Section A-Research paper
ISSN 2063-5346

Gene Expression Analysis for Cancer Prediction: A Review of Machine Learning and Deep Learning Techniques

G. Sumalatha
PH.D scholar,
PPG college of Arts and Science,
Coimbatore, Tamilnadu, India.
Sumasekar29@gmail.com

Dr. N. Muthumani
Principal,
PPG college of Arts and Science
Coimbatore, Tamilnadu, India.
drmuthumani24@gmail.com

***Abstract*— Cancer is one of the main global causes of death. Presently, Gene Expression Microarray (GEM) data has been used to assist an accurate and rapid detection of cancer and its subtypes. In various areas of biological research, the evaluation of gene expressions (GEs) is crucial for acquiring the relevant information. DNA microarray technology offers the ability to retrieve information from the expression levels of thousands of genes in a solitary experiment. Early detection of cancer and its subtypes can be directed in an ideal manner by the collection of relevant genes to increase the diagnostic accuracy. GEM data often generates tens of thousands of genes for each data sample. This results in lower sample size, high dimensionality issues and data complexity for detecting the cancer from GEM data. There is a need for computationally efficient and speedy methods to address these types of problems. So, an advanced Artificial Intelligence (AI) techniques such as Machine Learning (ML) and Deep Learning (DL) algorithms have been developed to deal with these issues. These models have achieved success in several disciplines including image, video, audio, and text processing. Similarly, ML and DL models address the challenges observed in (GEs) analysis for various cancer detection tasks to identify the most suitable biomarkers for the various cancer subtypes. This study provides a comprehensive analysis of the many ML and DL techniques designed to detect cancer and its subtypes by analyzing GEM data. Initially, multiple cancer detection and categorization models developed by numerous researchers using ML and DL algorithms are examined briefly. Then, a comparative research is undertaken to comprehend the shortcomings of these algorithms and to propose a new method for accurately detecting cancer and its subtypes.**

Keywords*— *DNA Microarray Data, High Dimensionality Machine Learning, Deep Learning, Cancer Detection

I. INTRODUCTION

Cancer is connected with aberrant, uncontrolled cell development that can penetrate or migrate to different regions of the body. There are currently hundreds of different cancer types that can be lethal to humans [1]. Cancer is a fatal disease that reduces human lifespan expectancy, therefore early detection of the cancer is essential. The early identification of cancer necessitates a more precise and reliable procedure that provides information about the patient's malignancy and hence enables improved clinical decision-making and treatment [2]. Normal cells are transformed into cancer cells when the genes responsible for cell proliferation and differentiation while undergoing mutation.

Identifying GEs using DNA microarrays is an efficient method for classifying, diagnosing, and predicting cancer. There may be thousands of GEs in GEM data, but just a handful are related with specific cancers [3]. Screening and extracting relevant genes, as well as analysing their impact on a disease, are difficult undertakings. Due to advancements in DNA microarray data and deep sequencing technology [4], the expression level of thousands of genes can be evaluated simultaneously. Microarray experiments give researchers with vast amounts of data, but without the correct tools and procedures, it is impossible to retrieve the critical information and knowledge hidden in this database.

A significant quantity of raw GEM data creates analytical and computational difficulties. The structure or pattern of the microarray data makes the analyst's work difficult. The finest statistical models heavily rely on the total number of potential gene combinations. Consequently, the viability of microarray technologies is contingent upon extensive data mining and diagnostic

techniques. The subject of data mining serves a significant aspect in resolving the dimensionality problem [5, 6].

Feature selection (FS) [7] is a data mining method employed to resolve high dimensionality problems. The FS approach distinguishes between pertinent and irrelevant characteristics and excludes the irrelevant ones. Several FS strategies were presented [8] in order to minimize the MD's dimensionality. Gene selection serves two primary functions [9]: (1) to discover important cancer-associated genes. (2) to identify a small gene set with discriminatory strength in order to generate a higher robust pattern classifier for generalization. However, because of the relatively high dimensionality and small sample size of GEM data [10], FS could not be efficiently used to the discovery of relevant genes.

Many researchers have been inspired to examine the application of ML approaches after identifying cancer from GEM data and then categorising cancer patients as high or low risk. ML approaches have been used to forecast the development and treatment of malignant diseases [11]. The potential of ML techniques is to recognize meaningful patterns in complex datasets. Support Vector Machines (SVMs), Genetic Algorithm (GA), K-Nearest Neighbour (KNN) Random Forest (RF), Navies Bayes (NB) Artificial Neural Networks (ANNs), Decision Tree (DT), Bayesian Networks (BNs) and other ML methods helps to resolve the high dimensionality issues on large micro array data with less susceptible error [12].

These techniques have been frequently used in cancer research to construct predictive models, resulting in effective and precise decision making. ML methods finds difficulty to extract meaningful information from massive databases. Also, a ML methods for cancer prediction from GEM datasets requires separate FSs before it is trained. The Figure 1 depicts the difference between ML and DL model.

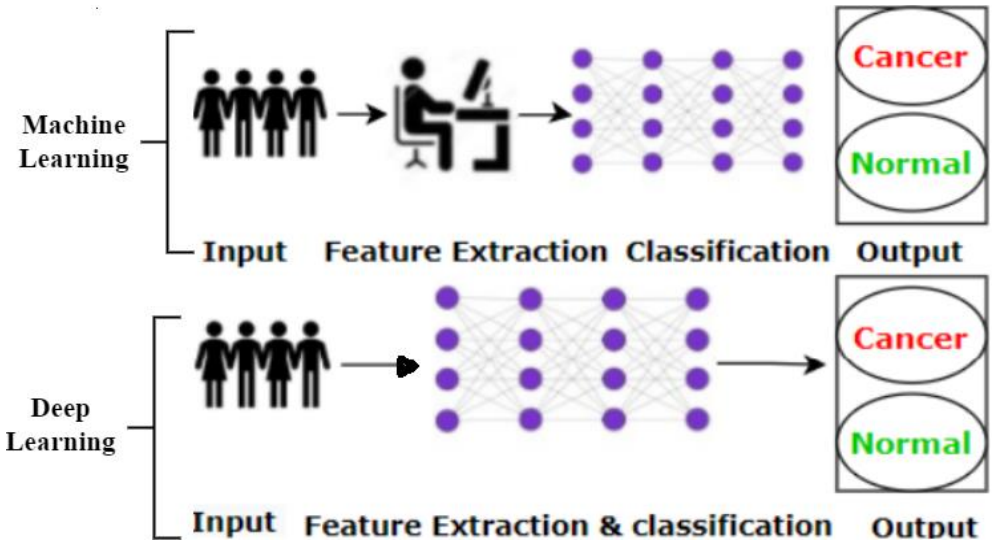

Fig. 1 ML model vs. DL model

In recent years, DL algorithms are applied for various application including computer vision, speech recognition, and natural language processing due to their strong and diplomatic performance. In addition, it is applied to diagnose a variety of other chronic conditions and to enable doctors in determining medical decisions. DL has had a considerable effects on the processing of microarray data [13]. With the new emergence of large datasets, DL methods are employed to accelerate data interpretation and enhance the efficiency of cancer diagnosis, prognosis, and therapy reaction. Using microarray data, there necessitates an immediate demand for reliable and rapid ways to automatically model GEs.

Integrating fastening approaches with rapid binding categories can simultaneously enhances the time duration and rank of identified genes [14]. Moreover, the implementation of an effective DL model can greatly and quickly increase the accuracy of GEs. DL architectures are essentially Artificial Neural Networks with

several nonlinear layers, and various varieties have been constructed based on the characteristics of input data and research target. The figure 2 depicts the structure of DL model. Here, DL structures into four groups [15] (i.e., Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Hybrid Architectures.

Each GEM dataset contains thousands upon thousands of genes. Analysing such a significant amount of GEM data is highly challenging. Moreover, only a minimal number of genes are involved in modulating the GEs levels. These few genes are referred to as characteristic genes. These distinctive genes are associated with unique biological processes of different forms of cancer. Recognizing these genes from vast arrays of GEM is an essential domain of research. This substantial group of genes can improve the accuracy of cancer and its subtypes and gives proper pathway for the early diagnosis.

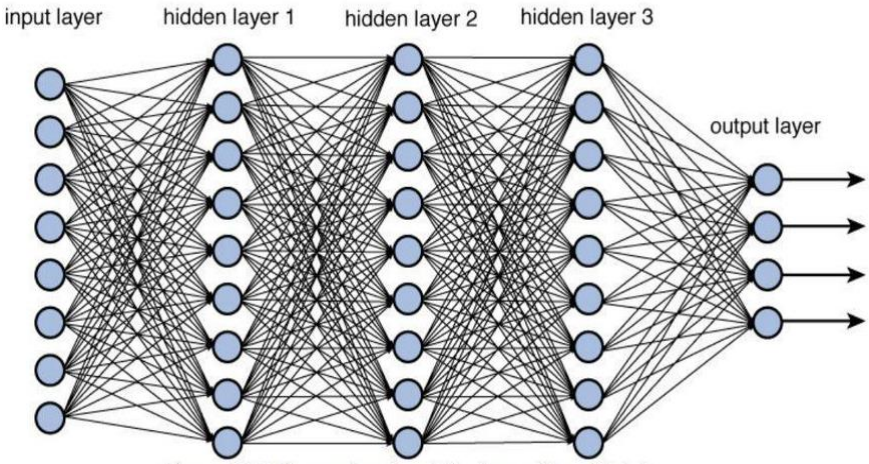

Figure 2 DL architecture

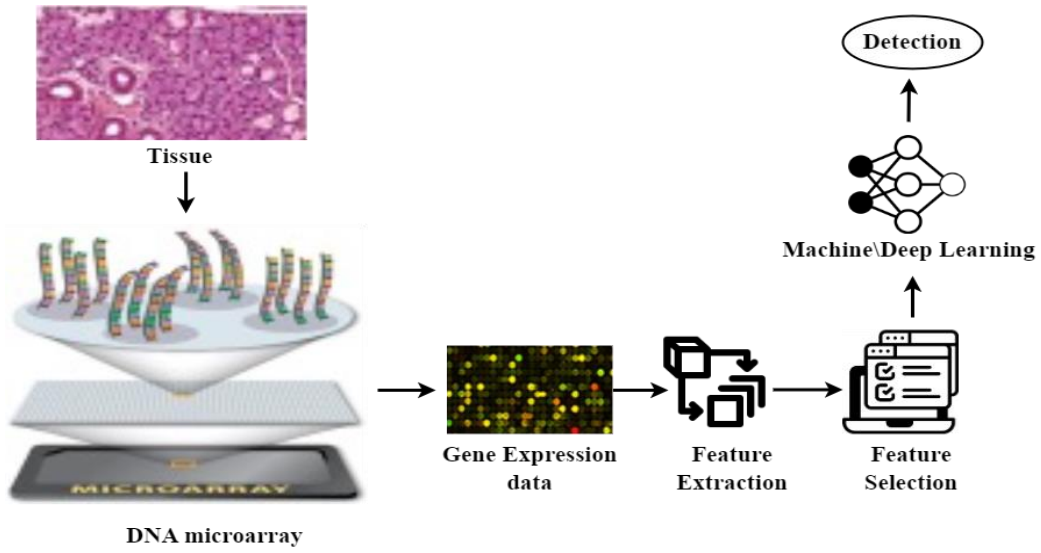

Figure 3 ML\ DL model used for cancer detection from microarray cancer dataset

The figure 3 provides the application of ML\ DL model used for cancer detection from GEM dataset. This paper aims to offer a detailed overview of various ML and DL algorithms and their applications in GEM data-based cancer detection. In addition, a comparative analysis of the merits and drawbacks of these models is provided to indicate their future extent. The remaining sections are formatted as follows:

Section II describes a variety of cancer detection algorithms based on GEM data. In Section III, a comparative study of these models is provided. Section IV provides a summary of the complete study and a preview of its future breadth.

II. SURVEY ON ML AND DL MODELS USED FOR CANCER DETECTION

Yuan et al. [16] developed a modernized classifier called DeepGene based on DL and somatic point mutations for cancer categorization. Initially, clustered filtering technique was utilized to aggregate the gene data based on the probability of mutation instance removing the vast percentage of unnecessary genes. Then, an indexing sparsity elimination was applied to transforms the gene data into indexes of its non-zero elements, thus mitigating the effect of data sparsity. Finally, the data gathered from filtering and sparsity elimination techniques were given as input to a DNN classifier, which retrieves high-level characteristics for reliable categorization of cancer and its sub-categories.

Tumuluru & Ravi, [17] developed a Grasshopper optimization model (GOM)-based deep belief neural (DBN) networks for cancer classification. This model utilizes Logarithmic transformation to pre-process the GEM data for eliminating the complexity associated with the classification. Then, the Bhattacharya distance was used to selects the highly informative genes form the pre-processed module. The selected genes were fed as input to the DBN network to for categorizing the cancer swub categories. The weight upgrade in DBN networks was based on a standardized error estimate utilizing GOM and Gradient Descent, which classifies cancer presence or absence with a lower error rate.

Guo et al. [18] developed a boosting cascade deep forest model called BCDForest for cancer sub-kinds categorization based on GEM data. BCDForest enhances the ensemble heterogeneity and adapting quality of each RF model in Multi-Grained Screening (MGS) to determine additional knowledge about an input data. Initially, a MGS model was designed to train several binary learners to enhance ensemble heterogeneity. Then, a boosting method was employed to identify highest significant features in cascade forests, hence propagating the merits of distinct features throughout cascade stages to increase categorization performance. Furthermore, an out-of-bagging method was used to assess the uncertainty in model fitting and assigns a confidence weight for each forest to modify the result predictions.

Kong and Yu [19] suggested a DNN algorithm employing RF (fDNN) to extract feature representation form GEM data to categorize cancer types. This model included two phases. In the first stage, training data with labels were applied to fit in the forest, and in the subsequent stage, estimations from each tree in the forest for all occurrences are provided for training to the fully-connected DNN. The categorization efficiency of the fDNN model was correlated with both hyper-parameters of the forest and DNN model. These parameter were trained to examine the fDNN performances under same similar settings of specific dataset for cancer prediction.

Ghosh et al. [20] suggested a genetic algorithm (GA) to identify the cancerous gene from GEM data using the ensemble models. In this approach, a two-stage model was constructed for FS. Initially, the ensembling filter methods were established by examining the union and intersection of the top-n characteristics of ReliefF, chi-square, and proportional unpredictability. This ensemble combines all of the information from the three rankings into a unified subset. Then, GA was used to the union and intersection to obtain fine-tuned findings for accurate cancer categorization.

Xu et al. [21] developed a deep flexible neural forest (DFNForest) model for cancer subkinds categorization based on GEM data. A multi-categorization problem was divided into numerous binary categorization problems in each forest by the DFNForest model. Then, the flexible neural tree (FNT) model's complexity was improved without the inclusion of new parameters owing to the optimization of DFNForest's cascade model. Additionally, this methodology offers the integration of fisher ratio with a neighborhood rough set to lessen the GEM data dimensionality and improve the accuracy of cancer classification.

Alrefai [22] presented an ensemble ML method for leukemia cancer diagnosis based on GEM dataset. In this model, an ensemble method was developed by combining four classifiers such as SVM, KNN, NB, and DT to initiate the fitness values used in Particle swarm optimization (PSO) as ideal result to identify all the search space in the limited time while ensuring the highest amount of meaningful genes that leads to improved leukemia cancer identification. Furthermore, once the output was combined using majority vote, the diversity of classifiers was leveraged to increase prediction efficiency.

Lee et al. [23] developed a cancer detector using ensemble model for accurate cancer classification system. The CPEM was based on mutation attributes and an ensemble of ML classifiers such as RF and DNN was generated using different forms of cancer somatic mutations and their consequent attributes as input. In this technique, the implications of diverse input parameters derived from different genetic data like mutation profiles, rates, spectra and signatures, and somatic copy values alterations were explored and employed to accurately detect cancer categories. The effects of input features were investigated and parameters related various cancer types were discovered to provide relatively significant importance in the initial prediction stage.

Khorshed et al. [24] suggested a DL framework called Gene eXpression Network (GeneXNet) system for cancer diagnosis based on whole-transcriptome GEM data. GeneXNet was designed to discover genetic defects that drive cancer growth by accumulating genomic fingerprints across multiple tissue types without the application of gene FS. The transfer learning (TL) approach was utilized to build classifiers for different types of cancers that lacked enough patient samples to be trained individually. Furthermore, this approach visualizes the molecular clusters created by the network's intermediate GEs feature maps which assists in identifying the genomic linkages of GEs determining tumor class prediction.

Khalifa et al. [25] developed an optimized DL approach using binary PSO with DT (BPSO-DT) and CNN to detect different cancer types based on cancer RNA sequence (RNA-Seq) GEM data. Initially, the high-dimensional RNA-seq data was adjusted to minimize its dimensions by determining the optimal features and excluding the unrelated information in order to attain a significant level of categorization performance with BPSO-DT. The improved RNA-seq data were subsequently incorporated into 2D-images. Various data enhancement approaches were applied to these embedded 2D-images in order to eliminate the overfitting issue and train the model to obtain greater precision. Finally, the acquired 2D image data were fed into a deep CNN structure for cancer classification.

Lopez-Garcia et al. [26] developed a TL with CNN for cancer survival prediction using GEM data. This model integrates the efficiency of the CNN method to retrieve high-level characteristics from structured input information with the efficiency of the TL method to overcome overfitting difficulties on small training datasets comprised of high-dimensional representative, which were employed for cancer prediction tasks with GEM data. In addition, this approach intended to reassemble GEM data by converting linear expression vectors into two-dimensional images of GEs from which a CNN might leverage local patterns to enhance lung-cancer survival forecasting efficiency.

Zhong et al. [27] developed a Cascade Flexible Neural Forest (CFNForest) for cancer subtypes classification on GEM data. CFNForest was built on a FNT model that utilized a bagging ensemble method to construct the model's topology and parameters autonomously during training. In addition, CFNForest improves the functional performance and reliability of the algorithm by employing a sample selection process among layers and adjusting the final weights for each layer. The categorization of cancer subforms using FNT Group Forest with multiple feature sets was appropriate for analysing limited sample level datasets.

Fathi et al. [28] developed a hybrid cancer classification and diagnosis approach called PCC-DTCV using different ML models. PCC-DTCV maximizes the adjusting variable (max-depth) of the DT classifier using Grid Search cross-validation (CV) and Pearson's correlation coefficient (PCC) for gene (feature) selection. High categorization scores were obtained using this model, which was utilized to determine an ideal or nearly ideal subset of instructive and significant genes. It assists in discovering the highest relevant genes to enhance the categorization of GEM data. The finding also indicates the model's efficiency in recognizing the majority relevant genes to minimize the complexity of GEM data.

Bae et al. [29] utilized a ML algorithms like K-Means Clustering (KMC) and the Modified Harmony Search Algorithm for selecting the features in for colon cancer detection. The actual data were initially Z-normalized using a data pre-processing approach. The Reynolds score was then used to select potential genes for normal and abnormal distribution. After candidate genes were grouped using KMC, a representative gene was chosen from each cluster. Finally, the updated harmony search method would be used for FS. In addition, the gene combination was produced using the FS approach, which was used to the classification model and validated using 5-fold CV.

Mahfouz et al. [30] introduced an ensemble of KNN (EKNN) based decision model for selecting the features in GEM data for cancer detection.

Along with the conventional KNN, four decision models were used in this technique like Density (KDN), Average linkage (KLN-Avg), Single linkage (KLN-Min), and complete linkage (KLN-Max). A KNN-table with a sequence of KNN for each sample was maintained throughout the training phase and statistics were derived for each class based on it. Using both the computed statistics and the data from the KNN-table, the resulting model's decisions were generated during the validation phase. According to the Bhattacharyya computation, the predicted probabilities of the complete and mean relation among each class were much more dispersed out than those of other similar approaches.

El Kafrawy et al. [31] created the composite FS method SVM-mRMRe which integrates SVM with minimum redundancy-maximum relevancy (mRMR) for the cancer detection. The relevant features (genes) from the gathered dataset were initially identified using the SVM. In the following step, SVM_ (Radial Basis Function) RBF _CV was coupled with the gene subset result. A voting mechanism was used to discover important genes with greater relevance and smaller redundancy. The next step was an ensemble mRMRe selection of non-redundant and relevant genes to the biological context providing more in-depth biological interpretations. Finally, other features were biologically calculated and the results were utilized to classify cancer.

Rukhsar et al. [32] presented a DL method for evaluating the RNA-Seq GEM data for cancer and its subtype categorization. A dataset for various cancer kinds was initially retrieved as a numerical sequence. Then, RNA-Seq results were transformed into 2D images using normalization and zero padding. Then, pertinent characteristics were retrieved and selected using DL. Finally, a variety of DL models including CNN, ResNet50, ResNet101, ResNet152, VGG16, VGG19, AlexNet, and GoogleNet, were deployed to the RNA-Seq dataset to produce numerous cancer categorization outputs.

Shen et al. [33] constructed a DL model called DCGN using high-dimensional GEM data for cancer subtypes classification. This model incorporates CNN with bidirectional gated recurrent unit (BiGRU) to perform nonlinear dimensionality reduction and trains patterns to eliminate unnecessary variables from GEM data. The BiGRU analyses deep features and retains their essential data whereas the CNN handles high-dimensional data and extracts significant local features. Data equalization was originally accomplished by DCGN using the artificial minority oversampling approach. Finally, it obtains the relevant features by integrating both neural networks to address the difficulties of limited sample numbers and sparse, high-dimensional features for better cancer subtype categorization.

Rezaee et al. [34] presented a DL based microarray cancer classification and ensemble gene selection method. The initial pre-processing of the gathered data included gene normalization which lowers computational costs and optimizes GEs. Then, ensemble soft voting was employed for the FS procedure. The weights of the genes were then assessed for gene selection, and a DL stacked auto-encoder method was employed to classify cancer. Finally, the anticipated label produced a classification of cancer types.

Zan et al. [35] developed a DL method named DeepFlu for predicting symptomatic influenza infection using pre-exposure GEs. Initial, human peripheral blood GEM datasets were collected. The leave-one-person-out CV was then performed to the fundamental multilayer perceptron dataset for system computation. Finally, deepflu structure which comprised of a feed-forward neural network assists to identify linearly non-separable characteristics to forecast the existence of flu symptoms.

Kanwal et al. [36] created a heterogeneous DL framework infused with artificial algae algorithm (AAA) to enhance the prognostic prediction for cancer. The AAA approach was utilized to identify the highest significant features from the dataset and construct a resilient model to exclude the noisy data. A combination of earlier and delayed fusion procedures was applied to enhance the ability for prediction. The AAA model was then augmented with Double DEEP Q-NETWORK (DDQN), Convolution eXtreme Gradient Boosting (CNN-XGBOOST), and Convolution Support Vector Machine (CNN-SVM) models for cancer prognostication.

III. COMPARTIVE ANALYSIS

The previously examined AI models for cancer diagnosis and classification are compared in this part, and their advantages and disadvantages are shown in Table 1.

Table.1 Evaluation on various methods for cancer diagnosis and classification

Ref No.	**Algorithms**	**Merits**	**Demerits**	**Performance**
[16]	DNN, indexing sparsity elimination and clustered filtering technique	Even in large dataset the features space characters were stable and robust	Deepgene was tested only using somatic point mutations with known cancer types, unknown cancer type information were not evidently provided in the experiment	The maximum accuracy obtained by deep-gene was 63.9% on TCGA dataset
[17]	GOM and DBN	Because of its optimized performance, it was able to address the network complexity in a short amount of time.	The developed model can only identify only certain number of gene subsets.	The accuracy of GOA-DBN on colon cancer data and leukemia dataset was 95.34% and 94.59% respectively.
[18]	BCDForest, Boosting method MGS model and Out-of-bagging method	Lower computational time and less informative loss was resulted	On larger dataset, the efficiency of this model was less	This algorithm executes 96.4% on brain dataset; 91.6% on colon dataset and 92.8% on adenocarcinoma dataset
[19]	fDNN model	This model effectively mitigated the overfitting problems and provides less classification error rate	Regularization in parameter influences the accuracy of this two-stage approach with high dimensionality issues	This approach achieves 98.6% of Area Under Curve (AUC) score on GSE99095 dataset and 77.8% of AUC score on GSE106291 dataset
[20]	GA and Ensembling Filter Methods	Efficient performance on larger class featured datasets	High computational complexity	This model achieves 100% of accuracy on colon cancer dataset ; 96.07% on Lung cancer dataset; 100 on Leukemia cancer dataset; 98.03% on prostate dataset and 100% on SRBCT dataset
[21]	DFN Forest and FNT	DFNForest was well adjusted for analysing small-scale biology data Because the range of levels in may be adjusted adaptively	This model have high computational cost issue	This model achieves 93.6% of accuracy on BReast Invasive CArcinoma (BRCA) dataset; 84.2% of accuracy on Glio-Blastoma Multiforme (GBM) dataset and 88% of accuracy on lung dataset
[22]	Ensemble ML method and PSO algorithm	Better classification accuracy	Even after a certain number of repetitions, the computing gene selection model's convergence speed was lower	This model achieves 100% of classification on Leukemia dataset
[23]	Ensemble model RF	This model was suitable to	This model requires large	This algorithm obtains a

	and DNN	biological significance of neural type.	amount of datasets for the efficient performance.	mean categorization efficiency of 84% for 31 cancer categories from the TCGA database.
[24]	GeneXNet and TL model	This system was good in genetic features classification without utilizing any discrete input features	This system lacks to provide better result on performing with larger dataset	CFNForest obtained categorization efficiency of 90.9% and 94.4%, respectively on the Lung and the BRCA dataset,
[25]	BPSO-DT and CNN and Data Enhancement approaches	It was easier to understand the correlation among samples since they were more defined.	The datasets was imbalanced the efforts were not given to resolve this issue in the dataset	This model yields 98.9% categorization accuracy on human samples comprising 33 distinct cancer types across 26 organ locations,
[26]	TL – CNN model	There was no repeat in the distribution of the characteristics throughout the various clusterings since they are dispersed fast.	Slow convergence rate was considered limitations	The obtained accuracy attains 98.30% on BRCA, 98.20% on kidney renal clear well carcinoma (KIRC), 97.7 % on Lung Squamous Cell Carcinoma (LUSC) and 96.4% on Uterine Corpus Endometrial Carcinoma (UCEC).
[27]	CFNForest and FNT model	Robustness and flexibility was high for the cancer survival prediction	This method necessitated a high range of records for effective sampling	This approach results 72.69% of accuracy and 73.88% of sensitivity on pan cancer dataset
[28]	PCC-DTCV and Bhattacharyya computation model	This strategy was beneficial in discovering an adequate or near-optimal selection of relevant and significant genes and it generated good categorization performances.	This model results with high dimensionality issues	This model attains 94% of accuracy; 88% of AUC, , 97% of sensitivity, and 79% of specificity with DT classifier on Gordon dataset for lung cancer
[29]	KMC and the Modified HSA	The computational time was reduced in the final subset.	Less performance on smaller datasets	This model obtains the classification accuracy of 93.46% on colon cancer dataset
[30]	EKNN	It takes less classification time to correctly classify the sample datas	It was not well performed on larger datasets	This model achieves 93% on Leukemia dataset; 98.16% on prostate dataset and 78.11% on CNS dataset
[31]	SVM-mRMRe and SVM_RBF _CV	The high dimensionality issues were highly	It had a less number of gene subset comparatively	This model achieves 1.00±0.00% on breast

		resolved		cancer, lung cancer and brain cancer dataset
[32]	DL methods	This method efficiently extracts high-level features for classification	This method computes more time to train the data	This method results with 97% of total accuracy on RNA-Seq data
[33]	DCGN, BiGRU and CNN	Limited sample sizes and sparse, high-dimensionality issues are effectively overcome using DCGN	It struggled to discover an adequate search space for excellent classification precision due to a lack of prior knowledge of datasets.	DCGN achieves 98.6% and 99.3% of accuracy on BRCA dataset Bladder Urothelial Carcinoma (BLCA) dataset respectively.
[34]	DL stacked auto-encoder method and Ensemble soft voting	This model effectively eliminates the high dimensionality issues	Even on evaluating smaller type of dataset, this model have high time computation	The testing were performed using three data sets, this model achieves 97.51% accuracy on DiffuseLarge B Cell Lymphomas. 99.36% accuracy in leukemia and 96.34% in prostate cancer
[35]	Deep flu architecture and FFNN	Consistent result was achieved in selecting the expression patterns to identify the Flu infection	Due to the scarcity of influenza GEM data, DL finds difficulties in training the model.	DeepFlu scored the best overall performance using the 22,277 H1N1 and H3N2 characteristics collected. DeepFlu had 70% accuracy, 0.787 AUROC, and 0.758 AUPR for H1N1. It achieved 73.8% accuracy, 0.849 AUROC, and 0.901 AUPR for H3N2.
[36]	AAA, DDQN, CNN-XGBOOST, and CNN-SVM	Good prediction accuracy and significantly lesser training time	Utilizing more number of learning features might degrade the performance results	This model achieves optimum accuracy of 99% for the Brain dataset; 91% for Prostate Cancer dataset and 95% for Metabric dataset

From the above table, the article [16-36] is studied and it is concluded that the article [36] yields better detection result on cancer and its sub-types based on GEM dataset. In the article [35], AAA was utilized to identify beneficial features from diverse data modalities, and findings were further infused by using DDQN, CNN-XGBOOST, and CNN-SVM based algorithms for cancer prognostic detection. It improves cancer prognostic mortality evaluation and prediction by merging multidimensional features through early and late fusion procedures by utilizing DL and Reinforcement Learning (RL) techniques. Furthermore, the topology of this model dramatically reduces training time on larger datasets while still generating a reliable framework for cancer categorization on GEM datasets.

IV. CONCLUSION

In this paper, a comprehensive evaluation of ML and DL algorithms employing GEM data across major cancer types like lung, breast, CNS etc., are studied for cancer classification. The discussion focused on different ML and DL structures and their benefits in diagnosing different forms of cancer using GEM data. Many constraints like low sample size, high dimensional and class

imbalanced data, processing power and time etc., were stated. Also, it is observed that DL approaches are overcoming the issues of traditional ML techniques in analysing GEM data for cancer. The discussed challenges and performances are key to develop fully functional models that could help in improving in cancer for prognosis and diagnosis and provides ultimately personalized treatments for cancer patients.

REFERENCES

[1] A. S. Nath, A. Pal, S. Mukhopadhyay and K. C. Mondal, "survey on cancer prediction and detection with data analysis," Innovations Syst Softw Eng, vol. 16, no. 3, pp. 231-243, 2020.

[2] H. Q. Wang, G. J. Jing and C. Zheng, "Biology-constrained gene expression discretization for cancer classification," Neurocomputing, vol. 145, pp. 30-36. 2014.

[3] M. F. Ochs and A. K. Godwin, "Microarrays in cancer: research and applications," BioTechniques, vol. 34, no. S3, S4-S15, 2003.

[4] L. Liu, A. Y. L. So and J. B. Fan, "Analysis of cancer genomes through microarrays and next-generation sequencing," Transl. Cancer Res., vol. 4, no. 3, pp. 212-218, 2015.

[5] Y. Peng, Z. Wu and J. Jiang, "A novel feature selection approach for biomedical data classification," JBI, vol. 43, no. 1, pp. 15-23., 2010.

[6] E. Elsebakhi, O. Asparouhov and R. Al-Ali, "Novel incremental ranking framework for biomedical data analytics and dimensionality reduction: Big data challenges and opportunities," J. comput. sci. syst. biol., vol. 8, no. 4, pp. 203, 2015.

[7] M. Kumar, N. K. Rath, A. Swain and S. K. Rath, "Feature selection and classification of microarray data using MapReduce based ANOVA and K-nearest neighbor," Procedia Comput. Sci., vol. 54, pp. 301-310, 2015.

[8] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J. M. Benítez and F. Herrera, "A review of microarray datasets and applied feature selection methods," Information sciences, vol. 282, pp. 111-135.

[9] F. Yang and K. Z. Mao, "Improving robustness of gene ranking by multi-criterion combination with novel gene importance transformation," Int J. Data Min Bioinform, vol. 7, no. 1, pp. 22-37, 2013.

[10] F. Yang and K. Z. Mao, "Robust FS for microarray data based on multicriterion fusion," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 8, no. 4, pp. 1080-1092, 2011.

[11] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput. Struct. Biotechnol. J., vol. 13, pp. 8-17, 2015.

[12] S. Osama, H. Shaban and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review," Expert Syst. Appl. pp. 118946, 2022.

[13] S. Gupta, M. K. Gupta, M. Shabaz and A. Sharma, "Deep learning techniques for cancer classification using microarray gene expression data," Front. Physiol. vol. 13, pp. 1-13, 2022.

[14] B. Hanczar, V. Bourgeais and F. Zehraoui, "Assessment of deep learning and transfer learning for cancer prediction based on gene expression data,' BMC Bioinform. BMC Bioinform., vol. 23, no. 1, pp. 1-23, 2022.

[15] M. F. Mridha, M. Hamid, M. M. Monowar, A. J. Keya, A. Q. Ohi, M. Islam and J. M. Kim, "A comprehensive survey on deep-learning-based breast cancer diagnosis," Cancers, vol. 13, no. 23, pp. 6116, 2021.

[16] Y. Yuan, Y. Shi, C. Li, J. Kim, W. Cai, Z. Han and D. D. Feng, "DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations," BMC Bioinform., vol. 17, no. 17, pp. 243-256, 2016.

[17] P. Tumuluru and B. Ravi, "GOA-based DBN: Grasshopper optimization algorithm-based deep belief neural networks for cancer classification", Int. J. Appl. Eng, vol. 12, no. 24, pp. 14218-14231, 2017.

[18] Y. Guo, S. Liu, Z. Li, and X. Shang, "BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data," BMC Bioinform., vol. 19, no. 5, pp. 1-13. 2018.

[19] Y. Kong and T. Yu, "A deep neural network model using random forest to extract feature representation for gene expression data classification," Scientific reports, vol. 8, no. 1, pp. 1-9, 2018.

[20] M. Ghosh, S. Adhikary, K. K. Ghosh, A. Sardar, S. Begum and R. Sarkar, "Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods," Med Biol Eng Comput, vol. 57, no. 1, pp. 159-176, 2019.

[21] 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.

[22] N. Alrefai, "Ensemble Machine Learning for Leukemia Cancer Diagnosis based on Microarray Datasets," International Journal of Applied Engineering Research, vol. 14, no. 21, pp. 4077-4084, 2019.

[23] K. Lee, H. O. Jeong, S. Lee and W. K. Jeong, "CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network," Scientific reports, vol. 9, no. 1, pp. 1-9, 2019.

[24] 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.

[25] N. E. M. Khalifa, M. H. N. Taha, D. Ezzat Ali, A. Slowik and A. E. Hassanien, "Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach," in IEEE Access, vol. 8, pp. 22874-22883, 2020.

[26] 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, pp. e0230536, 2020.

[27] L. Zhong, Q. Meng and Y. Chen, "A Cascade Flexible Neural Forest Model for Cancer Subtypes Classification on Gene Expression Data," Comput. Intell. Neurosci., vol. 2021, pp. 1-12, 2021.

[28] Fathi, H., 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, pp. 1-14, 2021.

[29] J. H. Bae, M. Kim, J. S. Lim and Z. W. Geem, Feature selection for colon cancer detection using k-means clustering and modified harmony search algorithm. Mathematics, vol. 9, no. 5, pp. 570, 2021.

[30] M. A. Mahfouz, A. Shoukry and M. A. Ismail, "Eknn: Ensemble classifier incorporating connectivity and density into knn with application to cancer diagnosis," Artif Intell Med., vol. 111, pp. 1-24, 2021.
[31] 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.
[32] L. Rukhsar, W. H. Bangyal, M. S. Ali Khan, A. A. Ag Ibrahim, K. Nisar and D. B. Rawat, (2022). "Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification," Appl. Sci., vol. 12, no. 4, pp. 1-17, 2022.
[33] J. Shen, J. Shi, J. Luo, H. Zhai, X. Liu, Z. Wu, …., and H. Luo, "Deep learning approach for cancer subtype classification using high-dimensional gene expression data," BMC Bioinform., vol. 23, no. 1, pp. 1-17, 2022.
[34] K. Rezaee, G. Jeon, M. R. Khosravi, H. H. Attar and A. Sabzevari, "Deep learning-based microarray cancer classification and ensemble gene selection approach," IET Systems Biology, vol. 16, no. 3-4, pp. 120-131, 2022
[35] A. Zan, Z. R. Xie, Y. C. Hsu, Y. Chen, T. H., Lin, Y. S. Chang and K, Y. Chang, "DeepFlu: a deep learning approach for forecasting symptomatic influenza A infection based on pre-exposure gene expression," Comput Methods Programs Biomed., vol. 213, pp. 1-8, 2022.
[36] S. Kanwal, F. Khan and S. Alamri, "A multimodal deep learning infused with artificial algae algorithm–An architecture of advanced E-health system for cancer prognosis prediction, " J. King Saud Univ. - Comput. Inf. Sci., vol. 31, no. 6, pp. 2707-2719, 2022.