Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

An Artificial Intelligent System for Breast Cancer Sub typing: A Deep Learning Approach

Deeba Khan*

Computer Science and Engineering, Ramaiah Institute of Technology,
Bengaluru, Karnataka,560054, India
deebamajeed@gmail.com

Seema S

Computer Science and Engineering, Ramaiah Institute of Technology,
Bengaluru, Karnataka,560054, India
seemass@msrit.edu

Article Info
Page Number: 2833 - 2856
Publication Issue:
Vol 71 No. 4 (2022)

Article History
Article Received*: 25 March 2022*
Revised*: 30 April 2022*
Accepted*: 15 June 2022*
Publication*: 19 August 2022*

Abstract

Classification of invasive Breast Cancer, based on source, into: Ductal and Lobular is an important clinical problem since it promises targeted therapy rather than a common treatment regime for all the patients. These predominant types exhibit highly distinguishable genomic characteristics with genes like TBX3 and GATA3 mutated in Ductal and Lobular cases respectively making them targets for the treatment. In this work we propose an AI system that exploits deep learning techniques and multiomics data for classification of patients based on tumor source. The proposed system is based on a hybrid optimization model - Hunger Customized Narrowed exploration (HCNAE) that combines the traditional Aquila Optimizer and Hunger Games Search in a conceptual way. The hybrid deep learning model pools optimized features from LSTM and MSE-CNN models that results in higher accuracy. The loss function of CNN is modified to reduce the classification losses. Finally, a comparative evaluation is performed to ensure that the proposed model is effective in identifying BC subtype based on the source.

Keywords: Invasive Breast Cancer; Median-based Z-score normalization method; Hunger Customized NArrowed Exploration (HCNAE); Mean Square Error Based (MSE) based Convolutional Neural Network (MSE-CNN), Long-Short Term Memory (LSTM)

1. Introduction

A huge women population across the globe is being highly affected by Breast Cancer (BC), and it is said to be the second major cause for the death for women. In India alone, there were approximately 260,000 new instances of female BC and 40,000 fatalities in 2019 [1] [2] [3] [4]. These figures indicate that developing strong knowledge-based diagnostic and prognostic technologies capable of generating phenotype estimations for a person is crucial as they assist in tailored therapy. BC has been reported to be a heterogeneous disease with several molecular subtypes, each with diverse molecular properties as well as clinical signs [2]. Personalized medicine attempts to give the most appropriate treatment plan depending on the patient's medical history, genetic traits, and responsiveness to medication to resolve the issue [5, 6]. During the last few years, cancer research has undergone a steady transformation, resulting in clinically meaningful subtyping [1] [5] [6] [7] [8].

Researchers have used many approaches, including early-stage screening, to detect tumor forms before they start showing symptoms. They have also introduced new opportunities for predicting cancer therapy outcomes earlier [9] [10] [11] [12]. Three biomarkers, ER, PR, and HER2, have customarily been used to distinguish BC in terms of prognosis. Such biomarkers are indeed effective therapeutic targets [7]. The Cancer Genome Atlas (TCGA), a popular project related to cancer research has also put efforts in understanding the nature of two types of invasive breast cancer: Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC) [13]. IDC is the most common type of breast cancer that develops in the milk ducts of the breast, it comprises of about 65-85% of total cases. The other variant, ILC constitutes 10% of all cases is considered to be an advanced type. ILC develops in the breast milk-producing lobules. The distinctive genomic characteristics of ILC distinguishes it from the IDC, providing prospects for targeted therapy.

ILC type exhibit CDH1 and PTEN loss. Also, mutation is observed in PTEN, TBX3, and FOXA1 in case of ILC. PTEN loss is associated with increased AKT phosphorylation, which was highest in ILC among all breast cancer subtypes. Spatially clustered FOXA1 mutations correlated with increased FOXA1 expression and activity. Conversely, GATA3 mutations and high expression characterized luminal A IDC, suggesting differential modulation of ER activity in ILC and IDC.

Identifying these genes and labeling patients as ILC or IDC cases provide more personalized treatment opportunities. In this work we exploit unique opportunity provided by modern high-throughput technologies in classifying BC patients based on genetic features represented by multi-omics profiles. The correct diagnosis and prognosis of clinical outcomes is amongst the most intriguing as well as difficult challenges for clinicians [14] [15] [16]. “Numerous interim analyses

of gene expression, somatic mutation, CNV, and protein expression" data have been published in the literature as a result of large-scale collaborative efforts such as TCGA and International Cancer Genome Consortium" [17] [18] [19]. While these platforms have providedaccess to a large amount of curate data, it is critical to solve the long-standing bottleneck of omics integration in order to gain a better understanding of cancer prognosis and phenotype. In numerous cancer researches, "multi-omics data" integration has emerged as a potential technique for predicting clinical outcomes and identifying biomarkers [14] [20]. In cancer research, modeling of survival and treatment response clinical outcomes can serve as a stepping stone toward individualized therapy [21] [22] [23]. The integration of omics enables to investigate the genome sequence at numerous degree of sophistication at same time as well as establish concluding remarks. Because of the significant dimension and variability associated with omics datasets, linear forecast models for these kinds of research frequently fail. As a result, a sophisticated integrated method is necessary to manage this disparate information in a logical manner.

As a consequence, medical researchers are increasingly using Machine Learning (ML) approaches. Such algorithms are capable of discovering and identifying patterns and links in massive data, as well as successfully forecast cancer subtypes [24] [25]. It seems obvious that using machine learning techniques to predict cancer susceptibility, relapse, and mortality could enhance treatment outcomes. Researchers have indeed devised a number of computer algorithms for predicting metastasis as well as its important attributes. Some approaches used clinic pathological features, while others used "image-based features or text-based features", and the most modern techniques have used "omics-based data". This article highlights *in silico*source-based subtype prediction algorithms that include omics information as features. To enhance the classification accuracy, the features needs to be extracted precisely [26].

Recently, in computational biology and bioinformatics,"Deep Learning-based technologies" have been extensively used. The advantages of learning non-linear functions and recovering low - dimensional feature representations [3] show how Deep Learning (DL) models have progressed. We combine multi-omics data with Deep Learning-based forecast models as a result of these considerations. Although most current methodologies use only one or few of types of omics data, such as "mRNA-seq and miRNA-seq data", we believe that combining greater varied data might contribute to stronger prediction, especially when deep learning is used.

The key contribution of this research work is:

(a) To introduce a new Median-based Z-score normalization method for data pre-processing purpose.

(b) To extract Tangent Weighted Entropy (TWE) and higher order statistical features"

(c) To introduce a new Hunger Customized Narrowed Exploration (HCNAE) for optimal feature selection. This HCNAE is the conceptual blend of standard Aquila Optimizer (AO) and Hunger Game Search (HGS), respectively.

(d) To develop a new hybrid deep learning model with standard LSTM and MSE-CNN [27] [28] models for BC classification.

In this sectionwe discussthe literature on BC-omics data. In the following sections the data sources, the proposed classifier, feature extraction and feature selection techniques used, the BC classifier using a hybrid DL model are discussed. Finally, we present the results and discuss our findings.

1.1. *Related work*

In 2015, Giovanni Ciriello et al. [29] in their study highlighted that ILC and IDC are clinically and molecularly distinct diseases. They summarized that ILC patients suffered loss of biomarkers like CDH1 and PTEN, AKT was found to be activated and genes such as TBX3 and FOXA1were mutated.Conversely, GATA3 mutations and high expression characterized IDC, suggesting differential modulation of ER activity in ILC and IDC.

In 2021, Liu et al. [25] have embarked on a project new computational technique for subtyping BC. The BTF has been leveraged with multi-omics data from 762 BC patients collected from: "The Cancer Genome Atlas, including RNA-sequencing expression profiles, copy number variation, and DNA methylation". Using the BTF's factorized latent characteristics, they used a "consensus clustering method" to identify BC subtypes. KM estimators were being used to look at the subtype-specific survival patterns of BC patients. Several state-of-the-art methodologies for "cancer subtyping" were contrasted with suggested method. Even though, the projected model has recorded better statistical outcomes, it still suffers from the drawbacks of lower detection accuracy and precision.

In 2021, Johnson et al. [30] have presented an OMS map derived from four serial biopsies taken over the course of 3.5 years of treatment in a woman with metastatic BC. This resource connects diagnostic and improvisational assessments, including such "comprehensive DNA, RNA, and protein profiles; images of multiplexed immunostaining; and 2- and 3-dimensional scanning electron micrographs", to in-depth, "longitudinal clinical metadata" that contains treatment duration and dosing frequency, anatomic imaging, and blood-based response metrics. Such findings disclose details on the tumor genome's variability and development, as well as signal transduction, the immunological microenvironment, cellular composition and structure, and ultra structure. They

show how integrated analysis of this data might identify potential mechanisms of responsiveness and tolerance, as well as uncover new therapeutic susceptibility. But, the computational complexity is higher in terms of response time and memory consumption.

In 2020, Chen et al. [31] Have generated an IMNA with the objective of identifying great promise essential genes in regulatory networks by assimilating interactions of molecules across numerous biochemical scales, such as "GWAS signals, gene expression-based signatures, chromatin interactions, and protein interactions from network topology". They used this method to prioritize essential genes implicated in regulatory networks in BC. They additionally created an AGES profile for BC depending upon that "gene expression deviation of the top 20 rank-ordered genes". AGES scores have been linked to genetic variations, tumor characteristics, and survival of patients. However, potential important genes without genetic support were not taken into consideration.

In 2020, Cui et al.[32] have used a "multi-omics integrated approach based on long non-coding RNAs (lncRNAs)" to evaluate medication reactions. Considering "lncRNA, microRNA, messenger RNA, methylation levels, somatic mutations, as well as the survival data" of cancer patients receiving with medications, they found DRlncs. To find DRlncs for various chemotherapeutic medicines in BRCA, researchers used an "integrative and quantitative multi-omics method". To find DRlncs for various chemotherapeutic medicines in BRCA, researchers used an integrative and quantitative multi-omics method. Adriamycin, Cytoxan, Tamoxifen, and all BRCA patient samples have been used to identify certain DRlncs. Such DRlncs exhibited distinct characteristics in terms of expression as well as computational precision.

In 2020, Rahman et al. [33] have analyzed GDF10's potential as a medicinal biomarker for human BC using a multi-omics approach. Using the Oncomine, GEPIA2, immunohistochemistry, and UALCAN datasets, they looked at GDF10 mRNA expression patterns in BC subtypes. The assessment' findings showed that GDF10 expression in BC subtypes was downregulated. By analyzing 16 BCE studies from the cBioPortal database, three additional missense mutations in the GDF10 protein sequence were discovered with a frequency of 0.62 percent –2.95 percent copy number alterations. Additionally, Kaplan-Meier plots demonstrated a link among GDF10 downregulation and a lower chance of survival in BC patients. GDF10's co-expressed gene expression has also been linked with the development of BC. However, interlink features between the genes has not been considered.

In conclusion, the ML approach was chosen differently in different studies. More comparative research, that is, research that employ more than one classifier and offer numerous accuracy metrics, are required, but for the time being, it appears that SVM is indeed the better performing

classifier, followed by RF. Because "microRNA and DNA methylation levels" impact gene expression, contemporary omics-based methodologies employing a system biology model for detecting metastasis are crucial, and should have been included when constructing a model. Such an investigation, unfortunately, necessitates a large volumes of information, and feature selection has become a time-consuming procedure that limits model generalization ability. A novel technique is required to improve the models on the basis of predictive accuracy and generalization. This technique might be DL, which combines an automated feature selection procedure and directly captures the nonlinear and complicated interactions of high - dimensional data or noisy biological data. Furthermore, when adapted to different cancer-related prediction tasks, DL techniques have outperformed or were on par with other ML approaches that need an outlier detection phase. Nonetheless, there are few up-to-date DL techniques established to enhance the metastatic prediction problem. To minimize over fitting and propose a more flexible prediction model, DL, requires more data than typical ML models. Nonetheless, emerging approaches like "zero-shot learning" and "few-shot learning" help to overcome this barrier to a certain degree. Methods like resampling and cost sensitive learning [8] can also be utilized to address the imbalance in accessible omics data. However, in order to produce algorithms that can generalize effectively over a wider range, additional comprehensive and varied training sets are usually required. Dealing with cases when the number of features exceeds the number of data samples for model training and validation is yet another problem. Therefore, the deep learning with optimization technology [24] can be used.

2. MATERIALS AND METHODS

In this section data used for the study and the proposed model is summarized.

2.1. *Data*

Multi omics data of 705 breast tumor samples from TCGA was used for the experiment. Of the total samples 574 and 131 are IDC and ILC cases respectively. The omics data type and the number of features for each type are summarized in Table 1. The data used for training the model has already been considered earlier for astudy on bio-markers of cancer based on multiomics data [34]. In total 1936 features across 4 omics types were considered.

Table 1. Summary of data used for training the classifier

Omics Type	Number of features	Description of Omics type
Copy number variations(DNA)	860	Number of copies of gene in the cancer.

Somatic mutations (DNA)	249	Whether the cancer has mutated in a given gene.
Gene expression (mRNA)	604	How much is a certain gene being expressed.
Protein levels	223	How much is a certain gene being expressed.

2.2. *PROPOSED BC DETECTION MODEL: AN OVERVIEW*

2.2.1. *Architectural Description*

A novel source-based BC classification model is introduced by following three major phases: "(a) data pre-processing, (b) feature extraction and (c) optimal feature selection and (d) detection". The architecture of the projected model is shown in Fig.1. Let the input data be denoted as $D_i^{inp}; i = 1,2,..n$ .The steps followed in the projected model are mentioned below:

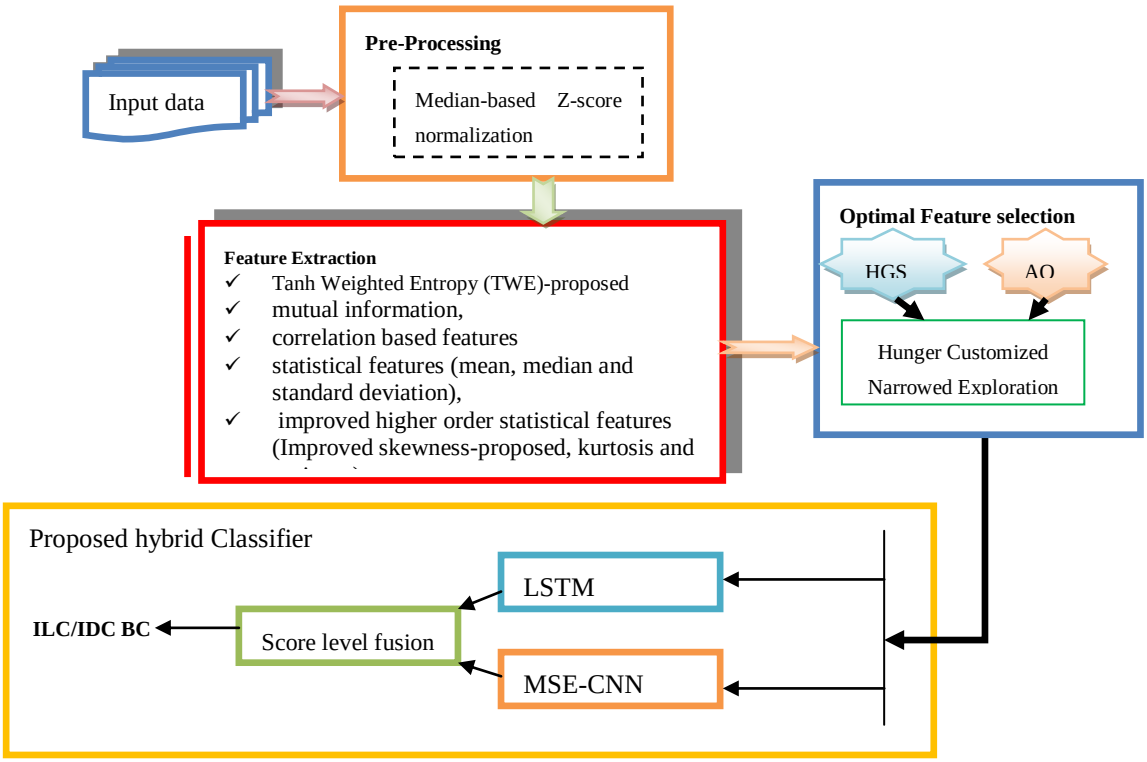


Fig. 1. Architecture of the projected BCSubtype Classifcation Model

Step 1- Initially, D_i^{inp} is pre-processed via median based Z-score normalization (proposed). The resultant pre-processed data is denoted as D_i^{pre} .

Step 2- Then, from D_i^{pre} , the features like "TWE g^{IE}, mutual information g^{MI}, correlation based features g^{C}, and statistical features g^{stat} (mean, median and standard deviation), improved higher order statistical features g^{stat-h} (Improved skewness, kurtosis and variance)" are extracted. These retrieved features are fused together, and it is pointed as = $G = g^{IE} + g^{MI} + g^{C} + g^{stat} + g^{stat-h}$

Step 3-The extracted feature G has both the relevant as well as irrelevant features. While using both these features, the computational complexity of the model will increase. Therefore, in optimal features are isolated, via HCNAE-a new hybrid optimization model. The projected HCNAE is the conceptual blends of the standard AO [26] and HGS [27], respectively. The selected optimal features are represented using the notation G^{opt}

Step 4- Ultimately, the detection is accomplished via a new hybrid deep learning model. The projected hybrid deep learning model is the conceptual amalgamation of the standard LSTM and MSE-CNN, respectively. These two classifiers will be trained via G^{opt} . The outcome from LSTM and MSE-CNN is out_{LSTM} and $out_{MSE-CNN}$, respectively. The ultimate outcome (regarding the BC)is the score level fusion of the outcomes acquired from LSTM out_{LSTM} and CNN $out_{MSE-CNN}$, respectively.

2.3. *Data Pre-Processing via Median-based Z-score normalization*

Pre-processing is the most important phase since it helps to prepare the data in the most useful way. Because data preprocessing can have a significant impact on the learning model's outcome, it's critical that all features are on the same scale. In such algorithms, normalization is crucial. The Z-score nomination enhances the model's numerical stability while also cutting down on training time. The normalization of Z-scores isn't very good. Outliners are not as well protected by data normalization. As a result, this study introduces a median-based Z-score normalization-based pre-processing model. This stage is diagrammatically shown in Fig.2.

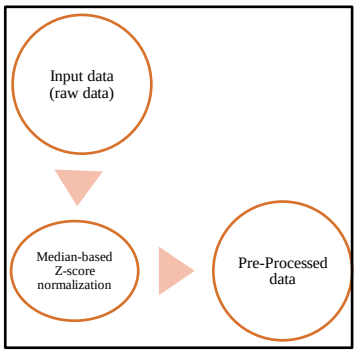


Fig. 2. Pre-processing Stage

Median-based Z-score normalization: This is based on the input data's median (Med) (proposed-median is determined rather than mean in conventional Z-score) and standard deviation σ. The use of the median value instead of the mean will aid in the translation of raw data into something meaningful. When there are multiple attributes but in various scales, this can lead to terrible data models. When using median based z score normalization, all of the attributes are brought to the same scale. The Median-based Z-score normalization may be expressed mathematically as Eq.(1) to Eq. (3), respectively.

$$D_i^{pre} = \frac{D_i^{inp} - Med}{\sigma} \qquad (1)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} D_i^{inp} (D_i^{inp} - \overline{D_i^{inp}}}{n}} \qquad (2)$$

Here, $\overline{D_i^{inp}}$ is the mean of D_i^{inp}.

$$\overline{D_i^{inp}} = \left(\prod_{i=1}^{n} D_i^{inp} \right)^{1/n} \qquad (3)$$

The pre-processed data acquired as a resultant is denoted as D_i^{pre}.

3. Feature Extraction and Feature Selection

3.1. *Feature Extraction*

The features like “Tanh Weighted Entropy (TWE), mutual information, correlation based features, and statistical features (mean median and standard deviation), and improved higher order statistical features (skewness, kurtosis and variance)” are extracted from D_i^{pre} This stage is diagrammatically shown in Fig.3.

Tanh weighted entropy (TWE) (proposed): the entropy based features provides information regarding the variables in the database. In order to acquire a much more reliable information regarding the variables, and its variations (i.e. between diseased and non-diseased), here is a requirement forthe variables, and its variations (i.e. between diseased and non-diseased), here is a requirement foran improved entropy based feature extraction. As a resultant, a new TWE has been introduced in this research work. Mathematically, this can be given as per Eq. (4) to Eq. (6), respectively.

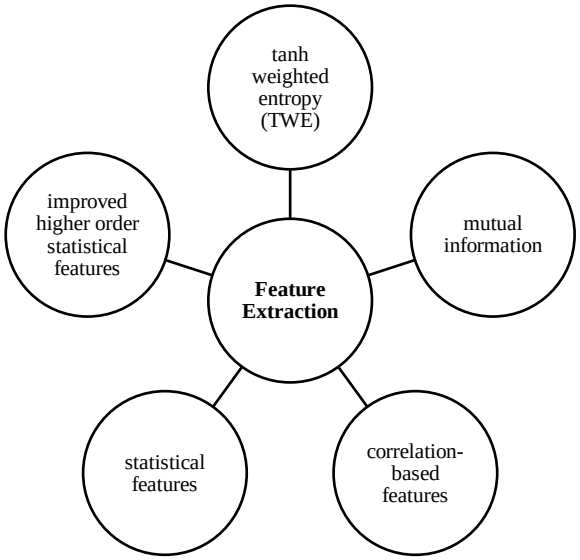


Fig. 3. Feature Extraction Stage

$$g^{IE}(D_i^{pre}) = W * E(D_i^{pre}) \quad (4)$$

Here, W is the weight function that is computed using the tanh function, and $E(D_i^{pre})$ points to the standard entropy function.

$$W = \left[\frac{2}{1 + e^{(-2(D_i^{pre}))}} \right]^{-1} \quad (5)$$

$$E(D_i^{pre}) = -\sum_{i=1}^{-\mu_i(D_i^{pre})} prob_i . \log(prob_i) \quad (6)$$

Here, $\mu_i(D_i^{pre})$ is the unique value of the feature attribute. The extracted features is pointed as g^{IE}.

Mutual Information: Mutual information is indeed a measurement of the decrease in uncertainty for one variable when another variable's value is known.The mutual information of two variables A, B within D_i^{pre} can be computed as per Eq. (7).

$$g^{MI}(A;B) = H(A) - H(A|B) \quad (7)$$

Here, $f^{MI}(A;B)$ is the mutual information of A,B. $H(A)$ is the entropy of A and $H(A|B)$ is the conditional entropy of A given B. The extracted features arerepresentedas g^{MI}.

Correlation: It provides information regarding the correlation between the variables (used to analyze relevance). This assists in precise detection outcome and also aids in minimizing the errors that takes place during the detection process. The correlation between variables a,b of D_i^{pre}. The extracted features arerepresentedas g^{C}.

Statistical Features: The statistical features like "mean, median and standard deviation" is extracted from D_i^{pre}. The extracted feature arerepresented as g^{stat}.

Higher order statistical features: The features like improved skewness (proposed) and kurtosis are extracted from D_i^{pre}. The extracted feature arerepresented as g^{stat-h}.

These retrieved features are fused together, and represented as $G = g^{IE} + g^{MI} + g^{C} + g^{stat} + g^{stat-h}$.

The extracted feature-set G has both the relevant as well as irrelevant features. The optimal features are isolated, via HCNAE that is discussed in next section.

3.2. *Feature Selection*

From G, the most optimal features are selected via a new optimization technique, HCNAE. The projected hybrid optimization model is the conceptual blend of the standard AO [26] and HGS [27], respectively. AO has been developed based on the Aquila's prey hunting capability. HGS is based on the social behavior of animals in hunting their prey.Since the AO and HGS are both effective at solving difficult optimization problems with high convergence, they have been combined to create a novel hybrid optimization model. The HGS model is implied with AO's solution updating phase. The input to the projected model is the extracted features G. The phases of the projected hybrid optimization model are: "Solutions initialization, expanded exploration, narrowed exploration, expanded exploitation and Narrowed exploitation", respectively. The steps followed in the projected hybrid optimization model are demonstrated below:

Phase 1: **Solution initialization**:The population of candidate solution population solutions G (given in Eq. (8) and Eq. (9)) are initialized. These are generated based on the problem's lower bound (LB) as well as upper (UB) limits.

$$G=\begin{bmatrix} g_{1,1} & \cdots & g_{1,q} & g_{1,Dim-1} & g_{1,Dim} \\ g_{2,1} & \cdots & g_{2,q} & \cdots\cdots & g_{2,Dim} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ g_{n-1,1} & \cdots & g_{n-1,q} & \cdots & g_{n-1,Dim} \\ g_{n} & \cdots & g_{n,q} & g_{n,Dim-1} & g_{n,Dim} \end{bmatrix} \quad (8)$$

Here, g_i symbolises the positions of i^{th} solution and n signifies the total number of candidate solutions, & Dim relates to the problem's dimension size.

$$G_{iq}=rand\times\left(UB_q-LB_q\right)+LB_q \qquad i=1,2,\ldots,n; q=1,2,\ldots,Dim \quad (9)$$

Phase 2: Extensive exploration (G_1)

In this phase (G_1), the Aquila uses a greater soar with a vertical stoop to determine the better hunting location and identify the prey area (row of optimal features). This behaviour is given in Eq. (10).

$$G_1(itr+1)=G_{best}(itr)\times\left(1-\frac{itr}{\max^{itr}}\right)+\left(G_{mean}(itr)-G_{best}(itr)*r1\right) \quad (10)$$

Here, $G_1(itr+1)$ points to the solution at the subsequent iteration of itr (current iteration); $G_{best}(itr)$ best-obtainedsolution at itr and $G_{mean}(itr)$ is the mean value corresponding to the current location of the search agent (given in Eq. (11)). Through the iterations, the expanded search is controlled by $\left(\frac{itr}{\max^{itr}}\right)$ and $r1$ is a random number generated between [0, 1]. In addition, itr and $\max^{itr}$ points to the current and maximal iteration counts, respectively.

$$G_{mean}(itr)=\frac{1}{n}\sum_{i=1}^{n}G_i(itr); \quad \forall q=1,2,\ldots,Dim \quad (11)$$

Phase 3: Proposed Narrowed exploration (G_2)

When an optimal feature (targeted prey) is identified by the search agent, from a high soar; it prepares the land and then attacks the prey. The foraging behavior based on HGS (HGS is implied within AO) is mathematically given in Eq. (12).

$$G(itr+1)=\begin{cases} G(itr).(1+randn(1)) & rand1<l & Game1 \\ W1.G_{best}+R.W2.\left|G_{best}-G(itr)\right| & rand1>l; rand2>E & Game2 \\ W1.G_{best}-R.W2.\left|G_{best}-G(itr)\right| & rand1>l; rand2<E & Game3 \end{cases} \quad (12)$$

Here, $rand1, rand2$ is a random number generated between [0, 1]. $randn(1)$ is a random number that satisfies the normal distribution. The value of $W1$ in hunger role is computed using Eq. (13).

$$W1(i)=\begin{cases}1 & if\ \ rand4>1\\ hungry(i).\dfrac{N}{shungry}(rand4) & if\ rand4<l\end{cases}\quad(13)$$

Here, $rand4$ is a random number generated between [0,1]. The value of $W2$ in hunger role is computed using Eq. (14).

$$W2(i)=\left[1-\exp(-\left|hungry(i)-shungry(i)\right|)\right]*rand5*2 \quad (14)$$

Here, $rand5$ is a random number generated between [0, 1]. The variation control for all positions E using the newly projected expression is given in Eq. (15).

$$E=\tanh\left[\left|obj(i)-B\right|\right] \quad (15)$$

Here, $obj(i)$ is the fitness of the search agent and B is the best fitness acquired so far. R is updated using Eq. (16) and Eq. (17), respectively.

$$R=2*shrink*rand-shrink \quad (16)$$

$$shrink=2*\left(1-\frac{itr}{\max^{itr}}\right) \quad (17)$$

Phase 4: Expanded exploitation (G_3)

Once, the prey are (optimal feature's location) is identified, the search agent is ready to land and attack it. This behaviour is mathematically given in Eq. (18).

$$\begin{aligned}G_3(itr+1)=&\left(G_{best}(itr)-G_{mean}(itr)\right)\times\alpha-rand3\\&+\left(\left(UB-LB\right)\times rand4+LB\right)\times\mu\end{aligned} \quad (18)$$

In Eq. (19), $G_3(itr+1)$ symbolises the solution of the 3rd search technique's (G_3) in next iteration, $G_{best}(itr)$ the estimated position of the prey till i^{th} iteration, $rand3, rand4$ is randomly generated between [0, 1], α and μ are the exploitation modification variables reduced to (0.1). LB denotes the lower bound & UB specifies upper bound of the problem

Phase 5: Proposed Narrowed exploitation (G_4)

The prey is attacked (i.e. optimal solution is identified) in this phase (G_4), this mechanism is shown mathematically in Eq. (19).

$$\begin{aligned}G_4(itr+1)=&QF\times G_{best}(itr)-\left(M_1\times Z(itr)\times rand5\right)\\&-M_2\times Levy(\beta)+rand6\times M_1\end{aligned} \quad (19)$$

Here, $rand5$, $rand6, rand7$ are the random numbers generated between [0,1]. QF is a quality function that is used to maintain the equilibrium of this phase (G_4). Therefore, QF is computed using a new expression given in Eq. (20) to Eq. (22). Here, $\sigma = rand7 * 10; \sigma \in (1,10)$.

$$QF(itr = itr^{\frac{2\times ran-1}{\left(1-itr/\max^{itr}\right)^{(1/\sigma)}}} \quad (20)$$

$$M_1 = 2\times ran - 1 \quad (21)$$

$$M_2 = 2\times\left(1-\frac{itr}{\max^{itr}}\right) \quad (22)$$

The selected optimal features are represented using the notation G^{opt}

4. BC detection via hybrid deep learning model

Ultimately, the final classification step is accomplished via a new hybrid deep learning model. The projected hybrid deep learning model is the hybrid of the standard LSTM and MSE-CNN. These two classifiers are trained using the optimal feature-set G^{opt}. The outcome from LSTM and MSE-CNN is out_{LSTM} and $out_{MSE-CNN}$, respectively. The ultimate outcome (regarding the BC) is the score level fusion of the outcomes acquired from LSTM out_{LSTM} and CNN $out_{MSE-CNN}$, respectively. The major objective of this research work is to minimize the classification errors c_{error}. Therefore, MSE-CNN and LSTM has been hybridized for highly accurate classifier.

4.1. *MSE- CNN*

A convolutional neural network [29] is a neuron-based hierarchical structure. A function with input G^{opt} and output out is used to represent a single neuron. This is mathematically shown in Eq. (23) and Eq. (24), respectively.

$$out^{(l)} = act(H^{opt(l-1)} * \omega^{(l-1)} + Z^{(l-1)}) \quad (23)$$

$$act(H^{opt}) = \frac{1}{(1+e^{-H^{opt}})} \quad (24)$$

The weight function is ω, the scalar bias function is Z, and the sigmoid activation function is $act(.)$ (proposed). The training procedure involves experimenting with a range of constraints (such as weights and biases) in order for the neural network to match the connection among outputs and inputs. The backpropagation algorithm is indeed a popular learning model wherein a loss function is built using the improved mean square error (MSE) (proposed-MSE is used instead of entropy function in standard CNN). The newly formulated loss function is shown in Eq. (25) respectively.

$$L = \frac{1}{n} \frac{\sum_{e=0}^{n} \omega_{(e)} (pre_{(e)} - obs_{(e)})^2}{\sum_{e=0}^{n} \omega_{(e)}} \quad (25)$$

$$\omega_{(e)} = \log(obs_{(e)} + 1) rand \quad (26)$$

Here, $rand$ is a random value that is computed using the sinusoidal map.

4.2. *LSTM*

LSTM [28], which is a modified Recurrent Neural Network approach, operates well on a wide range of situations and therefore is currently frequently exploited. By including gate units and memory cells inside the neural network design, LSTM addresses the problem of finding out where to remember input over time. Cell states in memory cells store info that has been previously witnessed. When information enters a memory cell, the result is determined by a mixture of cell states, as well as the cell state is then updated. The LSTM has cell consists of three gates: “update, output and forget”. The input that needs to be processed is updated by the update gate. The forget gate stores the necessary information in the memory. The output gate determines what will be output. Mathematically, LSTM can be given as per Eq. (27) -Eq. (31), respectively.

$$\Gamma_u = \sigma(W_u[j^{<t-1>}, H^t] + bias_u) \quad (27)$$

$$\Gamma_f = \sigma(W_f[j^{<t-1>}, H^t] + bias_f) \quad (28)$$

$$\breve{c}^t = \tanh(W_c[j^{<t-1>}, H^t] + bias_c) \quad (29)$$

$$C^t = \Gamma_f * C^{<t-1>} + i^t * C^t \quad (30)$$

$$\Gamma_o = \sigma(W_o[j^{<t-1>}, H^t] + bias_o) \quad (31)$$

Here, Γ_f , Γ_u and Γ_o points to the forget gate, update gate and output gate, respectively. σ points to the sigmoid function. in addition, W_u , W_f and W_c represents the weight corresponding to the update gate, forget gate and output gate, respectively. $j^{<t-1>}$ is the input of the previous state , C^t is the new cell state and $C^{<t-1>}$ is the previous state of the cell. The outcome from LSTM is out_{LSTM} .The final result is a score level fusion of the results from the LSTM and CNN, respectively.

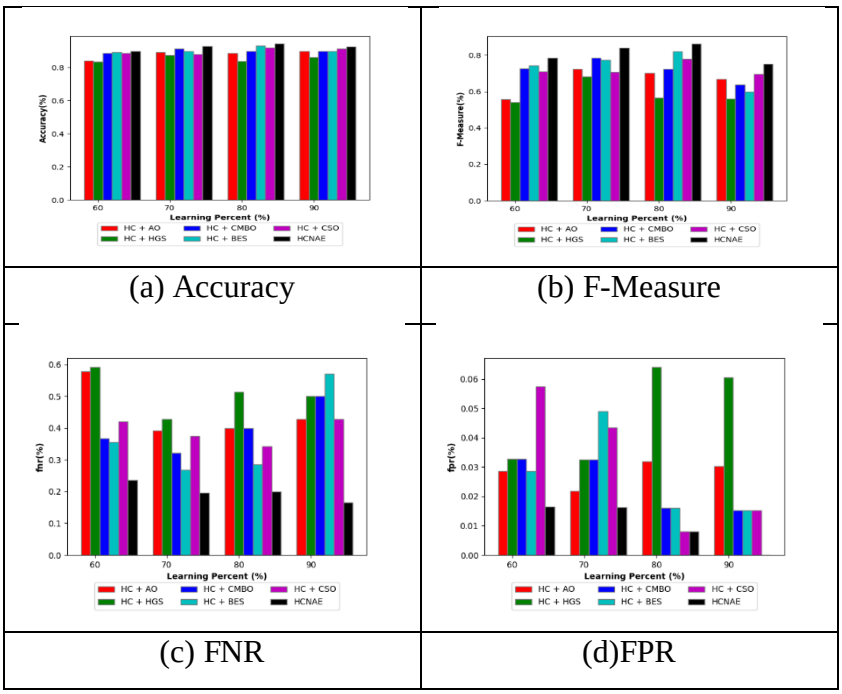
5. Result and Discussion

Python was used to implement the proposed BC classifier. Seventy percent of the data was set for training, while the remaining thirty percent was used for testing. The proposed model was evaluated in terms of "accuracy, F-measure, FNR, FPR, MCC, NPV, Precision, Recall, Sensitivity, and

Specificity". The positive measures of "accuracy, sensitivity, specificity, and precision" should be high, whilst the error measures of "FPR and FNR" should be kept as low as possible.

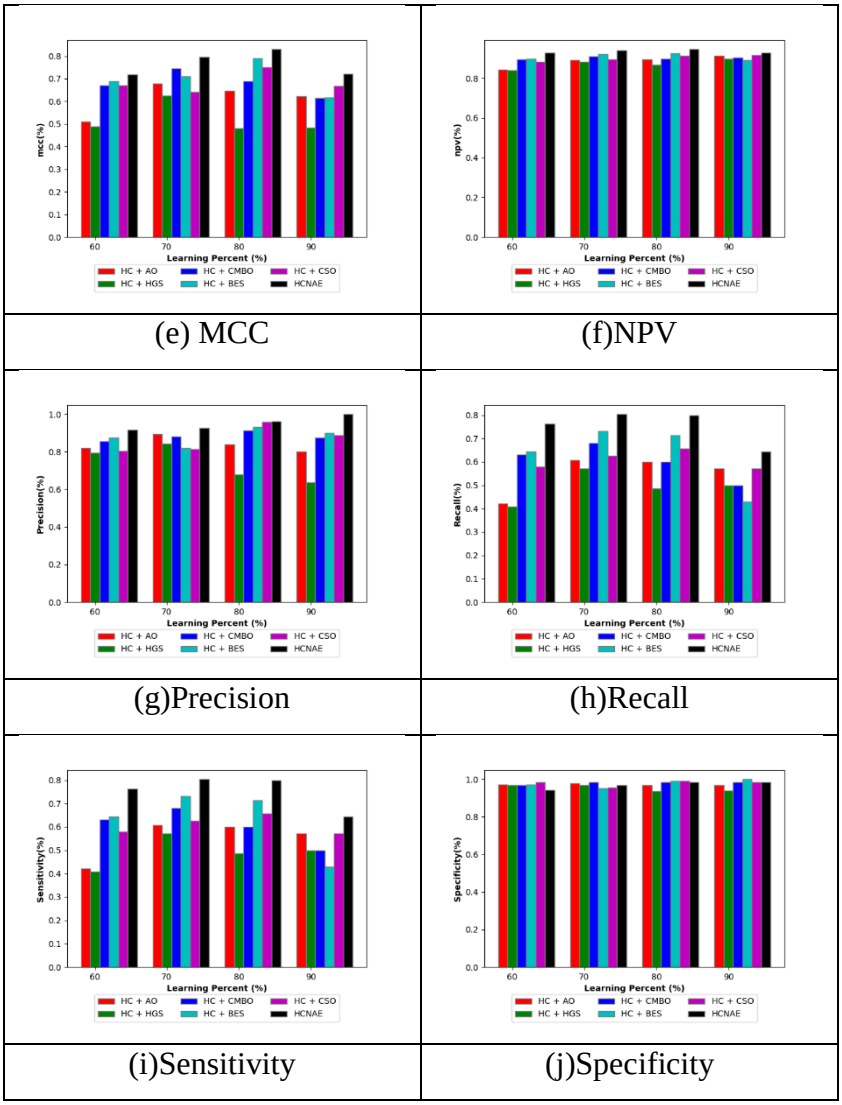
5.1. *Performance Analysis*

The proposed model (HCNAE+ Hybrid deep learning classifier) is compared to current models such as AO+HC, HGS+HC, CMBO+HC, BES+HC, CSO+HC, and so on. The predicted model records a notable performance over every change in the learning rate when assessing the obtained results. The maximum accuracy was reported by the proposed model. The hybridization of two classifiers for classification purposes is the main cause for the improvement in prediction accuracy. The predicted model achieves the greatest accuracy of 97.5 percent when training the model with 70% learning. The proposed model has a greater sensitivity, specificity, and accuracy than existing models. Furthermore, when compared to traditional models, the projected model has the greatest F-measure. The projected model has the greatest F-measure (miscellaneous measure) of 0.985 at learning rate of 0.8, which is higher than AO+HC=0.75, HGS+HC=0.58, CMBO+HC=0.78, BES+HC=0.83 and CSO+HC=0.82 Furthermore, for any modification in learning percent, the projected model's NPV and MCC are greater than the conventional models. The projected model, on the other hand, has the fewest misclassifications (i.e., lowest FPR and FNR).The estimated model's FNR is lower than that of existing models. For any modification in the learning percent, the estimated model's FNR is less than 20%. The anticipated model, on the other hand, shows FNR values exceeding 35%. The selection of optimum features for training the deep learning classifiers employed in the detection phase is the main cause for the decrease in FNR. The proposed model is claimed to be very important for BC subtype classification as a whole.



(a) Accuracy (b) F-Measure

(c) FNR (d)FPR



(e) MCC (f)NPV

(g)Precision (h)Recall

(i)Sensitivity (j)Specificity

Fig. 4. Performance analysis of the projected Model over existing models

5.2. *Convergence Analysis*

The BC detection model is considered as an optimization problem, and it solved using a new HCNAE model. The HCNAE is formulated by blending HGS and AO, respectively. The HCNAE model has been theoretically said to be higher convergent, due to the consideration of the tanh (hyperbolic tangent) as well as proposed Narrowed exploitation and proposed Narrowed exploration. The objective function or fitness function of this research work is minimization of the detection accuracy. So, the approach that records the least cost function or fitness is said to be the highly convergent one. As per the recorded outcomes, the projected model has recorded the least cost function than the existing models. Initially, the cost function of the existing as well as projected model is higher, and as the iteration count increases there is indeed a steep decrease in cost function over increase in iteration count. The cost function recorded by HCNAE is much lower than HGS alone and AO alone, respectively. Thus, by hybridization, there is increase in convergence speed.

At 50th iteration (highest value), the projected model has recorded the least cost function as 1.088. Thus, HCNAE is said to be highly convergent over the existing models.

5.3. *Analysis on Feature Selection Performance*

The features selection plays a vital role in reducing the computational complexity of the model. The results acquired by the projected BC detection framework in terms of feature selection are shown in Table 2. The projected framework could achieve 78.3%, when no feature selection is applied. In addition, with Principle Component Analysis (PCA) and Linear discriminant analysis (LDA) based feature selection, the projected model has recorded the accuracy as 67.5% and 80%. When, a new optimization model has been used for selecting the optimal features, the projected model recorded 92.9% accuracy. In addition, the errors (FPR and FNR) have also been reduced with optimal feature selection. Thus, the projected model is said to be much applicable for enhancing the BCsubtype classification.

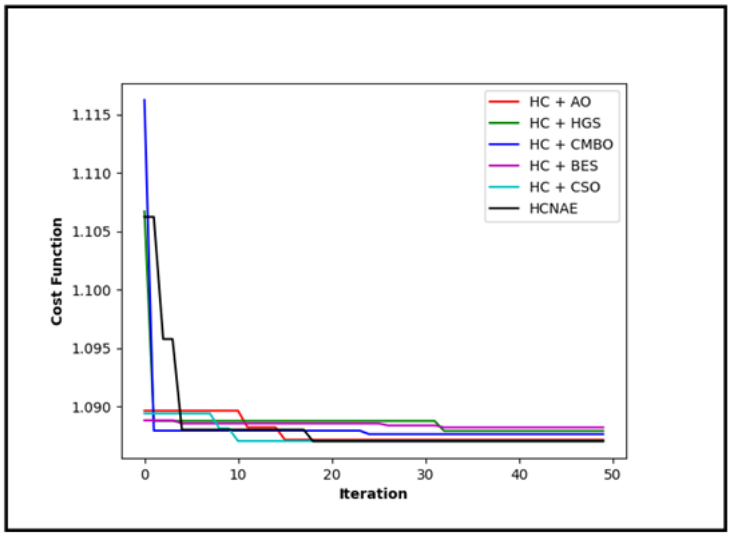


Fig. 5. Convergence Analysis of HCNAE

5.4. *Classifier Performance Analysis*

The classifier is the ultimate decision maker in this research work. The projected HC is compared over the existing classifiers, and the results acquired are shown in Table 3. The proposed hybrid deep learning model (HC) combines the standard Long Short Term Memory (LSTM) and Mean Square Error Based (MSE) based Convolutional Neural Network (MSE-CNN) models for BC detection in this research work. The accuracy recorded by the projected model with HC is 92.9%, which is indeed a best score compared to Bidirectional Gated Recurrent Unit (BI-GRU) = 0.8625; Classification- Deep Belief Network (DBN)= 0.779167; Recurrent Neural Network (RNN)= 0.729167; Support Vector Machine (SVM) [13] = 0.85 and Logistic Regression (LR) [6] =0.875.

The MSE-CNN has highly contributed in enhancing the detection performance. Moreover, HC has recorded the highest specificity as 96.7%. In addition, hybrid deep learning model has exhibited least error values (FPR and FNR), and this is due to the improvement made within CNN in terms of loss function.

Table 2. Analysis on the projected BC Detection Framework in terms of Feature Selection Performance

Measures	Proposed work with score based data normalization + proposed multiple feature + optimized feature selection and hybrid classifier	Proposed work without optimization	Proposed work with PCA based feature selection	Proposed work with LDA based feature selection
Specificity (%)	0.967391	0.755435	0.798913	0.73913
Sensitivity (%)	0.803571	0.875	0.267857	0.783333
Recall (%)	0.803571	0.875	0.277857	0.538462
Precision (%)	0.882353	0.521277	0.288462	0.538462
NPV (%)	0.941799	0.952055	0.781915	0.781915
MCC (%)	0.797114	0.546264	0.0685496	0.630867
FPR (%)	0.0326087	0.244565	0.201087	0.26087
FNR (%)	0.196429	0.125	0.732143	0.532143
F-Measure (%)	0.841121	0.653333	0.277778	0.7
Accuracy (%)	0.929167	0.783333	0.675	0.8

Table 3. Analysis on the projected BC Detection Framework in terms of Classifiers

Measures	Hybrid deep learning model (HC-MSE-CNN+LSTM)	BI-GRU	DBN	RNN	SVM[13]	LR[6]
Specificity (%)	0.967391	0.913043	0.875	0.782609	0.891304	0.923913

Sensitivity (%)	0.803571	0.696429	0.464286	0.553571	0.714286	0.714286
Recall (%)	0.803571	0.696429	0.464286	0.553571	0.714286	0.714286
Precision (%)	0.882353	0.709091	0.530612	0.43662	0.666667	0.740741
NPV (%)	0.941799	0.908108	0.842932	0.852071	0.911111	0.913978
MCC (%)	0.797114	0.613323	0.356003	0.311532	0.59152	0.646406
FPR (%)	0.0326087	0.0869565	0.125	0.217391	0.108696	0.076087
FNR (%)	0.196429	0.303571	0.535714	0.446429	0.285714	0.285714
F-Measure (%)	0.841121	0.702703	0.495238	0.488189	0.689655	0.727273
Accuracy (%)	0.929167	0.8625	0.779167	0.729167	0.85	0.875

6. Conclusion

A unique BC diagnostic technique based on a hybrid deep learning model from multi-omics data has been proposed in this research. This model is the need of the hour since source-based classification can assist in tailored therapy. The most accurate model is the one with optimized feature selection methods that is a hybrid of AO and HCN optimization techniques. This model may be a part of AI system that can help clinicians with more accurate treatment strategies along with other models that aid in survival analysis and drug response prediction system. For this work, we initially considered, the multiomics data that was pre-processed using a Median-based Z-score normalization approach (proposed). Following that, features such as "Tanh Weighted Entropy (TWE)-proposed, mutual information, correlation-based features, and statistical features (mean, median, and standard deviation), as well as improved higher order statistical features (improved skewness (proposed), kurtosis, and variance)" were extracted. The HCNAE algorithm was then used to choose the best features from the recovered features. Finally, the source based with BC classification is carried out using a unique hybrid deep learning model. The proposed hybrid deep learning model combines the MSE-CNN and LSTM algorithms. The final result will is a score level fusion of the results from the LSTM and MSE-CNN. Finally, a comparative evaluation is performed to ensure that the proposed model is effective in detecting BC. The highest classification accuracy of 97.5% is recorded by the projected model.

Bibliography

There are no sources in the current document.

[1] Ming-Jun Shi, Xiang-Yu Meng, Jacqueline Fontugne, Chun-Long Chen, François Radvanyi & Isabelle Bernard-Pierrot , "Identification of new driver and passenger mutations within APOBEC-induced hotspot mutations in bladder cancer", *Genome Medicine*, 2020

[2] Daniele Raimondi, Antoine Passemiers, Piero Fariselli & Yves Moreau, "Current cancer driver variant predictors learn to recognize driver genes instead of functional variants", *BMC Biology*, 2021

[3] J. Xing, Y. Fang, W. Zhang, H. Zhang, D. Tang & D. Wang , "Bacterial driver–passenger model in biofilms: a new mechanism in the development of colorectal cancer", *Clinical and Translational Oncology*, 2022

[4] Seyed Mohammad Razavi, Farzaneh Rami, Seyede Houri Razavi & Changiz Eslahchi, "TOPDRIVER: the novel identifier of cancer driver genes in Gastric cancer and Melanoma", *Applied Network Science*, 2019

[5] Leila Mirsadeghi, Reza Haji Hosseini, Ali Mohammad Banaei-Moghaddam & Kaveh Kavousi , "EARN: an ensemble machine learning algorithm to predict driver genes in metastatic BC", *BMC Medical Genomics*, 2021

[6] Junrong Song, Wei Peng, Feng Wang & Jianxin Wang , "Identifying driver genes involving gene dysregulated expression, tissue-specific expression and gene-gene network", *BMC Medical Genomics*, Vol.12, 2019

[7] Junrong Song, Wei Peng & Feng Wang , "A random walk-based method to identify driver genes by integrating the subcellular localization and variation frequency into bipartite graph", *BMC Bioinformatics volume*, 2019

[8] Yahya Bokhari, Areej Alhareeri & Tomasz Arodz , "QuaDMutNetEx: a method for detecting cancer driver genes with low mutation frequency", *BMC Bioinformatics*, 2020

[9] Ege Ülgen & O. Uğur Sezerman , "driveR: a novel method for prioritizing cancer driver genes using somatic genomics data", *BMC Bioinformatics* , 2021

[10] Yun-Yun Tang, Pi-Jing Wei, Jian-ping Zhao, Junfeng Xia, Rui-Fen Cao & Chun-Hou Zheng , "Identification of driver genes based on gene mutational effects and network centrality", *BMC Bioinformatics*, Vol.457, 2021

[11] P. A. A. Iloshini, M. W. A. C. R. Wijesinghe, T. Kartheeswaran and W. M. P. S. Weerasooriya, "Dots Witer: Prediction of Potential Cancer Driver Genes Using Hybrid Approach," *2019 International Conference on Advancements in Computing (ICAC)*, Malabe, Sri Lanka, 2019, pp. 163-167. doi: 10.1109/ICAC49085.2019.9103401

[12] J. Song, W. Peng and F. Wang, "Identifying cancer patient subgroups by finding co-modules from the driver mutation profiles and downstream gene expression profiles," in *IEEE/ACM*

Transactions on Computational Biology and Bioinformatics. doi: 10.1109/TCBB.2021.3106344

[13] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

[14] J. Song, W. Peng and F. Wang, "An Entropy-Based Method for Identifying Mutual Exclusive Driver Genes in Cancer," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 3, pp. 758-768, 1 May-June 2020. doi: 10.1109/TCBB.2019.2897931

[15] J. Xie *et al*., "Prediction of Essential Genes in Comparison States Using Machine Learning," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 5, pp. 1784-1792, 1 Sept.-Oct. 2021. doi: 10.1109/TCBB.2020.3027392

[16] X. Lu, X. Wang, L. Ding, J. Li, Y. Gao and K. He, "frDriver: A Functional Region Driver Identification for Protein Sequence," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 5, pp. 1773-1783, 1 Sept.-Oct. 2021. doi: 10.1109/TCBB.2020.3020096

[17] P. Wang, D. Wang and J. Lü, "Controllability Analysis of a Gene Network for Arabidopsis thaliana Reveals Characteristics of Functional Gene Families," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 3, pp. 912-924, 1 May-June 2019. doi: 10.1109/TCBB.2018.2821145

[18] C. Liu, Y. Dai, K. Yu and Z. K. Zhang, "Enhancing cancer driver gene prediction by protein-protein interaction network," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
doi: 10.1109/TCBB.2021.3063532

[19] F. Li, L. Gao and B. Wang, "Detection of Driver Modules with Rarely Mutated Genes in Cancers," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 2, pp. 390-401, 1 March-April 2020. doi: 10.1109/TCBB.2018.2846262

[20] Y. Li, F. Zhang and C. Xing, "Screening of Pathogenic Genes for Colorectal Cancer and Deep Learning in the Diagnosis of Colorectal Cancer," in *IEEE Access*, vol. 8, pp. 114916-114929, 2020.
doi: 10.1109/ACCESS.2020.3003999

[21] J. Zhang and S. Zhang, "The Discovery of Mutated Driver Pathways in Cancer: Models and Algorithms," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 3, pp. 988-998, 1 May-June 2018. doi: 10.1109/TCBB.2016.2640963

[22] T. Tamura, T. Akutsu, C. -Y. Lin and J. -M. Yang, "Finding Influential Genes Using Gene Expression Data and Boolean Models of Metabolic Networks," *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, Taichung, Taiwan, 2016, pp. 57-63. doi: 10.1109/BIBE.2016.25

[23] L. D. Mora, O. Azofeifa, D. Diaz and J. A. Guevara-Coto, "Machine learning approaches for the identification of new driver-like genes using microarray expression profiles," *2019 IV Jornadas Costarricenses de Investigación en Computación e Informática (JoCICI)*, San Pedro, Costa Rica, 2019, pp. 1-6. doi: 10.1109/JoCICI48395.2019.9105274

[24] Yang, Y., Chen, H., Heidari, A. A., & Gandomi, A. H, "Hunger games search:Visions, conception, implementation, deep analysis, perspectives, and towards performance shifts", *Expert Systems with Applications*, Vol.155, 2021

[25] Liu Q et al Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data. *J Biomed Inform*. 2022 Jan;125:103958. doi: 10.1016/j.jbi.2021.103958.

[26] Chou, J.S. and Nguyen, N.M, "FBI inspired meta-optimization", *Applied Soft Computing*, 2020

[27] M. A. Istiake Sunny, M. M. S. Maswood and A. G. Alharbi, "Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model," 2020 *2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 2020, pp. 87-92, doi: 10.1109/NILES50944.2020.9257950.

[28] L. Nie, Z. Ning, X. Wang, X. Hu, J. Cheng and Y. Li, "Data-Driven Intrusion Detection for Intelligent Internet of Vehicles: A Deep Convolutional Neural Network-Based Method," in *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2219-2230, 1 Oct.-Dec. 2020, doi: 10.1109/TNSE.2020.2990984.

[29] Giovanni Ciriello et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, 2015, Volume 163, ISSUE 2, P506-519, October 08, 201, https://doi.org/10.1016/j.cell.2015.09.033

[30] B. E. Johnson et al., "An omic and multidimensional spatial atlas from serial biopsies of an evolving metastatic breast cancer," *Cell Reports Med.*, vol. 3, no. 2, p. 100525, 2022, doi: https://doi.org/10.1016/j.xcrm.2022.100525.

[31] Y.-X. Chen et al., "An integrative multi-omics network-based approach identifies key regulators for breast cancer," *Comput. Struct. Biotechnol*. J., vol. 18, pp. 2826–2835, Oct. 2020, doi: 10.1016/j.csbj.2020.10.001.

[32] H. Cui et al., "Inferences of Individual Drug Response-Related Long Non-coding RNAs Based on Integrating Multi-omics Data in Breast Cancer," *Mol. Ther. Nucleic Acids*, vol. 20, pp. 128–139, Jun. 2020, doi: 10.1016/j.omtn.2020.01.038.

[33] F. Rahman et al., "A multi-omics approach to reveal the key evidence of GDF10 as a novel therapeutic biomarker for breast cancer," *Informatics Med. Unlocked*, vol. 21, p. 100463, 2020, doi: https://doi.org/10.1016/j.imu.2020.100463.

[34] Dataset:https://www.kaggle.com/samdemharter/multi-omics-integration-with-the-qlattice/data?select=data.csv

[35] V. Malik, Y. Kalakoti, and D. Sundar, "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer," *BMC Genomics*, vol. 22, no. 1, p. 214, 2021, doi: 10.1186/s12864-021-07524-2.

[36] D. Khan and S. Shedole, "Leveraging Deep Learning Techniques and Integrated Omics Data for Tailored Treatment of Breast Cancer.," *J. Pers. Med.*, vol. 12, no. 5, Apr. 2022, doi: 10.3390/jpm12050674.