Between 92% to 98% with NB and 93% to 97% with SVM	[73]
GEO Colon cancer, Acute leukemia , Prostate tumor, High-grade Glioma Lung cancer II, Leukemia2 data	,Indipendent Component Analysis (ICA) Artificial Bee Colony (ABC) ,Artificial Neural Networks(ANN)	93% to 98% with ANN	[74]
GEO Colon cancer, Acute leukemia , Prostate tumor, High-grade Glioma Lung cancer II, Leukemia2 data	,Indipendent Component Analysis (ICA) Genetic Bee Colony (GBC) ,Artificial Neural Networks(ANN)	96% to 99% with ANN	[75]

2) DEEP LEARNING APPROACHES

DL also called deep neural network (DNN) learning has shown some major breakthroughs in recent years due to the increase in computation power. DL can be defined as a sub-field of machine learning that works by generating structure that has multiple layers in which the next layer of input is the output of the previous layer (Figure 2) [4]. DL structure aims to mimic the mechanisms of the human brain by interpreting the various types of data including sound, text, and images [5]. DL uses principles similar to that of linear regression, where each neuron has a weighted value that is updated by applying a gradient descent algorithm through back-propagation to reduce global loss of function [6]. DL approaches contributed to overcoming the difficulties of cancer prediction by speeding up analysis whilst maintaining accuracy. Most common architectures are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Artificial Neural Networks (ANN).

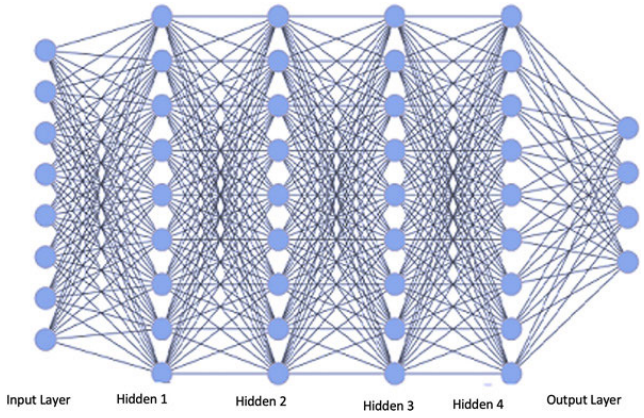

FIGURE 2. Deep learning structure.

Practically, DL is a subset of ML and functions in a similar way. Traditional ML and DL approaches both allow computers to learn from input data without the requirement of explicit programming, but they are technically different. DL does not require human effort to generate feature extraction in the way that traditional ML does (Figure 3) [7]. DL is also more efficient than ML in how it uses very large datasets. Recently, DL shows higher performance than human performance on tasks that involves images classifications (Figure 4) [7].

Deep Neural Network (DNN) was applied to analyze gene expression data for early pan-cancer prediction [24]. The system was tested for 37 different types of cancer. Again, TCGA data was used for the 37 cancer types. The dataset consisted of 10,663 samples (9,807 tumors and 856 normal samples) across 37 cancer types measuring expression of 10,000 genes. The system applied the DNN model with three different structures: 3NN, 5NN, 9NN, and compared the accuracy with SVM. The system applied two feature reduction methods: Prior Knowledge was applied to identify a set of genes known to be involved in relevant biological pathways, whereas an autoencoder was used for extracting informative genes from the input data (Table 5). Although, the study achieved good accuracy (i.e. with 5NN), more enhancements are needed due to the current sensitivity of the cancer classification field.

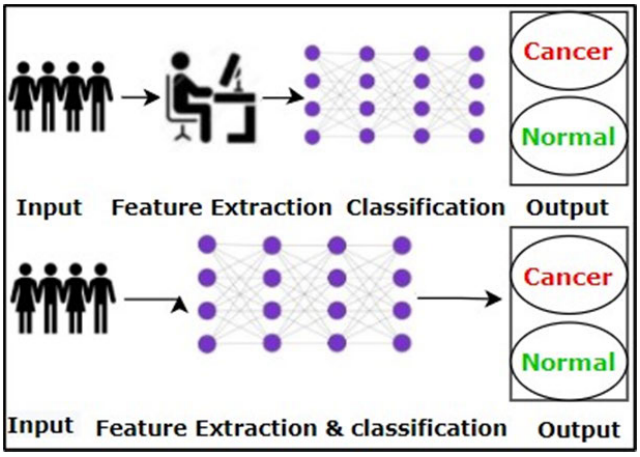

FIGURE 3. Deep learning vs conventional machine learning.

TABLE 5. Comparing the performance of DNN against SVM.

Methods	Accuracy	Precision	Recall	F1 score
SVM	82%	79%	83%	82%
3NN	84%	80%	89%	84%
5NN	90%	90%	90%	90%
9NN	83%	80%	88%	83%

Yawen *et al.* [71], has proposed a deep learning based multi-model ensemble for three types of cancer (Lung, Breast, and Stomach). The dataset was obtained from (TCGA) 162,271 and 878 samples from Lung, stomach and breast respectively. The accuracy was 98% for all the datasets used in this study. Although, this system has accomplished impressive results and employed good sample sizes for training and testing the proposed model, the ensemble model has high complexity and which requires more time for training. Additionally, the ensemble model is difficult to interpret.

Matsubara *et al.* [35], used CNN (combining spectral clustering information processing) to classify lung cancer using both protein interaction network data and gene expression data from 639 samples (152 benign and 487 malignant). The dataset is available from NCBI GEO datasets (ID GSE66499). This study achieved 0.81, 0.88, 0.78 and 0.74 accuracy, recall, precision and specificity, respectively. This study, as for the others described did not employ validation data to check the efficiency of the model. Moreover, 190 out of 487 cancer samples were randomly chosen and this explain the results obtained. (190) malignant samples were randomly selected, and that would not be efficient.

Zeebaree *et al.* [37], used a CNN deep learning algorithm with microarray gene expression data across eight cancer types. The largest tumor cohorts used in this study was 286 samples for breast cancer probing expression of 13321 genes, while the smallest cohort size was for

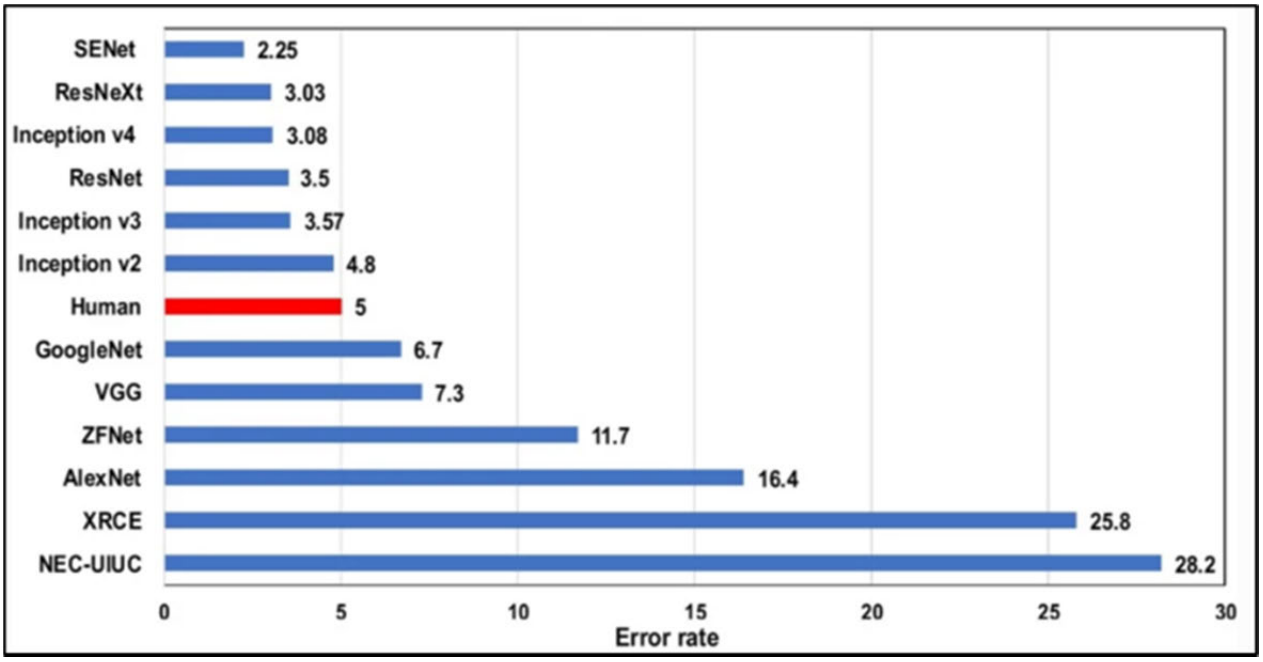

FIGURE 4. Deep learning performance vs human preformance.

brain cancer (42 samples probing 5597 genes). The overall sample size for most tumor types was small compared with prior studies. The study had the lowest accuracy of 0.41 for one of the two breast cancer datasets using CNN. However, in comparison to alternative approaches (mSVMRFE-IRF and varSelRF) CNN is typically superior in terms of accuracy and minimizing the gene numbers used for classification.

Sun *et al.* [22], proposed a novel multimodal deep neural network (MDNN) algorithm for breast cancer prediction. This system was applied to publicly available data [23] and examined 24368 genes across 2509 breast cancer and 548 normal samples. The proposed algorithm achieved higher accuracy when compared with SVM, Random Forest RF, and Logistic regression (LR). Minimum Redundancy Maximum Relevance (mRMR) was also applied as a feature selection method to reduce the number of features (genes) to enhance accuracy. The accuracy achieved was 0.82, 0.80, 0.79 and 0.76 for MDNN, SVM, RF, and LR respectively. However, recall values were low in this study (0.45, 0.36, 0.22 and 0.18 for MDNN, SVM, RF, and LR respectively) and precision was 0.95 for all algorithms. Although this study achieved satisfactory accuracy, further enhancements are required due to the sensitivity of the cancer classification domain. Additionally, recall values were very low negatively impacting the proposed model. The study was tested only on a breast cancer dataset whereas, most studies have used multiple cancer datasets to prove the validity of the results that have been obtained with their models.

Muhamed *et al.* [47], used Long Short-Term Memory (LSTM) as a classifier technique and Matthews Correlation Coefficient as a feature selection method for classification of five subtypes of kidney cancer using microRNA (miRNA) gene expression data. The dataset for the five subtypes of kidney cancer was obtained from TCGA dataset. The study achieved an overall accuracy between 0.88-0.92 identifying 35 miRNAs with strong discriminative ability for renal cancer subtypes. Limitations of this study were the lack of replication in a new dataset to test the efficiency of the suggested model and the models were used on datasets that were not balanced in terms of equal sample sizes between the different cancer subtypes.

Anika *et al.* [26], proposed a CNN-based model that uses gene expression data (from TCGA) for predicting 20 types of cancer. The dataset used 1,881 samples across these 20 cancers profiling expression of 60,383 genes. The model achieved accuracy between 0.52 to 0.78 when applied to a single cancer from the 20 types of cancer, whereas the accuracy was between 0.66 to 0.93 when applied across the pan-cancer dataset.

A novel Deep Flexible Neural Forest (DFNForest) model was tested as an alternative to deep neural networks to classify subtypes of three different tumors (Lung, Breast and Glioblastoma multiforme) using TCGA RNA-Seq data [56]. The study combined two feature reduction methods (fisher ratio and neighborhood rough set) to reduce the dimensionality of the data, avoid overfitting and select informative genes [57]. The Novel DFNForest model achieved 0.93, 0.88 and 0.84 accuracy in classifying subtypes of breast, lung and GBM cancers, respectively. The study highlights the variability in effectiveness of subtype classification across different tumor types.

A differential regulatory network embedded deep neural network (DRE-DNN) approach was developed from a canonical DNN and applied to predict liver cancer (hepatocellular carcinoma) outcomes using three datasets (GEO GSE10143 and GSE14520 and TCGA) [45]. The proposed model achieved 0.86, 0.74 and 0.72 average AUC values for GSE10143, GSE14520 and TCGA datasets respectively which was improved on conventional DNN AUC values. The study used different sources of data for validation and measuring the performance of the proposed method. It used sufficient dataset for train the DRE-DNN model, and whilst it was useful as a tool for prognosis it did not accomplished good results for classification purposes. However, it goes some way to addressing the overfitting problem of the model. A novel hybrid filter wrapper feature selection method has been described [53] for selecting a subset of informative genes for diagnosis and classification. CNN and ReliefF were applied for classifying different cancer microarray datasets (Ovarian, Leukemia and CNS). For the CNS data, 60 samples probed for 7129 genes were used for testing the feature selection and classifier technique. ReliefF was applied to select a subset of informative genes to increase the performance of the CNN and reduce time for training the model. The ReliefF-CNN method achieved 0.83 (increased from 0.65) accuracy with CNS data. Based on the studies that have been done for classifying CNS, there are some challenges that need to be addressed. Comparatively small sample cohort sizes were used from CNS patients as this was freely available. Therefore, multicenter studies are required that provide larger datasets for analysis. The accuracy scores obtained are suboptimal and may need improvement to inform clinical decision making. Techniques are required that have high efficiency with small datasets such us Learning vector quantization.

Guillermo *et al*. [70], has proposed CNN and transfer learning (TL) for lung cancer prediction. CNN has been used to extract features from high dimensional dataset. The dataset TCGA for 33 different types of cancer(10535 samples and top 20K most variably expressed genes) were used but the study focused on the lung cancer dataset to test the proposed model. The highest accuracy was 68 %, 72% and 69% CNN, densely-connected multi-layer feed-forward neural network(MLNN) and SVM respectively. The investigation had quite low accuracy scores and was limited to one type of cancer. Examination of other cancer types may not ac hive the same accuracy. Other evaluation measurements from this study were determine (Table 6).

Additional studies of deep learning approaches that have been used on cancer gene expression data in the cancer are (Table 7). The Table includes the datasets, techniques that have been developed, accomplished results, and references.

TABLE 6. Comparing the performance of CNN against MLNN and SVM.

Methods	AUC	Sensitivity	Specificity	Accuracy
CNN	73%	67%	68%	68%
MLNN	70%	61%	73%	72%
SVM	70%	64%	69%	69%

D. LIST OF COMMON DATASETS

Here, common data repositories for analyzing cancer gene expression data that helps researchers to test their proposed models (Table 8) are discussed.

III. THE CHALLENGES OF ANALYZING AND USING GENE EXPRESSION DATA FOR CLINICAL USE

- Although gene expression datasets are large in terms of volume, the datasets commonly have quite small cohort sizes but with a very large number of associated variables (e.g., gene expression values). This is a recognized issue for DL as well as (but less demanding) for ML algorithms [69]. In cancer genomics, there are several repositories providing access to high quality, curated public data allowing training of DL models. However, pre-processing and harmonization are required across these and newly developed datasets.
- Large public data resources for gene expression in cancer is limited to a few key sources (e.g., TCGA and GEO). Most Deep learning techniques require big data to develop accurate models that can be applied to new cancer datasets. Researchers have attempted to mitigate these existing issues using different techniques such as regularization methods (ridge and lasso or L1 and L2), dropout, data augmentation, reduction of NN complexity to enhance the performance of the model. However, they did not completely solve this issue.
- Early cancer prediction and classification enhancement with very high accuracy is necessary and could be achieved by developing mathematical methods with less complexity and are less computationally time consuming. For some tumor types (e.g., gallbladder cancer) classification, prediction or gene biomarker identification studies using ML or DL is limited.
- More studies on personalized treatments using deep learning approaches for complex genetic diseases are warranted.
- Identifying individual gene signatures for each cancer type is important because it may contribute to early diagnosis of disease. In addition, knowledge of these gene pathways could have a significant impact in determining the underlying pathology of these tumors and potentially druggable path-ways.
- Some gene pathways may be implicated across multiple cancer types and identifying these may aid risk prediction for individual patients.
- Cancer subtype classification algorithms need improving and extending across different tumor type to facilitate optimal patient treatment.
- A notable limitation of AI in analyzing gene expression data is known as ''The curse of dimensionality'' [68] in which higher dimensional data may reveal random effects that do not replicate in related patient cohorts.
- Only a few studies have focused on analyzing gene expression of Liver, Kidney and CNS types of cancer

TABLE 7. Summary of DL approaches applied to gene expression data.

Dataset	Techniques	Accuracy	Reference
TCGA(Breast cancer).	MDNN, SVM, RF, and LR	82%, 80%, 79% and 76% for MDNN, SVM, RF, and LR respectively	[22]
TCGA(37) types of cancer.	SVM,DNN(3NN,5NN,9NN)	82%,84%,90% and 83% SVM,3NN,5NN and 9NN respectively	[24]
TCGA(20 types of cancer).	CNN-based	Between 52% to 78% when applied to a single cancer %66 to 93% when applied across the pan-cancer dataset	[26]
TCGA(BRCA,COAD and KIPAN)	standard Lasso, DeepNeti and DeepNetii	65%,62%and 65% standard Lasso,DeepNeti and DeepNetii respectively for BRCA data. 77%,72%and 75% standard Lasso,DeepNeti and DeepNetii respectively for KIPAN data. 57%,58%and 57% standard Lasso,DeepNeti and DeepNetii respectively with COAD data.	[28]
TCGA	Ensemble of LASSO	68%	[29]
GEO(GSE66499).	CNN	81%	[35]
GEO(10 types of cancer).	mSVM-RFE-iRF,varSeIRF, CNN	Between 49% to 93% with varSeIRF %71 to 95% mSVM-RFE-iR with Between 41% to 100% with CNN	[37]
TCGA(mRNA,Methylation, Copy number variation(CNV).	SVM,Integrative deep learning(IDL)	80% and 60% with SVM and IDL respectively	[40]
GEO(GSE68465 and GSE8894)	Consistency Subset Feature Selection(CSF), CSFS Wrapper Subset(WS) SVM, Gene Expression Programming (GEP) Deep gene selection (DGS)	84%,83%,83%,85%, 85%,87%,CSF,CSFS,WS,SVM,GEP DGS respectively and five gene selected to be associated with lung cancer	[44]
TCGA,GEO(GSE10143 and GSE14520)	DRE-DNN	86%,74% and 72% average AUC values for GSE10143, GSE14520 and TCGA respectively	[45]
TCGA(five kidney cancer subtypes)	LSTM Matthews Correlation Coefficient	88% to 92%	[47]
GEO(Central Nervous System)	ReliefF,CNN	83%	[53]
GEO(copy number alterations(CNA), Gene expression) Breast cancer subtypes	Deep CNN(DCNN)	62% with CNA data 62% with Gene expression data and 96% with both	[54]
TCGA(Lung,Breast, Glioblastoma multiforme(GBM)) cancer subtypes	DFNForest Fisher ratio and Neighborhood rough set	93%, 88% and 84% Breast,Lung and GBM respectively	[56]
TCGA(PANCANCER,BRCA,GBM,LUNG) GEO(Adenocarcinoma,Colon,Brain)	Boosting Cascade Deep Forest(BCDForest) KNN, LR, RF, SVM ,gcForest	Highest accuracy was between (41% , 97%) with TCGA data between (92% , 96%) with GEO data BCDForest Applied on Both(GEO,TCGA)	[58]
TCGA(Breast Cancer subtypes)	CNN	88%	[59]
TCGA(Lung cancer)	CNN,MLNN,SVM TL,ANOVA	Highest accuracy was 68%, 72% and 69% CNN+TL,MLNN+TL and ANOVA+SVM respectively	[70]
TCGA(Lung,Breast,Stomach cancer)	Deep learning-based multi-model ensemble	98%	[71]
GEO	Incremental feature selection(IFS) Minimum redundancy maximum relevance (mRMR) Recurrent neural network (RNN)	73.9% for normal tissues and 63.9% for cancers	[76]

TABLE 8. List of common dataset websites.

Dataset name	Description	Number of samples	Reference
GEO	It is a public functional genomics data repository supporting MIAME-compliant data submissions	3635328	[65]
ArrayExpress	It archives the dataset from genomic experiments.	57004	[66]
The Cancer Genome Atlas (TCGA)	It is a dataset for different types of cancer that contained gene expression data	84,031	[67]

using machine learning. For instance, CNS classification achievements were meagre [52], [53].

IV. CONCLUSION

In this paper, we provide a detailed review of Machine Learning (ML) studies using gene expression data across major cancer types (e.g. Lung, Breast, CNS, Kidney, Ovarian, Liver and Gallbladder). These studies use ML for different purposes including cancer identification, cancer subtype classification, identification of gene biomarkers and prognosis. In exploring these studies, the most common tools for measuring gene expression between benign and malignant

tissues have been identified. We highlight datasets commonly employed when using gene expression data to test ML models under development. Several challenges remain in analyzing gene expression data and those challenges can be used as sign posts for researchers beginning their studies in this field and may lead to new results that could help in improving in cancer classification and ultimately personalized treatments. In conclusion, deep learning approaches are overcoming the issues of traditional machine learning techniques in analyzing gene expression data for cancer.

REFERENCES

[1] S. Shandilya and C. Chandankhede, "Survey on recent cancer classification systems for cancer diagnosis," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2017, pp. 2590–2594.
[2] J. Pati, "Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232–4238, 2019.
[3] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018.
[4] W. J. Zhang, G. Yang, Y. Lin, C. Ji, and M. M. Gupta, "On definition of deep learning," in *Proc. World Automation Congr. (WAC)*, Stevenson, WA, USA, Jun. 2018, pp. 1–5.
[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[6] 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, Mar. 2020.
[7] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, Dec. 2021.
[8] C. A. Ul Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," in *Proc. 24th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2018, pp. 1–6.
[9] J. Wu and C. Li, "Feature selection based on features unit," in *Proc. Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Jul. 2017, pp. 330–333.
[10] A. A. Shanab and T. Khoshgoftaar, "Filter-based subset selection for easy, moderate, and hard bioinformatics data," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2018, pp. 372–377.
[11] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam Res. Paper Bus. Anal.*, vol. 30, pp. 1–25, Mar. 2017.
[12] S. Liu, C. Xu, Y. Zhang, J. Liu, B. Yu, X. Liu, and M. Dehmer, "Feature selection of gene expression data for cancer classification using double RBF-kernels," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–14, Dec. 2018.
[13] M. M. Babu, "An introduction to microarray data analysis," *Comput. Genomics*, pp. 225–249, Nov. 2003.
[14] R. Govindarajan, "Microarray and its applications," *J. Pharmacy Bio Allied Sci.*, vol. 4, no. 6, p. 310, 2012.
[15] A. Wolff, M. Bayerlová, J. Gaedcke, D. Kube, and T. Beißbarth, "A comparative study of RNA-seq and microarray data analysis on the two examples of rectal-cancer patients and burkitt lymphoma cells," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0197162.
[16] M. S. Rao, T. R. Van Vleet, R. Ciurlionis, W. R. Buck, S. W. Mittelstadt, E. A. G. Blomme, and M. J. Liguori, "Comparison of RNA-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies," *Frontiers Genet.*, vol. 9, pp. 1–16, Jan. 2019.
[17] Y. Guo, Q. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr, "Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data," *PLoS ONE*, vol. 8, no. 8, Aug. 2013, Art. no. e71462.
[18] Y. Piao and K. H. Ryu, "Detection of differentially expressed genes using feature selection approach from RNA-seq," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2017, pp. 304–308.
[19] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," in *Proc. Pacific Symp. Biocomput.*, 2017, pp. 219–229.
[20] J. N. Weinstein, E. A. Collisson, and G. B. Mills, "The cancer genome atlas pan-cancer analysis project John," *Chin. J. Lung Cancer*, vol. 18, no. 4, pp. 219–223, 2015.
[21] S. Vural, X. Wang, and C. Guda, "Classification of breast cancer patients using somatic mutation profiles and machine learning approaches," *BMC Syst. Biol.*, vol. 10, no. S3, pp. 264–276, Aug. 2016.
[22] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 841–850, May 2019.
[23] TCGA. (2012). *The Somatic Mutation Profiles of 2,433 Breast Cancers Refines Their Genomic and Transcriptomic Landscapes*. [Online]. Available: https://www.cbioportal.org
[24] C. Li and M. Zhang, "Deep learning in pan cancer early detection based on gene expression," Stanford Univ., Stanford, CA, USA, Tech. Rep. cs230, 2018.
[25] S. Tarek, E. R. Abd, and M. Shoman, "Gene expression-based cancer classification," *Egyptian Inform. J.*, vol. 18, no. 3, pp. 151–159, 2017.
[26] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, Jul. 2019.
[27] S. H. Bouazza, N. Hamdi, A. Zeroual, and K. Auhmani, "Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers," in *Proc. Intell. Syst. Comput. Vis. (ISCV)*, Mar. 2015, pp. 1–6.
[28] D. Urda and F. Moreno, "Deep learning to analyze RNA-seq gene expression data," in *Proc. Int. Work-Conf. Artif. Neural Netw.*, vol. 3, 2017, pp. 50–59.
[29] D. Urda, L. Franco, and J. M. Jerez, "Classification of high dimensional data using LASSO ensembles," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–7.
[30] J. Li, T. Ching, S. Huang, and L. X. Garmire, "Using epigenomics data to predict gene expression in lung cancer," *BMC Bioinf.*, vol. 16, no. S5, pp. 1–12, Dec. 2015.
[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] S. Turgut, M. Dagtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods," in *Proc. Electric Electron., Comput. Sci., Biomed. Eng. Meeting (EBBT)*, Apr. 2018, pp. 1–3.
[33] M. D. Podolsky, A. A. Barchuk, V. I. Kuznetcov, N. F. Gusarova, V. S. Gaidukov, and S. A. Tarakanov, "Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels," *Asian Pacific J. Cancer Prevention*, vol. 17, no. 2, pp. 835–838, Mar. 2016.
[34] A. L. Pineda, H. A. Ogoe, J. B. Balasubramanian, C. R. Escareño, S. Visweswaran, J. G. Herman, and V. Gopalakrishnan, "On predicting lung cancer subtypes using 'omic' data from tumor and tumor-adjacent histologically-normal tissue," *BMC Cancer*, vol. 16, no. 1, p. 184, Dec. 2016.
[35] T. Matsubara, J. C. Nacher, T. Ochiai, M. Hayashida, and T. Akutsu, "Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles," in *Proc. IEEE 18th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2018, pp. 151–154.
[36] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S.-M. Ngai, and J. Shao, "Classification of lung cancer using ensemble-based feature selection and machine learning methods," *Mol. BioSyst.*, vol. 11, no. 3, pp. 791–800, 2015.
[37] D. Q. Zeebaree, H. Haron, and A. M. Abdulazeez, "Gene selection and classification of microarray data using convolutional neural network," in *Proc. Int. Conf. Adv. Sci. Eng. (ICOASE)*, Oct. 2018, pp. 145–150.
[38] F. Yuan, L. Lu, and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," *Biochimica et Biophysica Acta (BBA) Mol. Basis Disease*, vol. 1866, no. 8, Aug. 2020, Art. no. 165822.
[39] N. Garikipati, "Computational genomic algorithms for miRNA-based diagnosis of lung cancer: The potential of machine learning," *Genomics*, pp. 1–23, Nov. 2016.
[40] J. Lim, S. Bang, J. Kim, C. Park, J. Cho, and S. Kim, "Integrative deep learning for identifying differentially expressed (DE) biomarkers," *Comput. Math. Methods Med.*, vol. 2019, pp. 1–10, Nov. 2019.
[41] H. Azzawi, J. Hou, R. Alnnni, and Y. Xiang, "SBC: A new strategy for multiclass lung cancer classification based on tumour structural information and microarray data," in *Proc. IEEE/ACIS 17th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2018, pp. 68–73.
[42] J. W. Chen and J. Dhahbi, "Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods," *Sci. Rep.*, vol. 11, no. 1, Dec. 2021, Art. no. 13323.

[43] M. Jansi Rani and D. Devaraj, "Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification," *J. Med. Syst.*, vol. 43, no. 8, p. 235, Aug. 2019.

[44] R. Alanni, J. Hou, H. Azzawi, and Y. Xiang, "Deep gene selection method to select genes from microarray datasets for cancer classification," *BMC Bioinf.*, vol. 20, no. 1, p. 608, Dec. 2019.

[45] J. Li, Y. Ping, H. Li, H. Li, Y. Liu, B. Liu, and Y. Wang, "Prognostic prediction of carcinoma by a differential-regulatory-network-embedded deep neural network," *Comput. Biol. Chem.*, vol. 88, Oct. 2020, Art. no. 107317.

[46] P. N. Yeganeh and M. T. Mostafavi, "Use of machine learning for diagnosis of cancer in ovarian tissues with a selected mRNA panel," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 2429–2434.

[47] A. M. Ali, H. Zhuang, A. Ibrahim, O. Rehman, M. Huang, and A. Wu, "A machine learning approach for the classification of kidney cancer subtypes using miRNA genome data," *Appl. Sci.*, vol. 8, no. 12, p. 2422, Nov. 2018.

[48] S. Bhalla, K. Chaudhary, R. Kumar, M. Sehgal, H. Kaur, S. Sharma, and G. P. S. Raghava, "Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer," *Sci. Rep.*, vol. 7, no. 1, p. 44997, Jul. 2017.

[49] Y. El-Manzalawy, "CCA based multi-view feature selection for multi-omics data integration," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Jun. 2018, pp. 1–8.

[50] N. Guan, X. Zhang, Z. Luo, and L. Lan, "Sparse representation based discriminative canonical correlation analysis for face recognition," in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, Dec. 2012, pp. 51–56.

[51] O. Arandjelović, "Discriminative extended canonical correlation analysis for pattern set matching," *Mach. Learn.*, vol. 94, no. 3, pp. 353–370, Mar. 2014.

[52] C. A. Kumar and S. Ramakrishnan, "Binary classification of cancer microarray gene expression data using extreme learning machines," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, Dec. 2014, pp. 1–4.

[53] S. Kilicarslan, K. Adem, and M. Celik, "Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network," *Med. Hypotheses*, vol. 137, Apr. 2020, Art. no. 109577.

[54] M. Mohaiminul Islam, S. Huang, R. Ajwad, C. Chi, Y. Wang, and P. Hu, "An integrative deep learning framework for classifying molecular subtypes of breast cancer," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 2185–2199, Jan. 2020.

[55] J. Xu, P. Wu, Y. Chen, and L. Zhang, "Comparison of different classification methods for breast cancer subtypes prediction," in *Proc. Int. Conf. Secur., Pattern Anal., Cybern. (SPAC)*, Dec. 2018, pp. 91–96.

[56] J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, and M. M. Khan, "A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data," *IEEE Access*, vol. 7, pp. 22086–22095, 2019.

[57] J. N. Weinstein, C. J. Creighton, C. Davis, L. Donehower, J. Drummond, D. Wheeler, A. Ally, M. Balasundaram, I. Birol, and Y. S. Butterfield, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, Sep. 2013.

[58] Y. Guo, S. Liu, Z. Li, and X. Shang, "Towards the classification of cancer subtypes by using cascade deep forest model in gene expression data," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 1664–1669.

[59] M. Mostavi, Y.-C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," *BMC Med. Genomics*, vol. 13, no. S5, pp. 1–13, Apr. 2020.

[60] A. Cano. (Nov. 20, 2019). *Kent Ridge Biomedical Data Set Repository. Retrieved From ELVIRA Biomedical Data Set Repository*. [Online]. Available: http://leo.ugr.es/elvira/DBCRepository/

[61] A. L. Pineda, H. A. Ogoe, J. B. Balasubramanian, E. C. Rangel, S. Visweswaran, J. G. Herman, and V. Gopalakrishnan, "On Predicting lung cancer subtypes using 'omic' data from tumor and tumor-adjacent histologically-normal tissue," *BMC Cancer*, vol. 16, no. 1, pp. 1–11, 2016.

[62] E. Collisson *et al.*, "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, vol. 511, no. 7511, pp. 543–550, Jul. 2014.

[63] M. Meyerson, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, pp. 519–525, Sep. 2012.

[64] M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen, S. E. Murphy, P. Yang, A. C. Pesatori, D. Consonni, P. A. Bertazzi, S. Wacholder, J. H. Shih, N. E. Caporaso, and J. Jen, "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLoS ONE*, vol. 3, no. 2, p. e1651, Feb. 2008.

[65] R. Edgar, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, Jan. 2002.

[66] A. Athar, A. Füllgrabe, N. George, H. Iqbal, L. Huerta, A. Ali, C. Snow, N. A. Fonseca, R. Petryszak, I. Papatheodorou, U. Sarkans, and A. Brazma, "ArrayExpress update—From bulk to single-cell expression data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D711–D715, Jan. 2019.

[67] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. S. C. Shmulevich, C. Sander, and Stuart, "The cancer genome atlas pan- cancer analysis project," *Nature Genet.*, vol. 45, no. 10, p. 1113, 2013.

[68] D. L. Barbour, "Precision medicine and the cursed dimensions," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–2, Dec. 2019.

[69] L. Wang, Y. Wang, and Q. Chang, "Feature selection methods for big data bioinformatics: A survey from the search perspective," *Methods*, vol. 111, pp. 21–31, Dec. 2016.

[70] G. López-García, 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, Mar. 2020, Art. no. e0230536.

[71] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Comput. Methods Programs Biomed.*, vol. 153, pp. 1–9, Jan. 2018.

[72] R. Aziz, C. K. Verma, and N. Srivastava, "A novel approach for dimension reduction of microarray," *Comput. Biol. Chem.*, vol. 71, pp. 161–169, Dec. 2017.

[73] R. A. Musheer, C. K. Verma, and N. Srivastava, "Novel machine learning approach for classification of high-dimensional microarray data," *Soft Comput.*, vol. 23, no. 24, pp. 13409–13421, Dec. 2019.

[74] R. Aziz and C. K. Verma, "Artificial neural network classification of microarray data using new hybrid gene selection method," *Int. J. Data Min. Bioinf.*, vol. 17, no. 1, pp. 42–65, 2017.

[75] R. Aziz, C. K. Verma, and N. Srivastava, "Artificial neural network classification of high dimensional data with novel optimization approach of dimension reduction," *Ann. Data Sci.*, vol. 5, no. 4, pp. 615–635, Dec. 2018.

[76] L. Chen, X. Pan, Y.-H. Zhang, M. Liu, T. Huang, and Y.-D. Cai, "Classification of widely and rarely expressed genes with recurrent neural network," *Comput. Structural Biotechnol. J.*, vol. 17, pp. 49–60, Jan. 2019.

MAHMOOD KHALSAN received the B.Sc. and M.Sc. degrees from the University of Northampton, where he is currently pursuing the Ph.D. degree in computer science. His main research interest includes how machine learning and gene expression assist in cancer prediction. His Ph.D. thesis work aims to bridge machine learning with biology to bring out the most probable genes associated with cancerous samples in order to advance the process of detecting cancer at early stages.

LEE R. MACHADO received the bachelor's degree from the University of Warwick, and the Ph.D. degree in cancer studies from the University of Birmingham, U.K. He is currently a Professor of molecular medicine with the Division of Life Sciences, a Faculty Research and Enterprise Lead, the Co-Leader of the Molecular Biosciences Research Group, Physical Activity and Life Sciences Centre, University of Northampton, and an Honorary Research Fellow with the Department of Genetics and Genome Biology, University of Leicester. He did postdoctoral research at the Institute for Cancer Studies, Birmingham, the MRC Toxicology Unit, Leicester, and the Department of Genetics, Leicester, before working as a Senior Scientist with Cancer Vaccine Company, and Scancell, Nottingham. He joined the University of Northampton, as a Lecturer, in 2013. He was an Interim Head of Sport, Exercise and Life Sciences, from 2017 to 2018. He has three years of University Board level experience. His research interests include employing cellular and molecular genetic strategies to address how the host immune system responds to pathogens and cancer. The aim of this work is to increase our understanding of human health and disease and develop rational therapeutic approaches to harness the exquisite specificity and sensitivity of the immune system.

EMAN SALIH AL-SHAMERY received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Babylon, Iraq, in 1998, 2001, and 2013, respectively. After completing her M.Sc., she worked as an Assistant Lecturer at the Department of Computer Science, University of Babylon. She is currently a Professor with the Software Department, University of Babylon. Her current research interests include artificial intelligence, bioinformatics, machine learning, neural networks, deep learning, and data mining.

SURAJ AJIT received the B.E. degree (Hons.) in computer science from Bangalore University, India, and the Ph.D. degree in computer science from the University of Aberdeen. He has five years of industry experience that includes working in BAE Systems as a Research Scientist for three years. He worked as a Research Assistant for four years in a prestigious advanced knowledge technologies project at the University of Aberdeen. He joined the University of Northampton, as a Software Engineering Lecturer, in 2011. He is currently an Associate Professor and the Deputy Subject Leader overseeing the postgraduate courses in computing. His main research interests include software engineering, pedagogy (assessments and marking), constraints, and knowledge management. He is a fellow of the Higher Education Academy.

KAREN ANTHONY received the Ph.D. degree in neuroscience from King's College London. She undertook postdoctoral research at both King's College London and University College London, where she carried out the pre-clinical and clinical development of RNA-based gene therapies for rare neurodegenerative and neuromuscular diseases. Her work in this area led to the first FDA drug approval for Duchenne muscular dystrophy. She joined the University of Northampton, as a Lecturer, in 2015. She is currently a Professor of molecular medicine and the Co-Lead of the Molecular Bioscience Research Group, University of Northampton. She studied biochemistry at the University of Leeds. She leads a research team investigating the neurobiology of the dystrophin gene and its involvement in cancer. She also works as a consultant advising international pharmaceutical companies on methodology for clinical trial biochemical outcome measures.

MU MU (Member, IEEE) is currently an Associate Professor at the University of Northampton, U.K. He has authored more than 50 peer-reviewed papers published in international conferences and journals. His research interests include human factors in multimedia distribution, intelligent networks, and immersive media. He has been the chair and a program committee member of renowned international conferences. He has been a principal investigator and a lead researcher of various research programs funded by European Commission, U.K. Research Councils, and other funding bodies. He is an Associate Editor of the *Multimedia Systems* (Springer) journal.

MICHAEL OPOKU AGYEMAN (Senior Member, IEEE) received the Ph.D. degree from Glasgow Caledonian University, U.K. He is currently a Professor and a Program Leader of computer systems engineering at the University of Northampton (UoN), U.K. He represents the Research Community of UoN at the University Senate. He is the Postgraduate (PGR) Lead of the Faculty of Arts Science and Technology and Co-Chairs the University's PGR Supervisory Forum. He has over ten years experience in embedded systems engineering. Previously, he was a Research Fellow with Intel Embedded System Research Group, The Chinese University of Hong Kong (CUHK). He is the author of five books, two book chapters, and over 80 publications in major journals and conference proceedings. His main research interests include 3 main strands and disciplines: embedded systems and high-performance computing, such as VLSI SoC design, computer architecture, reconfigurable computing, wired and wireless NoCs, smart rehabilitation solutions, embedded systems and the Internet of Things (IoT); business administration, such as neuromarketing, advertising, and market research; and pedagogy. He is a fellow of the Higher Education Academy (UK). He is a Technical Committee Member of several conferences, such as IEEE ICCSN, IEEE ICBDA, and IEEE ICCT. His work on wireless NoC has attracted two best paper awards in IEEE/IFIP EUC 2016 and Euromicro DSD 2016, respectively. He was a recipient of the 2018 International Changemaker of the year award in the first U.K. Ashoka U Changemaker Campus. He serves as a Reviewer of several conferences and journals, including IEEE Access. He has been a Guest Editor of the *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*. He is a Chartered Engineer (C.Eng.) of the IET and a Chartered Manager (C.Mgr.) of CMI.