2021 6th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 02-04, 2021

Deep Learning Techniques in Cancer Prediction Using Genomic Profiles

Swati B. Bhonde
Research Scholar,
Smt. Kashibai Navale College of Engineering,
Pune, India
swati.bhonde@gmail.com

Jayashree R. Prasad
Professor,
Sinhgad College of Engineering,
Vadgaon, Pune, India
jrprasad.scoe@sinhgad.edu

***Abstract*— According to World Health Organization (WHO), cancer is the second major reason of death followed by cardiovascular diseases. Early diagnosis of cancer is essential to offer correct treatment to the patient. Traditional histology based examination is used as the gold standard for cancer identification but the accuracy of the result generated is questionable and it may vary amongst pathologists. Next-Generation Sequencing (NGS) technologies have produced a huge amount of cancer genomic data publicly available which elevates an idea of the identification of candidate genes contributing to uncontrolled cell growth resulting in cancer. Analyzing gene expression data is crucial to find out these harmful mutations and to avoid further consequences. Therefore there is a need for learning methods to explore this data for discovery of target genes, accurate disease identification and drug discovery. Despite research in cancer detection using genome sequencing, there exists a need to improve accuracy and find out driver mutations in genes. This paper presents a systematic literature survey of in-depth learning techniques used to predict cancer using genome sequencing. The system can be used by medical practitioners to have timely diagnosis & prognosis of diseases like cancer.**

Keywords*— *Cancer prediction, deep learning, machine learning, precision medicine, genome sequencing

I. INTRODUCTION

In the digital world, the availability of medical data has tremendously increased with higher volume and variety. Genome sequencing and personalized medicine have changed the way to treat a disease like cancer. In a traditional method, pathologists need to manually identify changes in genes by taking a reference of clinical literature which is highly time consuming[1]. By knowing the genetic structure of an individual, precise medical treatment can be offered. This is vital for lessening the unwanted side-effects of medicine resulting from the traditional one-size-fits-all approach. It will also reduce the cost of treatment by offering effective treatment plans based on an individual's risk score & characteristics.

One of the features of cancer is uncontrolled cell growth resulting into tumor. Finding cancerous cell variations from genome data is crucial to predict cancer type. The genomic profile contains a massive set of multidimensional data that should be analyzed with an appropriate statistical method to extract meaningful information. Research is also on-going on the classification of cancer patients into a high grade or low grade, whether cancer is progressing or deteriorating and to predict drug response.[2] Challenges are there in the transformation and representation of this genome sequence data into a machine-readable format. Learning techniques are used to learn patterns from this data which can be further used for analysis, interpretation, and decision making. Therefore, there is a need for a system that enhances accuracy in cancer prediction and provides personalized medicine with no side effects in an individual. Thus, resulting in substantial lifesaving of the patient. [3]

With the advent of the omics and non-omics data and the integration of holistic healthcare records, Deep Learning Algorithms are nowadays widely used in cancer prognosis and diagnosis. In this review paper, we have reviewed existing work related to deep learning used in building predictive models for different cancer types. The review paper is divided in six main sections. Section-II provides the theoretical background of cancer and genomic profiles while overviewing recent development in health informatics related cases on cancer prediction Section- III lays out a generalized health analytic system deployed for cancer detection and prediction. Section- IV illustrates use of Deep Learning techniques specifically for cancer prognosis & Section- V outlines key challenges in applying deep learning algorithms for cancer prediction. Finally, Section- VI gives insights on potential limitations, practical implications, and future directions for researchers interested in healthcare analytic and genomic approaches to cancer prediction.

II. THE GENETICS OF CANCER

Genes are found in the DNA (DeoxyriboNucleic Acid) of each cell in a human body. They control functions of a cell like cell growth, division and lifetime. The DNA sequence forms a gene and if any alteration is made in the precise DNA sequence, then it incurs genetic mutation. A small number of mutations may or may not be harmful but if found largely then it can lead to cancer.[4]. Multiple mutations are single nucleotide variants (SNVs), structural variations and insertions or deletions (Indels). The mutation of genes in a cell causes cancer. These driver mutations available in cancer tumors are the root cause of rapid growth and spread of cancer cells in the human body. Currently, the task of differentiation is carried out manually, which is tremendously time consuming and produces less accurate results. In the traditional approach, these genetic mutations are reviewed and classified by pathologists by concerning to historical clinical literature. Cancer is a ghastly disease and very often it does not announce itself in an early stage. With the advanced Next Generation Sequencing (NGS) technologies, a huge volume of cancer genome data is generated. Decoding of such an enormous volume of data leads to many opportunities and challenges to clinicians and biologists for understanding cancer initiation, progression, and development.

The genomic profile of an individual is used to investigate the occurrence of certain diseases in an individual. It identifies the genetic characteristics of a person

2021 6th International Conference for Convergence in Technology (I2CT) | 978-1-7281-8876-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I2CT51068.2021.9417985

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.

and tries to discover the response of a person to a specific drug. Genomic profiling can be used by doctors to diagnose a patient with high-risk genetic variation causing a particular disease. Advances in processing genome data using High Throughput Sequencing techniques have replaced conventional methods used in genomics. It is now possible to investigate these genomic profiles using computational approaches. Even biomarkers causing the particular disease to emerge can be identified. They can be further used to predict and evaluate the treatment responses and proves as an indicator of any kind of disease. Though radical research is on-going in the field of oncology, there is not a single test that can accurately diagnose cancer. Also, the same drug is not effective across people suffering from the same type of cancer type. So the current trial and error based drug treatment system incurs a lot of expenditure that demotivates the patient. To avoid this, precision medicine can be used in which personalized treatment can be offered to the patient based on the genetic characteristics of an individual to know the onset of the disease. Thus it will help mankind from an adverse drug effect and thus resulting in significant lifesaving and even cost-saving in treatment.

III. GENERALIZED CANCER PREDICTION SYSTEM

Knowledge of changes in genes i.e. mutations plays an important role in predicting cancer. There are two types of mutations, first is acquired mutations caused by damage of DNA in a cell contributed by various factors like lifestyle, radon, UltraViolet (UV) radiation, genetic defects, gender and age, etc. Second is germline variation which is hereditary and occurs in spermor egg. The following figure shows the general system architecture used for cancer prognosis using Deep Learning Algorithms.

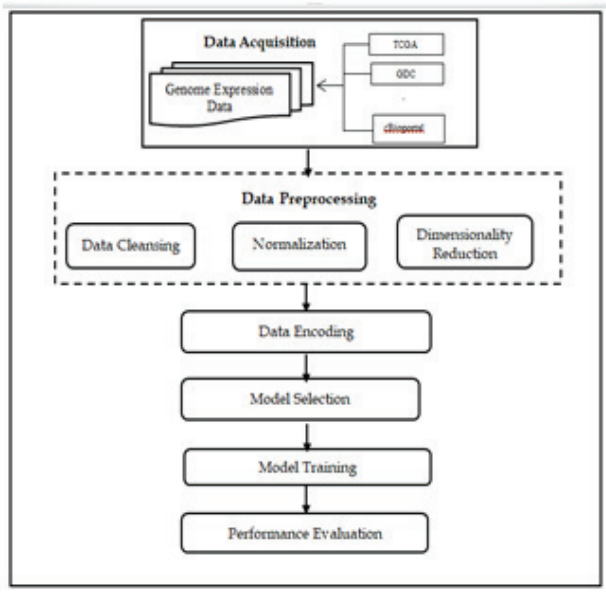

Fig. 1. General steps in handling genome data using deep learning

A. Data acquisition

Massive amount of genomic data is available publicly in databases like TCGA, NCBI, UCI, GeneBank, etc. These datasets cannot be used directly for deep learning algorithms as they are stored in fasta, fastq, gff2 format, which must be converted in vector or matrix format. Proper preprocessing techniques used in the initial stages of the model can significantly increase accuracy and improve the speed of a Deep Learning Algorithms. This phase comprises of following three subphases:

1) Data cleaning: This aims at missing value imputation, handling outliers, noise removal and removing duplicate values. It is one of the important steps contributing to a generation of accurate results because if data is not cleaned properly then it may increase the cost of training a model.

2) Data normalization: In normalization, values in the numeric column are scaled up to a certain scale to avoid local optima.

3) Dimensionality reduction: To remove unwanted features for proper training, dimensionality reduction techniques like auto-encoder and principal component analysis can be used.

B. Data encoding

Generally genomic data is stored as three types of sequences such as – DNA, RNA and amino acid sequences. To encode this data in machine-interpretable format, methods like one-hot encoding, position-specific scoring matrix, point accepted mutation, blocks substitution matrix, etc. are used. The output of this step is a normalized numeric vector or matrix, which can be further fed as input to Deep Learning Algorithms.

C. Model selection

So far, we don't have any model that performs well for all types of problems. Instead of using a single model one may integrate features of multiple models to get better results. Wnuk et.al. [17] got more precision after combining CNN with LSTM. Also, the use of new technologies like parallel programming can be induced to achieve less computation time and more accurate results.

D. Model training

Training of Deep Learning Algorithms is a challenging task to train such massive data, ample time is needed. Therefore, powerful computational resources like GPU and the correct programming paradigm must be chosen. Also, the training, validation and testing dataset must be chosen properly.

An appropriate selection of hyperparameters can have a great impact on deep architecture as they control the behaviour of the training phase. In this view optimization parameters or model, specific hyperparameters can be used to adjust the performance of the model. Grid search and random search approach can be used to evaluate the algorithm for each combination of hyperparameter. While training a model, we also need to set other parameters like accurate learning rate, dropout rate, weight initialization, batch size, number of epochs, hidden layers, hidden unit and activation function, etc. The effect of the overfitting and underfitting can be avoided if the model is well-tuned.

E. Performance evaluation

Generally k-fold cross- validation method is used to examine the accuracy of the model. Some other parameters like precision, accuracy, recall, f-measure, etc. are also used to check the performance of the model.

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.

IV. USE OF DEEP LEARNING TECHNIQUES IN CANCER PROGNOSIS

This section presents a systematic literature survey of the work carried out by the researcher worldwide to diagnose multiple types of cancers. We have used the publish and perish tool[4] to fetch all the papers using the keywords "Deep Learning Algorithms" + "Genomic profiles/gene expressions" + "Cancer prediction/prognosis". Around 1000 papers were retrieved covering broad topics related to this domain. Out of which 253 papers were identified reflecting the use of only Deep Learning techniques in cancer prognosis from the year 2016 to 2020. We have exclusively considered those sub-domains where Deep Learning Algorithms were used to predict clinical outcomes. We came across research publications related to molecular analysis of DNA(142 papers), introductory papers exploring applications of deep learning in cancer diagnosis(37 papers), tumor image analysis to determine progression or deterioration of cancer(11 papers), drug discovery, drug response - patient survival prediction((25 papers) and in many of the work an integrated system was developed to predict multiple cancer types(38 papers) by analyzing gene sequence. Then we sorted these papers based on citation count to retrieve the most relevant papers. In figure 2, the horizontal axis represents a spectrum of topics covered and vertical axis represents the citation count of papers.

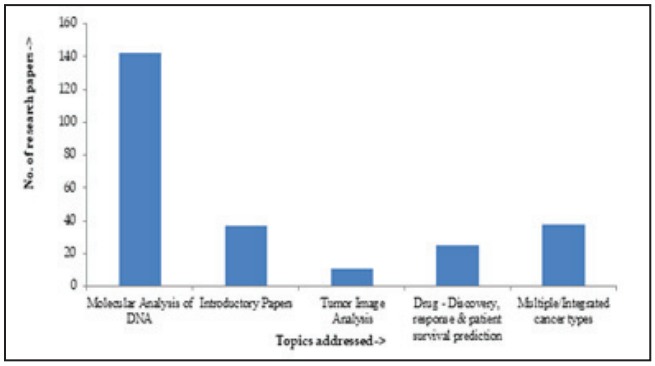

Fig. 2. Different topics covered by researchers

The table1, below shows the survey carried out to predict different cancer types using genome sequencing. Here, we have selected the top-30 papers based on their citation count.

TABLE I. LITERATURE SURVEY

Sr.No	Ref. no	Cancer types predicted	Method	No. of samples	Type of data	Accuracy	Validation Method	Important features	Limitation
1	[5]	BLCA, RCA,COAD, GBM, KIRC, LGG, LUSC, OV, PRAD, SKCM, THCA, UCEC	Genome Deep Learning - DNN - Decay method - L2 regularization method - Sliding average model	6083 - WES' tumor samples & 1991 - healthy samples	TCGA - Tumor samples, IGSR - Healthy samples	Accuracy- 94.70%, Sensitivity - 97.30%, Specificity - 85.54%	80% training & 20% testing	Identifies cancer risk using gene expression before it is diagnosed	Non-Omics data can be integrated & need to have generalized model to predict more cancer types.
2	[6]	BRCA, OV, UCEC	Bayesian optimization method, Deep Neural Networks, Risk back propagation	Transcripti onal feature set(RNA) - 17,000 Integrated - 300 to 400 features	TCGA Integrator	Median c-index ranging from 0.664 to 0.599	60% training 20% validation & 20% testing	It works on small datasets & uses transfer learning Approach to improve prognostic accuracy.	Larger dataset is needed to avoid overfitting and for non-linear models risk back propagation analysis should be done to improve for non-linear models
3	[7]	ACC, BLCA, BRCA, CESC,HNSC, KIRP, LGG, LUAD, PRAD, PAAD, STAD & UCS	Deep-Gene, an advanced deep neural network (DNN) based classifier	22,834 genes from the 3122 samples	TCGA	Optimal accuracy of 60.1%	10-fold cross validation method	This model generated more accurate cancer classification based on somatic point mutation.	This model can be extended for complex genotype- phenotype, for large scale data
4	[8]	BRCA,GBM, COAD, LHSC & LUNG	Boosting Cascade Deep Forest	3594 Samples	TCGA Pan cancer	Accuracy- 80.6%	5-fold cross validation method	It disseminates the benefits of selective features to remaining layers to improve the classification performance.	System lacks to handle imbalanced and high-dimensional small-scale data and improve stability of algorithm.
5	[9]	LUAD, STAD & BRCA	Stacked sparse auto-encoder (SSAE) based classification using RNA- seq data	21,000 genes	TCGA	Accuracy- 97.03%	5-fold cross validation method	This model processes high- dimensional data with good accuracy.	Algorithm suffers from imbalanced calculation accuracy, stability and computational cost is more.

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.

6	[10]	Detection of new gene function	Stacked Denoising Autoencoder Multi-Label Learning	1144 genes	KEGG Pathway & PubMed Central	Average Precision is 0.577	10-fold cross validation method	Algorithm discussed here provides dimension reduction and multi-label classification	Annotation performance can be improved by integrating more pathway information.
7	[11]	ACC,BRCA, CESC,DLBC ,KICH,KIRC ,KIRP,OV,P CPG,PRAD, SARC,SKC M,TGCT,TH CA,THYM, UCEC	Dedicated CNN with fully connected layers	60383 genes	TCGA Pan cancer	C-index of 0.78	80 Epochs	Model induced here trains feature encodings and Predict single cancer, overall survival and capable of handling multimodal data.	There is a need to use deeper architecture and advanced data expansion is needed to improvise accuracy.
8	[12]	BRCA, COAD & LUAD	deepDriver - CNN	2082 genes	NCI Genomic Data Commons (GDC)	AUC = 0.984 for BRCA and AUC= 0.976 for colorectal cancer	10-fold cross validation method	Algorithm learns information hidden within mutation data and similarity networks concurrently.	Accuracy can be improved by extracting predictive feature
9	[13]	GBM	Pathway-Associated Sparse Deep Neural Network (PASNet)	522 samples & 12,044 genes	TCGA	AUC-0.66, F1-score - 0.39	5-fold cross validation	This network prepares predicts patient's prognosis & describe complex biological process regarding biological pathways for prognosis.	Future scope is there to reduce overfitting & dimensionality
10	[14]	Acute Myeloid Leukemia (AML)	Deep profile - Autoencoder	13,237 genes	NCBI & GEO	Classificati on error rate is 0.24 for Deep-profile	5-fold cross validation	This model extracts a feature representation from a vast expression data to predict complex disease phenotype	There is a need to handle multiple cancer types by using DeepProfile - using multi- omics data
11	[15]	Thyroid cancer	Denoising Autoencoders and Stacked Denoising Autoencoders	510 cancer samples with 60,483 genes	TCGA	Feature extraction accuracy - 97.36	5-fold cross-validation	Algorithm used here selects highly predictive features from this high dimensional data	More biologically meaningful information needed to be extracted
12	[16]	Find outs oncogenic mutations in gene sequence	CNN	18330 genes	c-Bioportal database	Accuracy-86.68%	3-fold cross-validation	Performs robust classification with limited features	Need to do testing on larger architecture
13	[17]	LUAD, LUSC, KICH, KIRC, KIRP & BRCA	CNN & LSTM	3172 genes	TCGA	Precision= 93.40 Recall= 63.6	80% training & 20% testing	It can handle new sample types without re-training.	Need to consider other clinical parameters with DNA sequence.
14	[18]	BRCA, BLCA, CESC, LAML & LIHC	Network-based deep learning method, called G2Vec	2500 genes	GDAC Firehose	AUC-ROC value = 0.009–0.049	10-fold cross validation	This system find out prognostic gene signatures using distributed gene representations generated by G2Vec	To apply G2Vec framework to multiple omics datasets and to have integrative analysis is a challenge
15	[19]	24 cancer types	Fully connected, feed-forward neural network	2436 genes	TCGA	Accuracy= 91%	4- fold cross validation	somatic mutation based cancer prediction	To improve the classifier performance by training with larger numbers of samples.
16	[20]	COAD, READ, KICH, KIRC, KIRP, LUAD & LUSC	CNN	6045 genes	TCGA	Accuracy= 77.5%	5- fold cross validation	Finds out the most relevant gene contributing in development of specific type of cancer.	Need to provide more strong classification performance to predict multiple cancer subtypes.
17	[21]	Tested on 24 different tissues	DNN	9883 genes	TCGA	Accuracy= 99.70%	90% training & 10% testing	This is trained DNN model used to classify most oncogenic genes.	To develop a common classifier to predict multiple cancer types.

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.

18	[22]	33 cancer types	CNN	10267 samples	ICGC/TCGA	Accuracy= 95.65%	10- fold cross validation	This model is able to accurately distinguish among 23 major cancer types using information derived from somatic mutations alone.	There is a need to improve the interpretation & understanding of deep learning algorithm works.
19	[23]	BRCA, COAD, LUAD, KIRC, Brain & Uterus	CNN	4174 samples	TCGA	Accuracy= 77.65%	70% training & 30% testing	DeepCues algorithm discussed here integrates germline variants and somatic mutations interactively	It is capable of handling expression level, copy number variation, methylation
20	[24]	33 cancer types	CNN	10,340 samples	TCGA	Accuracy= 95.7%	5-fold cross validation	This is a new CNN –based model for cancer prognosis based on gene expression profiles.	Model makes few mistakes while identifying cancer subtypes with the same tissue origin.
21	[25]	Blood, brain, prostate, endometriu m, multi- tissues	Multi-class classification using DNN	23,450 samples	Tomlins-2006-v1, Khan-2001, Lapoint-2004-v2 and Tomlin-2006-v2	Accuracy= 85.5%	80% training & 20% testing	Achieves multi- class classification of gene expression data with the aim to predict the type of cancer.	Transcriptomic and proteomic data could improve upon the classification model
22	[26]	12 cancer types	Autoencoder	59,774 samples	TCGA	Max- C index =0.87	10-fold cross validation	Algorithm selects most informative features to construct the prediction model by Xgboost	Model performance can be improved by combining multi- omics data with multimodal data.
23	[27]	Lung cancer	Feed forward neural networks	72 samples with 12,600 genes	BioLab portal	Accuracy= 0.83 Sensitivity =0.86	10-fold cross validation	Algorithm provides optimized framework for cancer classification using hybrid combination of deep learning and genetic algorithm	Model should be scaled further to deal with large size of oncological data
24	[28]	Lung cancer	CNN	10535 sample with 7509 genes	TCGA	AUC=0.73 Sensitivity = 0.67 Specificity = 0.68	5-fold cross validation	This algorithm selects only high level featured using CNN by transforming RNA- seq samples into gene-expression images.	Same transfer learning approach can be used for pan cancer dataset.
25	[29]	18 cancer types	Deep sequencing using whole genome sequencing	344 plasma samples from 200 patients	Addenbrooke's Hospital, Cambridge, UK	AUC=0.989 for high ctDNA cancers, and AUC=0.891 for low ctDNA cancers	4-fold cross validation	This method provides earlier diagnosis & study of tumor biology	To improve sensitivity by combining fragment segment analysis with other entities in blood such as microvesicles and tumor-educated platelets
26	[30]	Lung cancer	Deep sequencing	42 patients	Private hospital dataset	Sensitivity =98%	Leave-one-cut-out cross validation	Robust algorithm to differentiate early- stage lung cancer patients from matched control group.	To develop CLiP methods for a diverse range of malignancies.
27	[31]	19 tumor types	CNN	20 patients	TCGA	AUC=96.52%	N=427	Classifies glial tumors & can be integrated with somatic variation detection.	Requires more computational power to train CNN network.
28	[32]	Breast cancer	Stacked Autoencoder	33564 features of 305 patients	Integrated dataset	Accuracy= 93.48%	4-fold cross validation	Reduces high dimensions of multi-omics data.	Handling high volume data with multiple dimension is difficult.
29	[33]	Hepatocellular carcinoma	Ultra deep sequencing	11079 samples	TCGA	Accuracy= 92.27%	4-fold cross validation	Typical alterations in genome sequence causing malignant tumor are captured precisely.	Small genetic variations in tumor may not be addressed.
30	[34]	Multiple cancer type	CNN	2318 samples	Cosmic	Accuracy =86%	4-fold cross validation	Predicts the oncogenic potential of a protein sequence resulting from a gene fusion.	Reduction of false positive rate by training large samples

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.

From the above survey, it is clear that a lot of research is on- going in this area. The last column highlights the gaps identified in the development of a predictive modelling system using genome sequencing. The following figure-5 shows a variety of cancer types addressed by different researchers.

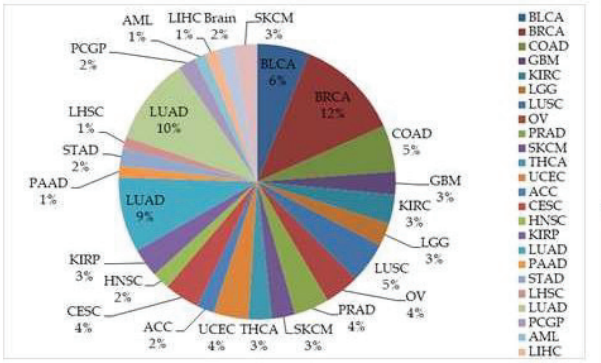

Fig. 3. Cancer types covered in various research studies

We also tried to figure out the most popular algorithms used by researchers in literature. Table 2 enlists different types of Deep Learning Algorithms used by researchers along with serial numbers referred in table 1.

TABLE II. DEEP LEARNING ALGORITHMS USED BY RESEARCHERS

Sr. No.	Name of algorithm	Reference no w.r.t table 1
1.	Deep Neural Networks (DNN)	[1,2,3,9,14,17,21]
2.	Convolutional Neural Network (CNN)	[7,8,12,13,16,18,19, 20,24,27,30]
3.	Backpropagation algorithm	[2]
4.	Deep forest algorithm	[4]
5.	Autoencoder (AE)	[10,22]
6.	Stacked Sparse Auto Encoder(SSAE)	[5]
7.	Stacked Denoising Auto Encoder(SDAE)	[6,11,28]
8.	Long Short Term Memory (LSTM)	[13]
9.	Feedforward Neural Network	[15,23]
10.	Deep sequencing	[25,26,29]

This section also covers types of Deep Learning architectures used in handling genomic profiles. :

A. Deep Neural Network

The term "deep" in name refers to use of multiple hidden layers making a network to learn more complex patterns from the input data with multiple levels of abstraction. In DNN, the input is fed to neurons of the first layer and an activation function is calculated for each neuron. This output is forwarded to the next layer's neurons. If the output of neurons is larger, obviously the significance of the input dimension is also larger. Further, these dimensions are combined in the next layer to form updated dimensions and hence system learns intuitively. The performance of DNN highly depends on parameters selected during the training phase. It is observed that the semi- supervised approach like combining fine-tuning of parameters using backpropagation and use of gradient descent technique to minimize prediction error rate can significantly improve the performance of an algorithm[35]

B. Convolutional Neural Network

It is one of the effective deep learning models because it can automatically perform adaptive feature extraction during the training phase. In the case of genomic profiles, CNN can be applied to discover meaningful recurring patterns. It simplifies the network model by assigning weights on a singular mapping of features so that overall weights can be reduced. CNN has been widely used to model sequence specificity of protein binding[36], to learn the effect of non-coding variants, and to study the functional activities of DNA sequence. One needs to optimize the ability of CNN and choose appropriate CNN architecture as per the need of an application. Specifically, when handling genomic data, deep learning models are always over-parameterized.[37] Thus, only varying network depth may not result in performance improvement. Many other parameters like kernel size, the number of feature maps, the design of convolutional kernel and choice of window size of input sequences must be taken into consideration.[38]

C. Backpropagation algorithm

When input data is fed to the neurons of a first input layer then the end output will be a prediction (sometimes it may be correct or incorrect). If we provide output again as feedback to the neural network to improve using some means to predict better, the system learns by updating the weight for the connections. To complete the process of providing feedback and defining the next step to make changes accurately, we use a backpropagation.[39] Reiterating the process multiple times step-by-step, with more and more data assists the network to update the weights appropriately and leads to a system where it can decide for predicting output based on the rules it has formed for itself through the weights and connections.

D. Deep forest algorithm

In gene expression analysis, the sparsity of effective features is unknown. Therefore to learn from sparse features and feed representation to a neural network, random forest algorithms can be integrated with it and named as Deep-forest algorithms[40]. This algorithm consists of two parts: the forest part which is used to extract features from raw sparsed input with the supervision of training outcomes and the second part is DNN which predicts learning outcomes with the help of new feature representation.

E. Auto-encoders

Auto-encoder is a type of neural network that finds a new representation of input nodes by using unsupervised learning techniques. It captures the significant features of the input data and restores the original data. They are widely used to extract meaningful features from data because they are capable to learn new presentation of data through encode-decode procedure.

F. Stacked Sparse Auto Encoder(SSAE)

If the number of free nodes are more then using the backpropagation and gradient descent approach would not be efficient.[41]

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.

Therefore, the number of hidden units to be activated by introducing sparsity-constraints at each hidden layer is restricted by using stacked sparse auto-encoders. Meaningful features are efficiently extracted from high dimensional data using SSAE.

G. Stacked Denoising Auto Encoder(SDAE)

A simple auto-encoder can successfully remember information from the inputs in new features within the hidden layer. But simply recalling information from the inputs does not assure that the extracted features are good. Denoising auto-encoder generates noisy representation depending on the input values e.g. setting values to 0 for a few input nodes or add a noise term with a Gaussian distribution and then feed that noisy term to the auto-encoder. More robust features can be constructed using noisy terms and thus unseen data samples can be handled effectively. An SDAE is a multi-layer auto- encoder. Each hidden layer is a representation of the previous layer obtained by a denoising auto-encoder with one hidden layer[41].

H. Long Short Term Memory (LSTM)

It is similar to Recurrent Neural Network (RNN). In RNN output from the last step is fed as input in the current step. RNN suffers from the problem of long – term dependencies i.e. it makes predictions based on only recent information and it cannot predict words stored in long-term memory. LSTM is used to predict genotype, grading, treatment plan (based on the risk score).

I. Feedforward neural network

This is a basic and simple type of artificial neural network in which information flows only in the forward direction from the input layer's nodes to output nodes through the hidden nodes. No cycles are formed during designing of the network.

J. Deep sequencing

These methods allow sequencing of billions of nucleotide in a single run. Whole-genome sequencers generate and analyze reads using massively parallel processing reactions that generate multiple reads covering each position in the genome. From the above survey, it is clear that a lot of research is on-going in this area. The last column highlights the gaps. It can be also observed that few researchers have integrated multiple algorithms to enhance the accuracy of the model.

Research gaps identified from the literature survey are mentioned below:

a) Due to the heterogeneity of tumors, constructing a reliable background mutation model is difficult.

b) Need to integrate omics and non-omics data to get better accuracy.

c) Identification of candidate genes that might explain significant response variations. To have a timely diagnosis of critical diseases, including cancers, with the least possible error, is the foremost expectation of patients.

d) The genomic profiles contain a massive set of multidimensional data, which should be analyzed with an appropriate statistical method to extract meaningful information.

e) There is a need to build a strong pre-processing model and appropriate feature selection to reduce the dimensionality and complexity of the dataset.

f) Need to achieve high efficiency in identifying matched target therapies for individuals. Patient response prediction, sensitivity, of a clinical drug toward the multiple diseases is a significant issue.

g) The scope is there to handle algorithmic unfairness model bias, model variance and outcome noise

h) Prediction of generic cancer driver genes using the pancancer dataset.

i) Identification of new pathways and networks needed to discover malignant transformation.

j) There is a need for deeper architecture to handle enormous genome data.

k) The system lacks to handle imbalanced and high-dimensional small-scale data.

l) A larger dataset is needed to avoid overfitting problem and risk backpropagation analysis should be performed for non-linear models.

V. CHALLENGES IN USING DEEP LEARNING ALGORITHMS

From the above literature survey, it is clear that though a lot of research is going on in the area of cancer prognosis using genome sequencing, there is a need of promising research in this study focusing on the following challenges :

1 The genomics profile contains a massive set of multidimensional data which should be analyzed with an appropriate statistical method to extract meaningful information. There is a need to have strong dimensionality- reduction techniques to remove unwanted data and to save computational power by processing appropriate data in subsequent phases. On the contrary side, deep-learning algorithms are capable of automatic feature extraction and encapsulate non-linear relations through convolutions or recurrences. If important features are selected in the initial stage, the accuracy of a model will enhance automatically.

2 There is a necessity to have more patient data for all tumor types. As the performance of Deep Learning Algorithms highly depends on the size of the input, an amount of input data should be more to avoid overfitting problems. Researchers are using techniques like transfer learning to deal with a small dataset.

3 The unbalanced nature of patient data adversely impacts the performance of model. E.g. If we consider clinical records of a patient, there must be an even number of survived and death cases to get better accuracy in the model. Also other performance measures like precision, recall, F- measures can be used additionally to test the results of a model. Guo et al.[8] & Xio et. al.[9] also discussed that a balanced dataset is needed to improve the stability and accuracy of the model.

Handling of missing values in the dataset poses another challenge that too when the size of a dataset

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.

is small. Many missing value imputation techniques like mean-mode- median based imputation, imputation using k-NN, imputation using Multivariate Imputation by Chained Equation (MICE), etc. can be used.

4 The availability of a dataset that integrates omics and non-omics data is crucial because when gene expression data is combined with additional health determinants, better accuracy in results is obtained. For this freely and publicly available datasets are in demand. Authors [5, 18, 22] have already taken this as a challenge as a part of their further expansion.

5 Building a single predictive model to handle multiple cancer types is a big challenge. The accuracy of a model differs from one cancer type to another. Therefore it is very difficult to comment on which model is best. A generalized deep learning model applicable across all types of cancer with a standardized validation method is crucial. Studies taken [5, 7, 8, 11 31, 34] have also highlighted that to deal with multiple cancer using one model is difficult.

6 There is a need to improve annotation performance by integrating more pathway information. For this, we urgently need more medical experts and researchers from the bio- medical field to label the data.

VI. FUTURE DIRECTIONS FOR RESEARCHERS

This section highlights some future directions for researchers in the field of cancer genomics. [42]

A. Advances in NGS:

Massively parallel and High Throughput machines are used for sequencing DNA data, which has raised many opportunities for the researchers. Open research issues are:

- To do DNA sequencing with speed and scalability
- To discover novel pathogens for the identification of diseases.
- To analyze epigenetic factors like DNA methylation and protein-protein interaction.
- To do the sequencing of somatic variations integrated with other types to detect rare diseases.

Development in NGS and in bioinformatics algorithms can be also used to predict drug response resulting in effective personalized cancer therapy.

B. Biomarker-based clinical treatments:

Cancer is a silent killer that propagates in the body without acute symptoms. Cancerous cells grow in an uncontrolled manner causing the healthy cells get converted into infected ones. The Next-generation of clinical treatment should precisely learn tumor heterogeneity using genomic profiles and offer suitable treatment.

C. Combinational immunotherapy

Immunotherapy can be combined with other types of treatment such as chemotherapy, targeted therapy and radiation therapy to acquire expected results.

VII. CONCLUSION

Analysis of genomic profiles using High Throughput sequencing techniques has just changed the way of how we should look at biomedicine. To mine meaningful insights from data are vital for noticing genes disorders which is useful in getting a prior warning to critical diseases like cancer and offer personalized treatment to a patient. Many deep learning algorithms have been proposed to predict harmful gene mutations in cancer. Though no single method is universal, the choice of whether and how to use Deep Learning Algorithms remains problem-specific. This paper has reviewed various research work done in the literature and discussed the motivation behind using Deep Learning Algorithms in cancer prediction using a genomic profile. From the analysis of the existing work, it is concluded that effectively extracting the insights from complex genome data is still a challenging task for time-sensitive decision making in healthcare services.

ACKNOWLEDGEMENT

The quote “Give me the wisdom to know what must be done & the courage to do it” mirrors my gratitude towards Dr. Jayashree R. Prasad under whom I worked as a research scholar. She has not only shaped my views on how the research process should go but also inspired me to know the importance of quality research with a promising impact. Also, I extend my heartfelt thanks to Dr. Rajesh S. Prasad whose kind advice and invaluable encouragement enlightened me to work with even more zeal and enthusiasm. Last but not least, thanks to my family members and friends for trusting and elevating me whenever it was in need.

REFERENCES

[1] Saproo V, Upadhyay R and Valera M 2019 Survey of Feature Selection and Text Classification Methods for Genetic Mutation Classification Int. J. Comput. Sci. Eng. 7 933–7

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

[3] Bhonde S B and Prasad J R 2020 Machine learning approach to revolutionize use of holistic health records for personalized healthcare Int. J. Adv. Sci. Technol. 29 313–21

[4] Anon Publish & Perish Sun Y, Zhu S, Ma K, Liu W, Yue Y, Hu G, Lu H and Chen W 2019 Identification of 12 cancer types through genome deep learning Sci. Rep. 9 1–9

[5] Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis J E, Song C, Gutman D A, Halani S H, Vega J E V, Brat D J and Cooper L A D 2017 Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models Sci. Rep. 7 1–11

[6] Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z and Feng D D 2016 Deepgene: An advanced cancer type classifier based on deep learning and somatic point mutations BMC Bioinformatics 17

[7] Guo Y, Liu S, Li Z and Shang X 2018 BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data BMC Bioinformatics 19 1–13

[8] Xiao Y, Wu J, Lin Z and Zhao X 2018 A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data Comput. Methods Programs Biomed. 166 99–105

[9] Guan R, Wang X, Yang M Q, Zhang Y, Zhou F, Yang C and Liang Y 2018 Multi-label Deep Learning for Gene Function Annotation in Cancer Pathways Sci. Rep. 8 1–9

[10] Cheerla A and Gevaert O 2019 Deep learning with multimodal representation for pancancer prognosis prediction Bioinformatics 35 i446–54

[11] Luo P, Ding Y, Lei X and Wu F X 2019 DeepDriver: Predicting cancer driver genes based on somatic mutations using deep convolutional neural networks Front. Genet. 10 1–12

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.

[12] Hao J, Kim Y, Kim T K and Kang M 2018 PASNet: Pathway-associated sparse deep neural network for prognosis prediction from high-throughput data BMC Bioinformatics 19 1–13

[13] Dincer A B, Celik S, Hiranuma N and Lee S-I 2018 DeepProfile: Deep learning of cancer molecular profiles for precision medicine bioRxiv 278739

[14] Teixeira V, Camacho R and Ferreira P G 2017 Learning influential genes on cancer gene expression data with stacked denoising autoencoders Proc. - 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017 2017-Janua 1201–5

[15] Agajanian S, Oluyemi O and Verkhivker G M 2019 Integration of random forest classifiers and deep convolutional neural networks for classification and biomolecular modeling of cancer driver mutations Front. Mol. Biosci. 6

[16] Choi J, Oh I, Seo S and Ahn J 2018 G2Vec: Distributed gene representations for identification of cancer prognostic genes Sci. Rep. 8 1–10

[17] Jiao W, Atwal G, Polak , 2020 A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns Nat. Commun. 11 1–12

[18] Lin C Y, Ruan P, Li R, Yang J M, See S, Song J and Akutsu T 2019 Deep learning with evolutionary and genomic profiles for identifying cancer subtypes J. Bioinform. Comput. Biol. 17 1–15

[19] Ahn T, Goo T, Lee C H, Kim S, Han K, Park S and Park T 2019 Deep Learning-based Identification of Cancer or Normal Tissue using Gene Expression Data Proc. - 2018 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2018 1748–52

[20] De Guia J M, Devaraj M and Leung C K 2019 DeepGX: Deep learning using gene expression for cancer classification Proc. 2019 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2019 913–20

[21] Zeng Z, Mao C, Vo A, Nugent J O, Khan S A, Clare S E and Luo Y 2019 Deep learning for cancer type classification bioRxiv 612762

[22] Mostavi M, Chiu Y C, Huang Y and Chen Y 2020 Convolutional neural network models for cancer type prediction based on gene expression BMC Med. Genomics 13 1–13

[23] Bigueras R T, Torio J O and Palaoag T D 2018 A data-driven architectural framework for LGUs in disaster preparedness and management system Int. J. Mach. Learn. Comput. 8 454–9

[24] Chai H, Zhou X, Cui Z, Rao J, Hu Z, Lu Y, Zhao H and Yang Y 2019 Integrating multi-omics data with deep learning for predicting cancer prognosis bioRxiv 807214

[25] Sharma A and Rani R 2017 An optimized framework for cancer classification using deep learning and genetic algorithm J. Med. Imaging Heal. Informatics 7 1851–6

[26] López-García G, Jerez J M, Franco L and Veredas F J 2020 Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data PLoS One 15 1–24

[27] Mouliere F, Chandrananda D, Piskorz A M, 2018 Enhanced detection of circulating tumor DNA by fragment size analysis (Europe PMC Funders Group) Sci. Transl. Med. 10 1–28

[28] Chabon J J, Hamilton E G, Kurtz D M, Esfahani, 2020 Integrating genomic features for non-invasive early lung cancer detection Nature 580 245–51

[29] Park H, Chun S M, Shim J, Oh J H, Cho E J, Hwang H S, Lee J Y, Kim D, Jang S J, Nam S J, Hwang C, Sohn I and Sung C O 2019 Detection of chromosome structural variation by targeted next-generation sequencing and a deep learning application Sci. Rep. 9 1–9

[30] Rakshit S 2018 Deep Learning for Integrated Analysis of Breast Cancer Subtype Specific Multi-omics Data TENCON 2018 - 2018 IEEE Reg. 10 Conf. 1917–22

[31] Campo D S, Nayak V, Srinivasamoorthy G and Khudyakov Y 2019 Entropy of mitochondrial DNA circulating in blood is associated with hepatocellular carcinoma BMC Med. Genomics 12 1–11

[32] Lovino M, Urgese G, Macii E, Di Cataldo S and Ficarra E 2019 A deep learning approach to the screening of oncogenic gene fusions in humans Int. J. Mol. Sci. 20 1–13

[33] Zeebaree D Q, Haron H and Abdulazeez A M 2018 Gene Selection and Classification of Microarray Data Using Convolutional Neural Network ICOASE 2018 - Int. Conf. Adv. Sci. Eng. 145–50

[34] Zeng H, Edwards M D, Liu G and Gifford D K 2016 Convolutional neural network architectures for predicting DNA- protein binding Bioinformatics 32 i121–7

[35] J Z and OG T 2015 Predicting effects of noncoding variants with deep learning–based sequence model Nat. Methods Kelley D R, Snoek J and Rinn J L 2016 Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks Genome Res. 26 990–9

[36] Koumakis L 2020 Deep learning models in genomics; are we there yet? Comput. Struct. Biotechnol. J. 18 1466–73 Anon Layman

[37] Kong Y and Yu T 2018 A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification Sci. Rep. 8 1–9 et al. C 2016 乳鼠心肌提取 HHS Public Access Physiol. Behav. 176 139–48

[38] Zhang H and Chen J 2018 J o u r n a l o f C a n c e r Current status and future directions of cancer immunotherapy

Authorized licensed use limited to: University of Prince Edward Island. Downloaded on June 03,2021 at 05:39:01 UTC from IEEE Xplore. Restrictions apply.