A Survey on Classification Analysis for Cancer Genomics: Limitations and Novel Opportunity in the Era of Cancer Classification and Target Therapies

Marwa Abouelkhir Abdelazim1*, Mona Mohamed Nasr 2, and Waleed Mahmoud Ead1

1Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt

2Faculty of Computers and Artificial Intelligence, Helwan University, Egypt

*Corresponding author: marwaaboalkheer@fcis.bsu.edu.eg (Abdelazim)

Abstract:
Advanced machine learning approaches are qualified for recognizing the too composite patterns in the massive datasets. We provide a perspective technical survey analysis in machine learning (ML), and deep learning (DL) approaches for genome analysis. It's quickly rising applications related to cancer diseases such as cancer diagnosis or subtypes of cancer through omics input data. It discusses effective approaches in the fields of genomics regulatory, pathogenicity, and variant calling. Moreover, the representation of ML’s potential benefits due to the several technological platforms involved in its diagnosis, prognosis, and treatment. We concentrate on the most up-to-date knowledge of cancer classification models, targeted therapy, and define how genetic mutations inspire targeted therapy’s responsiveness and highlight the different related issues in this era of precision medicine. Finally, we disuse limitations of the different approaches and hopeful ways of upcoming research in targeted therapy.

Keywords: Deep Learning, Genome Analysis, Precision Medicine, Cancer Classification Models, Classification Models, Omics


Introduction
DNA molecule is translated to mRNA for the synthesis of proteins. Proteins are the primary factors in the utmost cellular processes. The process via which a fragment of DNA is reading besides transformed addicted to a protein has excessive awareness in several therapeutic analyses and biological. This process may affect diverse phases; however, the process's basis is the creature’s DNA that converts according to the sequence of specific rules. Note to toward the full structure of DNA in a creature’s cell the genome. The DNA consisted of four forms of roots – “Cytosine” (C), “Guanine” (G), “Thymine “(T), and “Adenine” (A), the genome of human is made up of completed 3 billion of these genomic letters. The genome can be measured to keep a series cleared on the character set {A, C, G, T}. A sequence of the genome can deliver over the DNA sequencing process. In the latest years, novel sequencing technologies, named Next Generation and Third Generation sequencing (NGS, TGS). The principal reason for cancer has situated a mutation in the genome, each inherited or established through an individual's life.
Cancer disease has been stated as one of the lethal genetic diseases of the human genome. It has been the research attention pending today by clinicians, pathologists, biologists, and other life science and health specialists. The World Health Organization (WHO) informed cancer disease having 14 million novel cases in 2012. Cancer is an essential reason for sickness and death average that explanations equally the second primary reason for death universal causing 8.8 million deaths in 2015. The World Cancer Report well-defined cancer as a universal problem and proposed that cancer occurrence will cause growth to 20 million novel cases by 2025. Cancer disease research's significant aim is to recognize the particular genes that produced normal cells to mutate into cancer disease (Joseph, M., Devaraj, M., & Vea, L. A., 2018). Numerous procedures have been established toward exploring information gained over sequencing to mark variations (besides titled variants or mutations) compared to a recognized normal reference sequence. This allows symbolizing an individual’s (or a population’s) genome sequence(s) in terms of the variants. Variations can be of diverse types, from alterations in single positions to dissimilarities in the construction of thousands of bases' arrangements. Determining variants (or variant calling) has applications in various bioinformatics fields, exploring the study of diseases in humans (Ramachandran, A., Li, H., & Chen, D., 2018).

1.1 Overview of microarray technology and gene expression

Contrasted with the old method of genomic study, which has obsessed with the limited consideration and group of information on particular genes, microarray technologies have prepared it promising to display the expression stages for tens of thousands of genes similar. Upon the arena of molecular natural science biology, the examination of gene code information can deliver direction for cancer's initial identification. Expression gene information covers gene action data and can mirror the biological from the present cell; for instance, a cell is located in the typical or worst state. Though there is a great volume of expression gene information, further than 95% of them are unusable (Zhu, Q. X., Fan, Y., He, Y. L., & Xu, Y., 2018). Though, great dimension, slight sample scope, and askew spreading are altogether appearances of expression gene information. These appearances similarly encounter the taxonomy of subtype's cancer. Consequently; the situation stays vital toward discovery specific genetic factors correlated to the disease of cancer. Nevertheless, within the appearance of enormous genetic factors, characteristic choice stays actual essential. Presently(Zhu, Q. X., Fan, Y., He., & Xu, Y., 2018), the key issue within gene Microarray is grouping because of the thousands of quantities of genetic factor composed, the exclusion element is that growing quickness and the structure of heterogeneous create it greatly extra puzzling to switch biomedical information using such assets than traditional information exploration approaches as standard and that vast quantity of genetic factor can create the classification mission extremely challenging. And now, the period of “precision medicine,” cancer cell treatment stay able to tailor-made to a single patient founded on the outline of a tumor's genomic. Despite the ever-rising richness of genomic cancer information, they connected variation (mutation) profiles to treatment effectiveness remainders difficult (Chang, Y., Park, H., Yang, H. J., & Shin, J. M., 2018). A genomic drug’s purposes are defining how differences in the DNA of persons can impact the hazard of diverse illnesses and discover fundamental reasons that targeted treatments can be prepared. Here, we emphasize how progressive machine learning can use to resolve the main difficulties in genomic treatment. The genomics field is reading the job and data structure prearranged in live cells' DNA structures. Simultaneously, accurate medicine is the exercise of adapting drug
founded on wholly related data around the patient, containing a patient (Leung, M. K., Delong, & Frey, B. J., 2015). Subsequently, the levels of medical discipline and healthiness informatics have prepared major development lately. They have headed to detailed analytics required through generation, gathering, and huge information (Lan, K., Wang, D. T., & Dey, 2018). Solitary point while not capable to unnoticed is that the methods of artificial intelligence applications production an additional important part cutting-edge the achievement of bioinformatics investigation as of genetic facts opinion, then a relationship stands highlighted in addition to recognized toward connection these twofold analytics information procedures and bioinformatics on together manufacturing and academic circles it practical via construction models, building predictions, performing clustering and classifications, discovery “associated rules,” then lastly recognition desired features. Temporarily, “deep learning” be present an additional modern theory and outline, and has greatly well capability of trait demonstration in summary standard than overall machine learning, "One-size-fits-all" methodology toward medication might driving to certain of these sideways special effects, subsequently.

Wholly individuals remain dissimilar. Predicting disease status for difficult human diseases like cancer with genomic data is essential to review the relationship between gene expression profiles and disease states plays a vital role in biological and clinical applications. So, deep learning is appropriate for forecasting medicine -target contact with slight directing. Lately, numerous hopeful consequences have been elucidated via deep learning in medicine development, treatment repurposing (i.e., documentation of potential novel drives of accepted or new medicines), and medicine-target summarizing better forecast accuracy paralleled to other traditional machine learning approaches. And the passionate expansion of biotechnology has been cumulative in the past era of time, the exponential growth average of biomedical

Information produced thru numerous investigation and application regions be able to choose from the scale of micro molecular (gene roles, protein collaborations, etc.), the scale of organic tissue (brain connectivity chart, Magnetic reverberation images, etc.), the scale of a medical patient (“intensive care part,” EMR “electronic medical record,” etc.) and the scale of the whole population (therapeutic message panel, common mass media, etc.) (Lan, K., Wang, D. T., Fong, S., Liu, L. S., Wong, K. K., & Dey, 2018). The remainder of this survey is structured as follows: section two cancer taxonomy, section three bioinformatics input data, section four feature selection methods for informative genes, section five “Machine Learning” (ML) and “Deep Learning” (DL) in Genomics, section six common omics dataset sources, section seven effectiveness of genetic mutation on the susceptibility of targeted therapy and directions of upcoming research, Section eight discussions, and finally, it concludes the survey. The contribution of this work is a:

• We clarify quickly rising approaches related to cancer diseases. for representative the potential benefits of ML due to the various technological platforms involved in its diagnosis, prognosis, and treatment

• discover and analyze the most related works in cancer classification model and feature selection methods with gene expression and genetics information

http://doi.org/10.36295/ASRO.2020.232434
2. Cancer taxonomy

Cancers can be characterized either on the histological kind or their major spot (the location where cancer originated). From a perspective of histological, there are tens of diverse malignancies, which are clustered into six main groups (National Cancer Institute):

- Carcinomas: Greatest Carcinoma impact tissues or glands produce excretion, for example, the organ breasts produce milk, or the lungs, which filtrate slime, colon or bladder or prostate.
- Sarcomatoid malignancy: "Sarcoma" mentions malignance that initiates in caring and connective tissue disease as tendons, muscle, bones, fat, and cartilage.
- Myeloma cells: are malignancy that instigates in the cells of plasma of bone marrow. The plasma cells yield particular of the proteins institute in the blood. Sarcoma
- "Leukemia": ("blood cancers " or " liquid
- Cancers ") are malignancies of the bone marrow (the position of blood cell manufacture). Instances of leukemia contain Myelogenous or granulocytic leukemia (granulocytic white blood cell sequences and hostility of the myeloid):
  - Lymphatic, lymphocytic, or lymphoblastic leukemia (lymphocytic blood cell series and malignancy of the lymphoid)
  - erythremia (malignancy of several blood cell yields, however through red cells dominant)
  - Cancer of the lymph nodes: Lymphomas produce in the nodes of the lymphatic system or glands. Lymphomas could happen in particular tissues, for example, the breast, brain, or stomach.
- Diverse Types: The kind constituents might be within a single kind or from diverse kinds. Designated samples are:
  - Teratocarcinoma
  - Adenosquamous carcinoma

3. Bioinformatics input data

We classified the subjects of interest in the area of bioinformatics into three clusters ("Oomics," "biomedical imaging," "biomedical signal processing").
Genetic data, such as the genome, transcriptome, and proteome information, is applied to bioinformatics methodology issues through omics investigation. The greatest public raw data in omics stand fresh genetic structures (i.e., RNA, DNA, and amino acid structures). Supplementary, mined features from sequences are often used as inputs for deep learning processes to ease complications from difficult genetic information and progress outcomes. Any of the greatest explored difficulties in protein sequence forecast, which targets to forecast the subordinate structure or interaction diagram of a protein gene expression guideline, containing RNA binding proteins or splice junctions, and protein taxonomy including subcellular localization or great family, are similarly dynamically considered. Additionally, variance classification methods have been utilized with omics information to discover malignancy (Min, S., Lee, B., & Yoon, S., 2017).

4. Feature selection methods for informative genes

Gene classification or mining of features in a supervised or unsupervised style like the region of investigation attitudes is a novel encounter for “DNA Microarray expression,” covering above to 25000 genetic factors by the matching time. Separately from imprecating dimensionality, numerous other difficulties are handled in gene choice as supernumerary information, mislabeled data, noisy and unrelated data. Several gene choice procedures and algorithms are applied to decrease dimensionality by removing inappropriate, duplicate, and noisy genes (Hira, Z. M., & Gillies, D. F, 2015). Effective pattern selection has some advantages in such circumstances where thousands of features are complicated. First, dimension reduction is hired to decrease the computational cost. Second, a discount of noises is performed to improve classification accuracy. Lastly, the extraction of additional interpretable features or characteristics that can be supportive of recognizing and monitoring the target diseases. Here are three common groups of feature selection procedures: “filter methods,” “wrapper methods,” besides “embedded methods” (Zhang, D., Zhou, X., & He, F, 2018) as shown in (Fig. 1).
Fig. 1. Feature Selection Methods
5. Machine learning (ML) and deep learning in genomics

Machine learning techniques permit automatic and scalable methods in which visions from huge, multi-dimensional data can be gathered. Computer-based decision-support systems founded on machine learning (ML) can modernize medicine by performing difficult tasks that are presently allocated to specialists to improve diagnostic accuracy, recover clinical workflow, increase the effectiveness of throughputs, improve treatment selections, and reduce human resource costs. There are four common groups of Machine learning algorithms: Supervised algorithms, unsupervised algorithms, Semi-Supervised algorithms, Re-enforcement algorithms, as shown in (Fig. 2). (Table 2) contains all papers which used machine learning algorithms.

<table>
<thead>
<tr>
<th>Cancer type</th>
<th>Datasets</th>
<th>features Selection</th>
<th>ML algorithms</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>Microarray</td>
<td>Scoring method</td>
<td>• (K-Nearest Neighbor) and L- KMOD function</td>
<td>(Zhou, R. &amp; Hu, K., 2017)</td>
</tr>
<tr>
<td></td>
<td>Gene expression data</td>
<td></td>
<td>• Fuzzy SVM</td>
<td></td>
</tr>
<tr>
<td>Colon</td>
<td>microarray</td>
<td>Mutual Information, genetic algorithm</td>
<td>• decision tree (C4.5) classifier</td>
<td>(Pavithra &amp; Lakshmanan, 2017)</td>
</tr>
<tr>
<td>ovarian</td>
<td>TCGA RNAseq, gene expression data set</td>
<td>PGS statistic</td>
<td>• the univariate proportional-hazard model</td>
<td>(Ahn &amp; Park, 2018)</td>
</tr>
<tr>
<td>prostate</td>
<td>genetic factor expression data, protein-protein interaction network</td>
<td></td>
<td>• cluster score (CS)</td>
<td>(Lee &amp; Yang, 2018)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• predicting score (PS)</td>
<td></td>
</tr>
</tbody>
</table>
### Leukemia, Acute Lymphoblastic Leukemia (ALL), Leukemia (AML) 
- Auto microarray genetic factor expression

### Mixed cancer type 
- DNA methylation, gene expression joint sparse canonical correlation analysis (CCA)

### Mixed cancer type 
- DNA microarrays gene expression data mutual information maximization and relief, Genetic algorithm
- SVM
- The (RoF) algorithm
- (CSD-ELM)

### Mixed cancer type 
- capture sequencing, Exome sequencing, Genome sequencing data
- Logistic regression
- random-forest
- deep learning

### Mixed cancer types 
- Gene expression, RNA Sequencing, Protein expression
- Bayesian optimization
- ANN with Cox proportional hazards results in a layer (Yousefi&Cooper, 2017)

### Lung 
- Expression level, RNA-seq
- SVM (Li& Zhang, 2018)

### Lung 
- RNA-seq mutational spectrum information
- SVM (Xiang& Xing, 2019)

### Colon 
- microRNA
- SVM (Lu& Lin, 2018)

### Breast 
- gene expression profile data
- incorporating DNN and (SVM) classifier (sun&li, 2017)
<table>
<thead>
<tr>
<th>Dataset Type</th>
<th>Features</th>
<th>Algorithms Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>rRNA sequence</td>
<td>• (SVM) (Kishk&amp; El-Hadidi, 2018)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Random Forest</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• (NLP) methods such as kmer frequency</td>
</tr>
<tr>
<td>Mixed cancer types</td>
<td>gene expression data</td>
<td>• (SVM) (Roberts&amp; Kennedy, 2018)</td>
</tr>
<tr>
<td></td>
<td>differential variance</td>
<td>• Multiple SVM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Recursive Feature Elimination (Hasri&amp; Kasim, 2017)</td>
</tr>
<tr>
<td>Mixed cancer types</td>
<td>Microarray gene expression dataset</td>
<td>• Fast Correlation-Based Filter</td>
</tr>
<tr>
<td></td>
<td>recursive feature elimination</td>
<td>• (SVM). (Kavitha&amp; Gopi, 2017)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• SVM-Recursive Feature Elimination</td>
</tr>
<tr>
<td>Mixed cancer types</td>
<td>gene expression data</td>
<td>• Mining Synthetic Lethals algorithm (Sinha&amp; Bernards, 2017)</td>
</tr>
<tr>
<td>Mixed cancer types</td>
<td>Copy number and genetic factor expression data</td>
<td>• incorporating Rao’s score test (Han&amp; Lu, 2019) and supervised ML</td>
</tr>
<tr>
<td>pan-cancer</td>
<td>copy number variation (CNV) data and Genetic factor expression</td>
<td></td>
</tr>
</tbody>
</table>

### 5.1 Supervised learning

Contain all machine learning algorithms which map input data to a known class label(s) or target value. There are many supervised machine learning algorithms, such as random forest, linear and logistic regression, and support vector machines that are generally used for classification and regression steps. However, for the extremely large dataset, deep neural network-based classification and regression algorithm are commonly essential to deal with data non-linearity, some advantages, and disadvantages of classification algorithms presented in (table 3).
Fig 2. Machine learning algorithms taxonomy
Table 3. Advantages and Disadvantages of Various Supervised Machine-Learning Algorithms

<table>
<thead>
<tr>
<th>Technique</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-nearest neighbor</td>
<td>Simplicity and fast speed of accomplishment</td>
<td>Noise sensitively aware</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High space complexity required</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Less time consumption and higher accuracy for huge dataset</td>
<td>Cannot give accurate results if there exists dependency among variables</td>
</tr>
<tr>
<td>Decision tree</td>
<td>No domain limitation of the knowledge to construct decision tree</td>
<td>Restrict to one output attribute</td>
</tr>
<tr>
<td></td>
<td>Can easily process the data with high dimension</td>
<td>Performance of classifier is depended upon the type of dataset</td>
</tr>
<tr>
<td></td>
<td>Can handle both numerical and categorical data</td>
<td></td>
</tr>
<tr>
<td>Support vector</td>
<td>Better accuracy compared with other classifiers</td>
<td>Should select different kernel function for different dataset</td>
</tr>
<tr>
<td>machine</td>
<td>Easily handle complex nonlinear data points</td>
<td>The comparison with other methods training process takes more time</td>
</tr>
<tr>
<td>Neural network</td>
<td>Can simulate almost any functions for complex applications and problems</td>
<td>The black box nature, hard to interpret the structure</td>
</tr>
<tr>
<td></td>
<td>High performance of accuracy</td>
<td>High computational cost</td>
</tr>
<tr>
<td></td>
<td>The availability of multiple training</td>
<td>May over-fitting after times of training</td>
</tr>
<tr>
<td>Ensemble</td>
<td>Improvement in predictive accuracy</td>
<td>Difficult for understanding</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Using the wrong ensemble method will get bad results</td>
</tr>
</tbody>
</table>

5.2 Deep Learning Pathway In Genomics

The DL technique subdivisions ML methods related to a superficial “neural network” then by several hidden layers. Also, extra composite factors applied in training; the most distinctive deep learning models contain “deep belief network”
“DBN”), “stacked auto-encoder” (“SAE”), “recurrent neural network” (“RNN”) and “convolutional neural network” (“CNN”). The DL method permits great standard intellection as a huge capacity of dissimilar and high dimensional raw datasets than conventional ML methods (Yu, K. H., & Kohane, I. S., 2018). Over the past seven years, deep neural networks have run to multiple performance breakthroughs in speech recognition, machine translation, and computer vision.

**Table 4. Deep Learning Algorithms**

<table>
<thead>
<tr>
<th>Cancer type Datasets</th>
<th>Selection features</th>
<th>ML and DL Methods</th>
<th>Reference</th>
</tr>
</thead>
</table>
| Breast               | genetic factor expression PCA approach datasets | • autoencoder neural network  
• AdaBoost algorithm to make an ensemble classifier for the concluding prediction job | (Zhang & He., 2018) |
| Lung                 | gene expression data, Protein-protein interaction network convolutional filtering | • (CNN) the method that integrated clustering spectral data handling toward categorizing lung malignancy | (Matsubara & Nahcher, 2018) |
| Mixed cancers types | Genetic factor Expression Omnibus TCGA, Genotype-Tissue Expression, (microarray and RNA-seq) | • (DNN) | (Ahn & Park, 2018) |
| Breast               | gene expression profile data | • incorporating DNN and (SVM) classifier | (Sun & Li, 2017) |
| acute myeloid leukemia | de novo AML were retrieved from the TCGA database, Demographic information | • Deep learning neural network with stacked (multi-layered) autoencoder, in which high-level features are compressed, organized, and extracted | (Lin & Nguyen, 2018) |
Formative studies in 2015 established the applicability of deep neural networks to DNA sequence data, as shown in (table 4). As presented in (Fig. 3) A, a raw dataset must remain arbitrarily divided into training, validation, and test groups. The negative besides positive sets must stay stabled to prospective confusion (e.g., the sequence position and content); therefore, the forecaster acquires distinguished features by preference baffle. B, the suitable construction remains chosen and taught on the source of field acquaintance. As “CNN's” seize conversion Constancy, then “RNNs” take extra elastic spatial communications. C, “false positive” (FP), “false negative” (FN), “true positive” (TP), and “true negative” (TN) averages are estimated. Once there are additional “negative” than “positive” samples, recall and precision be there frequently measured. D, The learned model is inferred via processing in what manner altering every nucleotide in the input impacts the forecast (Zou, J., Huss, M., Abid, A., & Telenti, A., 2018).
### Table 5. Bioinformatics datasets source

<table>
<thead>
<tr>
<th>Source</th>
<th>Source URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>“The European Genome-phenome Archive (EGA)”</td>
<td><a href="https://ega-archive.org/">https://ega-archive.org/</a></td>
</tr>
<tr>
<td>“International Cancer Genome Consortium (ICGC)”</td>
<td><a href="https://dcc.icgc.org/">https://dcc.icgc.org/</a></td>
</tr>
<tr>
<td>“Genotype-Tissue Expression - GTEx Portal”</td>
<td><a href="https://gtexportal.org/">https://gtexportal.org/</a></td>
</tr>
<tr>
<td>“NIH Roadmap Epigenomics Project”</td>
<td><a href="http://www.roadmapepigenomics.org">http://www.roadmapepigenomics.org</a></td>
</tr>
<tr>
<td>“the 1000 Genomes Project”</td>
<td><a href="http://www.internationalgenome.org">http://www.internationalgenome.org</a></td>
</tr>
</tbody>
</table>

7. **Effectiveness of genetic mutation on the susceptibility of targeted therapy and directions of upcoming research.**

Genetic alterations in malignancy are created from together inherited and lifestyle sides, but many malignancy-related mutations are because of randomized DNA repetition faults. Remarkably, mutations in cancer treatment targets can significantly influence medicine sensitivity. Therefore, the effectiveness of directed treatment is principally reliant on the profile mutation of malignancy in patients. The core challenge of directed treatment nowadays that considers new future investigation ways (1) confirm and recognize chief altered genetic factor and proteins in malignancies such as novel objectives; (2) classify patients maximum unlikely and likely to benefit from confident directed treatments; (3) assess the technique of mutation-driven treatment confrontation [Jin, J., Wu, X., Li, J., … & Cho, C. H., 2019]. In [Chang, Y., Park, H., … & Shin, J. M., 2018], "report Cancer Drug Response profile scan" (CDRscan) the novel deep learning approach which forecasts anticancer medicine responsiveness depended on a great-domain medicine showing analyze data surrounding the genomic profiles of 787 humanoid malignancy cell lines and mechanical profiles of 244 medicines. This precise and vigorous drug response prediction model symbolizes a significant breakthrough for the awareness of correctness cancer drug over its presence in the medicine progress procedures such as medicine repurposing and screening a slight biochemical reference library for novel anticancer medicine applicant. In the medical surroundings, “CDRscan” is predictable towards modernizing “patient-tailored anticancer” medicine choice like a medical “decision support system” DSS through more medical justification readings. In [Wen, M., Zhang, Z., Niu., & Lu, H., 2017] Recognizing interactions amongst well-known medicines and objectives are the main encounter in drug relocation, operative “DL” approach (DBN) stayed practical toward precisely forecast novel drug-target interaction (DTIs) among FDA accepted medicines and targets deprived of separating the target into diverse modules. The established methodology is called DeepDTIs. This procedure can accomplish comparatively great expectation performance. Supplementary exploration of novel predicted DTIs designated. In [Eraslan, G., Avsec., & Theis, F. J., 2019] assessment the ‘big data’ methods aimed at detecting the driver genes in malignancy and how huge-domain several-platform exploration of big associates of patients be able to apply to recognize driver genes that are possible malignancy medicine targets. Over modeling applying tackles as ML decision tree founded classifiers or other algebraic approaches, these raw datasets can analyze composed toward categorizing main
genetic factors that must make the probable to remain tumor medication objectives. In [Huang, C., Mezencev, R., & Vandenberg, F., 2017] at this juncture, an exposed source software platform smearing a greatly adaptable (SVM) approach, which employs typical (RFE) approaches to forecast tumor medicine reaction. Via using ovarian cancer (OC) patient genetic factor expression, raw data made forecasts greatly reliable by detected replies to various treatments.

8. Discussion

(Zhou, Rel at., 2017), introduced "Fuzzy Support Vector Machine" for breast tumor genetic factor classification, suggested a way to mine a great quantity of data distinguishing the genetic factor, choosing genes with maximum categorized data, also in the meantime eliminating the association appearances of them and considered a "geometric means" KNN membership functions and "L-KMOD" the main function "FSVM" to discriminate breast virulence genetic factor. The consequence error average is 3.89%; the uppermost precision percentage is 98.9%. The limitations of this paper are that the challenges of small sample sizes

(Zhang, D el at., 2018), the technique suggested here shows “unsupervised learning” and “supervised learning.” To enhance cancer diagnosis, predicate, and progress, an additional, comprehensive consequence classifier. In “the unsupervised learning level,” “PCA” procedure and DNN are pragmatic to acquire abridged patterns through gene actions, however in “the supervised learning stage,” an “ensemble classifier” is prepared to afford to the patterns learned previously. The limitations of this paper are that firstly, the model constructed is not easy to analyze. Also, identifying which features are most important to the prediction task is difficult. Secondly, due to the deep learning model's complex structure, the amount of data is less prone to over-fitting.

This is a common problem in neural networks.

(Pavithra, D el at., 2017), focuses on a different variable selection technique based on filter, where a theoretical examination of “Mutual Information” was built on Feature Selection. To choose the best essential gene by applying “Mutual Information.” Then Wrapper founded that feature selection was constructed on the “GA algorithm.” In the “GA algorithm,” main genes are designated for tumor micro-array information; formerly, the designated features are inputs to (C4.5) classifier. In” MI, “it's founded a Feature Selection method that selects a feature from 200 variables nominated, and then the precision will be 87.09677%. Nonetheless, there are eight features selected in the genetic algorithm; then, the precision will be 88.70968%. The limitations for this paper are that the challenges of small sample sizes and the low accuracy to increase the accuracy use the mix of other techniques of feature selection methods and local search methods

(Matsubara, T el at., 2018), novels approach that collective “spectral clustering” with “CNN” to classification lung malignancy, which mixes protein collaboration network and “genetic factor expression profiles” information. The methodology combines transcriptome and proteome information and can produce effective and precise predictions by handling network info and applying a “CNN” outline. The averaged consequences subsequently “k-fold CV” was achieved deliver as “accuracy”=0.81, “recall”=0.88, “precision”=0.78, and “specificity”=0.74. The limitations for this paper are that small sample size to improve the prediction combining datasets from different sources

(Ahn, T el at., 2018), analyzed RNaseq information from the ovarian malignant tumor patients; in this paper, the important genes were designated through the “p-value” gained beginning “the univariate Cox PH model.” The genetic factor thru “p-value” less than 0.01 stayed nominated in this examination. In the following phase, genetic factor was
categorized into “Poor genetic factor” “and “Good genetic factor” over the designated genetic factor’s measurement symbol. They designed (PGS) utilizing the mean and standard deviation of the matching in the train set’s genetic factor expression.

The limitations of this paper are that approach used in this paper is not enough to predict the survival outcome, which even causes an extremely opposite prediction (Lee, J. Y el at., 2018), Predictable PCa28 as the prospective genes for analytical biomarkers of prostate tumour utilize omics data. PCa28 be able to distinguish among the typical and tumour tissues and are particular for prostate tumour, Amongst PCa28, various novel prostate malignance correlated genetic factors, as COL17A1 and HIST3H2A, were recognized by applying “predicting score” (PS) and “cluster score” (CS) functions. The limitations of this paper are that the challenges of small sample sizes

(Ahn, T el at.,2018), to identify the difficult environment of cancer founded on the massive communal dataset, this paper proposed a common classifier by (DNN) of cancer information causing starting 24 tissues of origin to detection genes commonly donate to classify tumour in a separate sample and the classify cancer and typical data with an accuracy of 0.997. DNN appeared as a useful gene expression-based classifier tool, as it performed usually higher than other methods, including logistic regression, ridge, lasso, elastic net, and SVM, in their experiments. One of the main drawbacks of a DNN is the difficulty in interpreting an individual's specific contribution to the outcome.

(Zhu, Q. X el at.2018), proposed “multidimensional mutual information” (MMI) feature selection practice to choose the greatest explanatory features for classification and eliminate dirty data. Subsequently, feature selection applying the proposed MMI, “Extreme Learning Machine” (ELM), is applied as a resourceful classification. The suggested, “MII-ELM algorithm” accomplishes the “accuracy” of 98.06% on Leukemia raw inputs per 8 features. This paper's limitations are that the challenges of low sample sizes and the proposed algorithm are unstable.

(Fang, J el at.2017), proposed a different process to cooperatively identify difficult relations amongst “DNA methylation” and “genetic factor expression” stages from several cancers. The core impression is to smear “joint sparse canonical correlation analysis” to identify a minor set of methylated spots related to an alternative set of genetic factors, whichever communal crossways cancers or detailed to a specific cluster (group-specific) of cancers. This study's limitations: First, they assume that the standardized methylation level follows the multivariate Gaussian distribution. It is not always true in real data. Second, parameter selection for sparse joint CCA is still a challenging problem. Finally, more biological evidence is required to support the results.

(Yan, K el at., 2018), introduced a fresh mixture feature selection outline for genetic factor expression information classification. A mixture feature selection outline suggests incorporating filter-based approaches and wrapper based approaches through middle stage assessment via an ensemble classification. A supplementary comprehensive ensemble gene selection procedure founded on a comprehensive “GA algorithm.” Respectively classifier of “CS-D-ELM,” “SVM,” and” RoF” is applied equally as an assessment task to choose essential genetic factors founded on “the EGA algorithm.” The limitations of this paper are that the challenges of low sample sizes

(Ainscough, B. J el at., 2018), documentation of somatic variations in sequencing information thru applying “the random forest and DL models” understood great (percentage AUCs>0.95) “classification enactment” through wholly alternative refinement curriculums (somatic, ambiguous, and fail), although “the logistic regression” confirmed abridged enactment (percentage AUC=0.89). This study's limitation is that the training data probably contains a substantial amount of noise that might impact model performance.
(Shen, Y et al., 2018) Created a novel classification process that associates “fused lasso and elastic net” by way of regularization for “linear support vector machine” (SVM) that named it “oriented feature selection SVM” (OFSSVM)—recognized its efficacy in “binary classification” (i.e., formative, whether must a malignance or the sub-types of spite) and multi-class classification (i.e., formative numerous cancer categories composed) in the intelligence of all-inclusive assessment which not solitary the classification accuracy but also the interpretability is restrained. But it also brings some problems although features have been selected, they do not provide the same reliability as forward- and backward-stepwise selection.

(Yousefi, S et al., 2017) determine in what way DL and Bayesian optimization approaches that must be abnormally effective in common great dimensional forecast errands can be adjusted to the problematic of predicting tumor consequences. Deep survival models can effectively broadcast information through diseases to progress analytical accuracy. Extrapolative information as of multi-cancer raw datasets, recognize perfect hyperparameter surroundings, valid users’ substantial time and effort in selecting model parameters. A fundamental challenge in deep learning is determining the network design that provides the best prediction accuracy. This process involves choosing network hyperparameters, including the number of layers, transformation types, and training parameters.

(Li, T et al., 2018), this study forecast tumor prognosis in lung adenocarcinoma patients, expect five-year survivorship on or after the expression glassy, patients stood separated into twofold clusters, the cluster of patients through lesser hazard is indicated thru persistence age longer than five years, and the cluster of patients with a greater hazard is indicated by survival era littler than five years, applying a modest SVM model on eight applicants genetic factor founded in aforementioned molecular knowledge and feature selection. The Limited accuracy of cancer prognosis in this paper due to small patients sample size, growing the sample size will critically increase the accuracy of machine learning models, also an unlimited number of training sample because each cancer patient is unique in terms of their genetic background, health status, treatment plan, and other environmental factors.

(Sun, D et al., 2017), recommended an innovative way called “D-SVM,” incorporating “DNN” and out-dated (SVM) classification, and mRMR solitary stayed applied to select a slight measure of genetic factor and reduction of the features. At this juncture, a DNN was recycled by way of pattern extraction toward proficiently apprehension genetic factor expression features on or after breast cancer raw dataset. The variables remained consequently utilized towards the train and test the ML process for humanoid tumor prognostication. The limitation of this paper is that breast cancer is a complicated disease, and several different factors affect cancer survival rates. Using a single omics dataset format not enough to achieve a better robustness and scalability algorithm for human breast cancer diagnosis prediction, so the use of integrating multi-omics datasets enhance the prediction model

(Han, Y et al., 2019), DriverML recognizes cancer driver genetic factor through merging a biased score test and machine learning method, applied DriverML to the 31 cancer-specific mutation datasets from TCGA, “Rao’s score statistic” is arithmetically appropriate to association several mechanisms through a usual of weight factors to yield a biased universal indicator, next, the weightiness cutting-edge the slash statistics measure the purposeful influences of diverse alteration categories on the protein. A major challenge faced in this paper is that the cancer genome sequencing is to identify Cancer-associated genes with mutations that drive the cancer phenotype

(Wisely, U. N et al., 2017), the problem expanded through the classification procedure is the capacity of micro-array information variables is enormous. Here, applied “Deep Belief Network” aimed at data classification and Mutual
Information equally a feature selection technique. The constructed system can categorize the raw microarray dataset with a percentage of 90.84% accuracy and a percentage f1-score of 89.68%. The limitation of this paper is that the low sample size (Roberts, A. G el at., 2018), this study offered feature selection through difference variance to the supplementary overall problematic of classification of subtypes cancers. Predict cancer development or diagnosis via SVM and features designated by improved gene expression variance in tumor tissue paralleled with typical tissue. Then that merging the two methods frequently springs enhanced classification outcomes than either feature selection technique only. This paper's limitation is that all datasets applied here are microarray, so if their applied RNA-seq is more accurate and less noisy.

**Conclusion**

This survey discusses machine learning and deep learning algorithms that widely used for cancer classification using gene expression dataset; for example, the analysis of cancer or subtypes of cancers is a revolutionary field of precision medicine, and drug decisions are gradually dependent on the accurate molecular level and genetic profiling of tumor cells. Discuss effective applications in the arenas of regulatory genomics, pathogenicity, and variant calling. Simultaneously, modest and effective feature selection is useful for understanding the cause of cancer and affords a theoretic basis for pathologists to research cancer and select targeted therapies.

**Appendix A. key terms**

<table>
<thead>
<tr>
<th>Key term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Gene”</td>
<td>“A gene is a specific part of a DNA molecule that contains whole the coding data requiring to monitor a cell to manufacture a selected product, for example, an RNA molecule or a protein. Contain within the gene are parts that we identify as active in the coding process (exons), in addition to noncoding parts (introns).”</td>
</tr>
<tr>
<td>“Gene expression profile.”</td>
<td>“A research laboratory way that recognizes whole of the genetic factor in a cell or tissue that are producing messenger RNA. Messenger RNA molecules transmit the genomic data that stands required toward creating proteins on or after the DNA in the center of the cell to the cytoplasm wherever the proteins are synthesized (Stranger, Barbara E., et al., 2017)”</td>
</tr>
<tr>
<td>“Gene mutation analysis.”</td>
<td>“Malignancy is a genomic illness thru Physically gained genetic abnormalities. Driver alterations are wanted for the cancer occurrence, even though passenger alterations are unrelated to cancer growth and collect thru DNA duplication; numerous main cancer sequencing tasks, for example, “the International Cancer Genome Consortium” (ICGC), “the Therapeutically Applicable Research to Generate Effective Treatments” (TARGET) in addition to “The Cancer Genome Atlas” (TCGA), have shaped a general collection of somatic changes through whole several cancer categories (Yi Han, el at., 2019).”</td>
</tr>
</tbody>
</table>
Biomarkers”  “As said by “the National Cancer Institute,” “a biomarker” is “a biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease” (NCI) for example cancer.”

“personalized medicine.”  “Speedy advances in genomics and molecular biology, novel markers can be discovered for the presence of or sensibility to a disease, or response to therapy, and these advances can help to find special treatment for every patient.”

“Artificial intelligence.”  “A division of computer science that tries to do both recognize and construct intellectual objects, frequently programmed as software programs.”

“Deep learning.”  “A branch of the greater machine learning domain. Deep learning occupations artificial neural networks with several artificial neurons’ layers. The outcome of a single layer is served as input hooked on the following layer to get better elasticity.”

“Supervised machine learning.”  “Supervised machine-learning approaches effort through recognizing the input-output association in the ‘training’ level and via exhausting the known association to predict the accurate output of the novel suitcases.”

“Unsupervised machine learning.”  “A category of the machine-learning task that aims at concluding essential patterns in unlabeled data.”

References


