IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 October 2022, accepted 30 December 2022, date of publication 5 January 2023, date of current version 10 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234294

RESEARCH ARTICLE

Efficient Bioinspired Feature Selection and Machine Learning Based Framework Using Omics Data and Biological Knowledge Data Bases in Cancer Clinical Endpoint Prediction

IMENE ZENBOUT[1,2], ABDELKRIM BOURAMOUL[1], SOUHAM MESHOUL[3], AND MOUNIRA AMRANE[4]

[1]MISC Laboratory, Department of Fundamental Informatics and Its Application, Constantine 2 University Abdelhamid Mehri, Constantine 25016, Algeria
[2]National Biotechnology Research Center (CRBT), Constantine 25000, Algeria
[3]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[4]Biochemistry Department, Setif University Hospital, University Ferhat Abbas Setif 1, Setif 19000, Algeria

Corresponding authors: Imene Zenbout (imene.zenbout@univ-constantine2.dz) and Souham Meshoul (sbmeshoul@pnu.edu.sa)

This work was supported by the Princess Nourah Bint Abdulrahman University Researchers Supporting Project through Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, under Grant PNURSP2023R196.

ABSTRACT Cancer Research has advanced during the past few years. Using high throughput technology and advances in artificial intelligence, it is now possible to improve cancer diagnosis and targeted therapy, by integrating the investigation and analysis of clinical and omics profiles. The high dimensionality and class imbalance of the majority of available data sets represent a serious challenge to the development of computational methods and tools for cancer diagnosis and biomarker discovery. Taking into account multi-omics data further complicates the undertaking. In this paper, we describe a five-step integrative architecture for dealing with the three aforementioned problems by incorporating proteomics data, protein-protein interaction networks, and signaling pathways in order to identify protein biomarkers with a direct association to cancerous patients' overall survival (OS) and progression free interval (PFI). The core parts of this architecture are a cluster based grey wolf optimization algorithm (CB-GWO) for feature selection and a deep stacked canonical correlation autoencoder (DSCC-AE) for clinical endpoint prediction. A thorough experimental study was carried out to evaluate the performance of the proposed optimization algorithm for feature selection, as well as the performance of the deep learning model in terms of Mathew coefficient correlation (MCC) and Area under the curve (AUC) on breast, lung, colon, and rectum cancers. The results were compared to other methods in the literature. The results are very promising and show the effectiveness of the proposed framework and its ability to outperform the other algorithms and models in terms of AUC (0.91) and MCC (0.64). In addition, hub marker genes with the potential occurence of alterations in colorectal cancer, breast cancer, and lung cancer have been identified.

INDEX TERMS Biomarker discovery, Integrative omics, cancer classification, deep canonical correlation analysis, enrichment analysis, feature selection, grey wolf optimization, machine learning.

I. INTRODUCTION

Over the past few years, the precision of cancer diagnosis has increased. High throughput sequencing and screening technologies and artificial intelligence have been of great assistance in enhancing and improving the protocols used for diagnosis, prognosis, and treatment; consequently, the medical community is gradually migrating towards precision medicine by means of integrative exploration and analysis of clinical and omics profiles [1], [2]. Despite the success

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

of precision oncology, the variation of cancer symptoms and the unplanned and unpredictable events recorded along the constant evolution of the tumor make cancer patient diagnosis and monitoring more difficult [3]. In addition, the data scalability of genomics profiles requires consistent in-silico methods to define an accurate set of biomarkers that can be used as diagnostic and prognostic biomarkers and aid in the medical decision making [4], in addition to mining the selected genes biomarkers for variants that can be effective therapeutic targets for each individual.

In the past few years, biomarker-discovery has gained a lot of attention as an emerging research field due to the advances of next generation sequencing technologies and novel high throughput technologies [5], [6], with the aim to provide cost-effective, time-effective, and better performance at different omics data levels (genomics, transcriptomics, proteomics, metabolomics, etc.). In the early years of the previous decade, precision medicine and bioinformatics data mining researches were primarily based on the genomics and transcriptomics profile single level omics data analysis [7], and the first in-silico biomarker tools were developed primarily for mining gene expression to identify the most relevant genes responsible for cancer-driving mutation [8], [9]. The biggest obstacle computational tools face when analyzing mRNA expression data is the curse of dimensionality in gene expression data sets, which are characterized by a large number of irrelevant features and small patient samples [10].

In recent years, the scientific community has begun to integrate proteomics data into oncology-related in-silico biomarker discovery and cancer classification in order to overcome the curse of mRNA expression data [11], [12], [13] since it is known that multicellular organisms share the same set of genes, even if the set of generated proteins varies between organisms, the set of produced proteins studied under specific conditions implies knowledge of the specifically synthesized proteins in the disease state. This has led to the development of proteomics next-generation sequencing, specifically Reverse Phase Protein Arrays (RPPA). The RPPA technology is thought to address the two major challenges in mRNAseq data by profiling the output (functional protein) of cancer-coding genes, which can reflect and capture the tumor state, as well as how to follow a pathologically effective therapy [14]. Furthermore, the RPPA technology operates with greater precision on a smaller set of proteins, limiting the protein expression matrix's scalability, while, we usually need to process hundreds to thousands of profiles per patient. Thereupon, we propose in this paper to start the process of identifying biomarkers at the proteome level, where we select the most relevant proteins and determine their coding genes. These initiatives will ensure that we only concentrate on mRNA instances that have been fully transcribed into mature proteins with potential biological processes. However, as stated in [15], one disadvantage of RPPA technologies is that some of the profiled proteins may be novel and unreferenced in databases, implying that they have no clinically meaningful application.

Therefore, signaling pathways and interaction networks are two types of data that can be used to identify highly clinically relevant cancer biomarkers in order to address this issue. The former describes the set of functional proteins and, by extension, the driver coding genes that have a potential approved biological positive or negative impact on essential biological processes, such as apoptosis, cell cycle, or cell death [16]. The latter is the omics-omics interaction network, which captures the correlation between the studied omics instances (protein-protein) based on their co-expression, physical interaction, or inference of biological process. When integrated with omics data expression, these two types of biological knowledge data bases can improve the biomarker discovery process by assisting in the identification of instances with a hub biological interaction and potential clinical impact, as well as by enhancing the features learning and clinical endpoints prediction in predictive tools. [17], [18].

In this paper, we describe a five-step integrative framework based on machine learning and grey wolf optimization algorithm, that incorporates proteomics data, protein-protein interaction networks, and signaling pathways in order to identify protein biomarkers with a direct association to cancerous patients' overall survival (OS) and progression free interval (PFI). The key contributions introduced in this paper are summarized as follows:

- Handling the high dimensionality of gene expression data using a a bioinspired optimization algorithm for feature selection to identify proteomic biomarkers (the gene' product) based on RPPA data sets instead of mRNA data, protein-protein interaction (PPI) network and signalling pathways.
- Identifying genes, and miRNA biomarkers based on the selected proteomic biomarkers and the analysis of mRNA-gene target network.
- Proposing a deep learning based integrative model that integrates the selected biomarkers in predicting cancer clinical endpoints.
- Conducting an intensive in-silico study to identify the most relevant omic biomarkers associated with OS and PFI, as well as a list of relevant variants that may be a potential research target for clinical application.

The remainder of the paper is structured in the following manner: Section II, presents a review of the recent related works. Section III, describes the used methods and tools and explains the proposed integrative architecture. Section IV, presents the experimental study and a detailed results' discussion. A functional and enrichment analysis of the selected genetic biomarkers is detailed in section V. A brief discussion on the association of the selected miRNAs with colorectal cancer is presented in section VI. Section VII concludes with a summary of the key findings and plans for future research.

II. RELATED WORKS

In recent decades, many artificially assisted systems for cancer diagnosis have been thoroughly investigated, either

through medical imaging analysis [19] or omic data analysis [5], [9], [20]. Although the incredible advancements brought about by AI in clinical applications cannot be understated, another need that must be met to improve genetic diagnosis, prognosis, and drug development is to provide models that assist biologists, clinicians, and the pharmaceutical industry in selecting molecular biomarkers with potential diagnosis, prognosis, and therapeutic targets. Therefore, besides integrative omics, one of the most hotly debated topics in the bioinformatics community is the in-silico molecular biomarker discovery which represents the process of feature selection. Many studies have been conducted in order to identify biomarkers with potential clinical applications [13]. Several studies have also been established to integrate omics data in order to train intelligent models to predict clinical outcomes and aid in the improvement of cancer-related medical decisions [21]. Among feature selection models, bioinspired-based models, population based models, and iterative models have gained a lot of attention because their output can be easily biologically interpreted, whereas, for omics integration, machine learning and deep learning are thought to be the best tool for dealing with the particularity and the complexity of omics data by learning a set of features representations from multiple views with different information and to use these learned features to train an intelligent tool for data driven medical decision making [20].

Discussing here some of the recent works, we notice that a wide range of the introduced feature selection models in cancer classification and clinical outcomes prediction, have been used on the mRNA gene expression data. Hybrid bioinspired models based algorithm have been introduced to select a subset of relevant genes with cancer prediction performance relevancy, like the work of Coleto-Alcudia and Vegas-Rodrigues [22], that presents a hybridization between teaching models and artificial bee colony (ABC), by first shrinking the space scalability using the ranking method and then ABC selects the most relevant gene subset. Similarly, M.Sobhanzadet et al. [23] proposed a genetic and world competitive contest (WCC) algorithm, where genetic algorithm is used to limit the number of genes while WCC is then used to select the best genes. Another hybrid metaheuristics for feature selection has been introduced in the work of Shukla et al. [24] based on teaching-learning algorithm (TLA) and gravitational search algorithm (GSA), where the authors used minimum redundancy maximum relevance to keep only genes with high relevance, then a GSA has been incorporated in the teaching phase to select the most relevant genes. Because of the aforementioned gene expression problem and the large number of noisy and irrelevant data in negative and neutral features, the authors used a two-step features selection model to select the top relevant genes. Another interesting graph theory based feature selection method has been recently introduced by Azadifar et al. [25], where the authors construct a gene-gene similarity based network, then the graph undergo a set of iterations where at each iteration a maximum clique is used to identify the optimal genes subset.

As previously stated, RPPA and proteomic data are increasingly becoming a target for molecular biomarker discovery due to their consistency of being a transcription proof of a specific gene expression and by extension the expression of a potential mutation, which allows for robust and effective multiomic integration. Takahashi et al. [26] introduced a parallel omic prediction of survival subtypes in lung cancer based on RPPA data, the results presented in this research exhibit and confirm the consistency of protemics in cancer classification. Another work presented by Isik et al. [27] that inspired our RPPA based omic biomarker discovery, uses the protein-protein interaction network to select the most correlated proteins, the gene expression of the selected proteins' coding genes have been used to predict the clinical outcome of patients. Kim [28], used RPPA with multiomic data for breast cancer survival prediction based on pathway activity inference to address the biological process and implication of learnt features. Despite the efforts made to reduce the dimensionality of omics data sets, the issue remains challenging, especially when multi-omics data must be processed using integrative approaches for improved prediction and interpretable findings. This prompted the work outlined in this paper.

III. MATERIALS AND MODELS

This paper describes an omics integrative study (Fig. 1) that is mainly based on biological data filtration and development of computational models. The proposed integrative framework yields an in-silico biomarker discovery model based on a bioinspired feature selection approach, as well as trained machine learning models that can be used as predictive tools in cancer. The protein expression data serve as the starting point for the proposed in-silico biomarker discovery model. The integrative study in phase two (Fig. 1.(B)) uses the expression of the filtered RPPA data from phase one (Fig. 1.(A)) and applies a bioinspired approach based on clustering, grey wolf optimization algorithm, signaling pathways, and protein interaction networks to select the most relevant expressed proteins that play the role of cancer proteomics biomarkers. In phase three (Fig. 1.(C)), we highlight the proteins' coding genes and miRNA targets based on the proteomics biomarker. The set of three omics biomarkers is used to train machine learning models to predict the clinical endpoints PFI, and OS in the fourth phase (Fig. 1.(D)) to test the predictive relevance of the selected biomarkers. Phase five (Fig. 1.(E)), like phase four, aims to reveal the biological and clinical interpretation and significance of the selected biomarkers.

A. PROTEIN EXPRESSION DATA COLLECTION AND PATHWAYS FILTRATION

The first step, shown in Fig. 1.(A), involves filtering the list of proteins based on their function in signaling pathways in order to retain only those instances that have a biological background reference based on their existence in the signaling pathways repositories. With the help of this filtering,

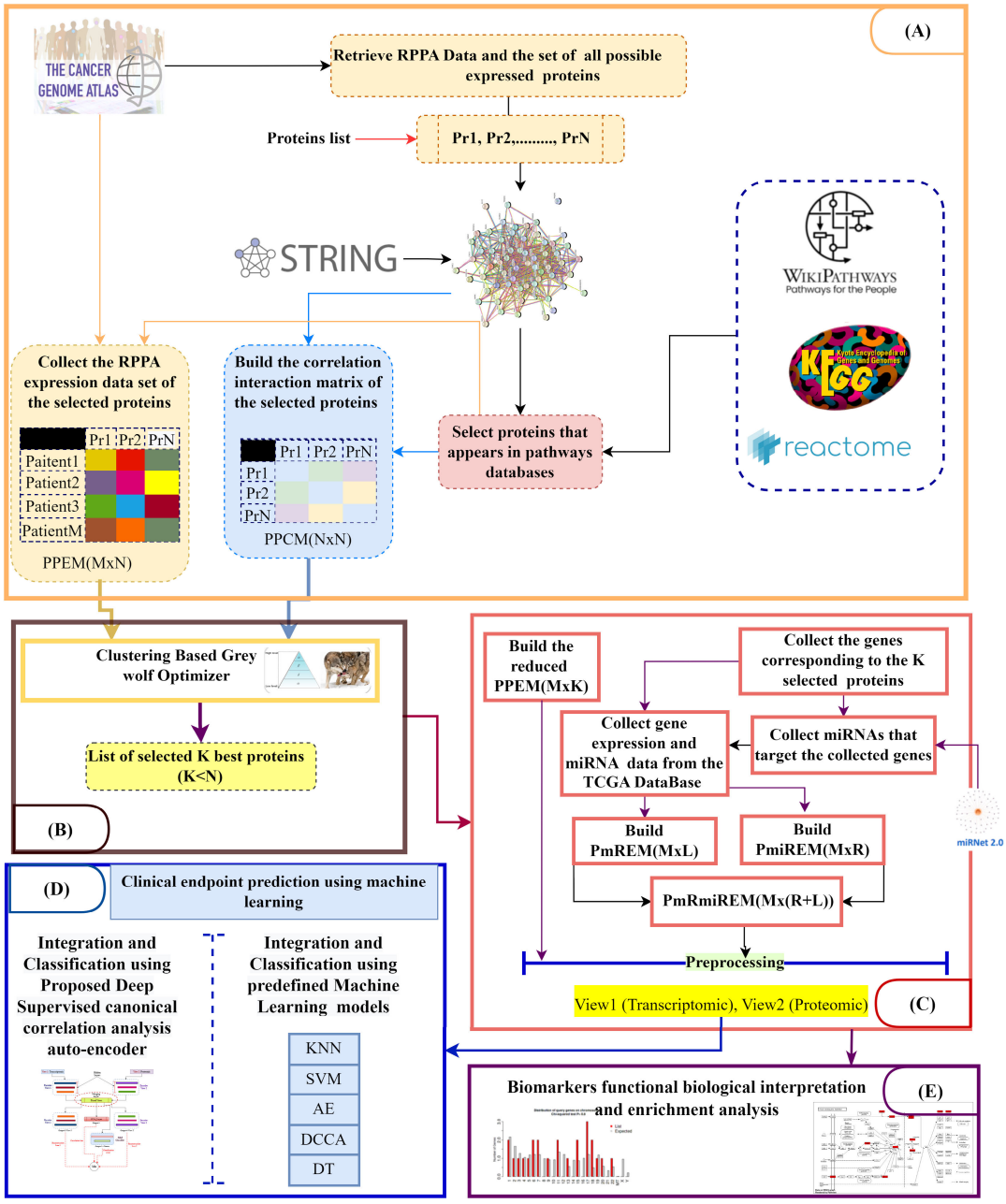


FIGURE 1. Integrative analysis workflow.

we are able to get around the issue with the RPPA technologies discussed in the introduction. Additionally, for the filtered proteins, we create a protein-protein interaction network to represent the biological, physical, and coexpression data pertaining to the interaction of the proteins. This network is then stored as a protein-protein correlation matrix, which is then used in accordance with the RPPA expression matrix to choose pertinent proteomic features.

For this purpose, RPPA profiles have been retrieved from the TCGA and TCPA portals [29], [30]. The protein list is mapped to a protein-protein network using the String data base multiple protein search query (https://string-db.org/). The Protein-Protein interaction matrix and a list of pathways from the three repositories wikipathway, KEGG, and reactome, associated to the list of proteins, were obtained. We filter the protein list in this step to keep only proteins that appear in the collected signaling pathways.We construct a protein-protein correlation matrix PPCM($N \times N$) from the protein-protein interaction network for the list of filtered proteins; this PPCM captures the correlation score between two proteins. We also collect the expression of a selected protein from available cancer samples from the TCGA data portal to

generate a patient-protein expression matrix PPEM($M \times N$), where each element x_{ij} in PPEM represents the abundance of protein (j) in the patient (i) sample. PPEM captures the patient-protein relationship in the same way that PPCM captures the biological correlation between the selected proteins. Following that, the two datasets are used to select cancer biomarkers with biological and signaling pathway backgrounds.

B. FEATURE SELECTION BASED ON GREY WOLF OPTIMIZATION AND BIOLOGICAL KNOWLEDGE DATA CLUSTERING

Advances in the integration of genomics and biological networks sparked our idea for this proposal. An integrative feature selection algorithm based on grey wolf optimisation (Fig. 1.(B)) is designed and implemented while incorporating the information from the biological interaction networks to select the best K expressed proteins with available clinical and literature evidence from patient-protein expression data. These incorporation aims to augment the correlation between the selected features by the grey wolf optimization algorithm and to enhance the convergence towards the best solution rapidly and precisely. And also to select proteins with available clinical and literature evidence.

1) GREY WOLF OPTIMIZATION

GWO is a bio-inspired metaheuristic proposed by Mirjalili et al. [31] that has gained a great deal of interest for solving optimization problems. It mimics the hierarchical social system depicted in Fig. 2.(A), which is related to the cooperative hunting behavior of grey wolves in the wild. The strategy of the GWO algorithm is inspired by the predatory nature of wolves and their cooperative intelligence when hunting large prey. Grey wolves tend to live in packs of 5 to 12 individuals and hunt in an authoritarian fashion. Each member of the pack is designated as alpha (α), beta (β), delta (δ), or omega (ω). Alpha refers to the strongest and most dominant pack member, who serves as pack leader. The alpha member will always make the final decision during a hunt, and the rest of the pack will defer to his authority. A beta member is second in rank and serves as an advisor to the alpha member, assisting him in decision-making. If the alpha dies, the beta will become the new pack leader. The deltas and omegas are the weakest wolves in the pack and must submit to the alphas and betas while deltas dominate the omegas. The interesting hunting behaviour of grey wolves can be summarized by the following main operations:

- Step 1: Locate the prey, then monitor pack members to chase and approach the chosen prey.
- Step 2: Pursue, encircle, and start harassing the prey until it stops moving.
- Step 3: Proceed with the attack.

When the superior wolf locates the prey, the pack members must obey their alpha and begin pursuing, encircling, and harassing the prey until it is isolated from the herd, at which point the attacking process begins. The formal modelling of this behaviour is described in the next section.

2) MATHEMATICAL MODELLING

GWO is a population based metatheuristic and nature inspired optimization algorithm. Like any other optimization algorithm, achieving a good balance between exploration and exploitation of the solution space is a crucial consideration for ensuring convergence to near-optimal, if not optimal, solutions. The two mechanisms used to achieve these search capabilities are the prey search and the attack. Formally, GWO is an iterative process (Fig. 2.(B)) that seeks to identify the optimal vector of decision variables for optimizing a specified objective function. In order to achieve this goal, a population of potential solutions is used. During search iteration t, the alpha wolf (w_α), represents the best solution, the second and third best solutions are beta (w_β) and delta (w_δ). The rest solution represent the omega wolves (w_ω) who are guided by (w_α, w_β, and w_δ). Omega wolves play the role of scapegoat and are guided by their superior counterparts.

a: ENCIRCLE PREY

When hunting a prey the alpha gives the order to the rest of the pack to encircle it, which is mathematically modelled by (1,2) [31].

$$\vec{\mathbf{D}} = |\vec{\mathbf{C}} \cdot \vec{\mathbf{X}}_\mathbf{p}(t) - \vec{\mathbf{X}}(t)| \tag{1}$$

$$\vec{\mathbf{X}}(t+1) = \vec{\mathbf{X}}_\mathbf{p}(t) + \vec{\mathbf{A}} \cdot \vec{\mathbf{D}} \tag{2}$$

where the two vectors $\vec{\mathbf{X}}$, and $\vec{\mathbf{X}}_\mathbf{p}$ represent the grey wolves's and the prey's current positions respectively. The coefficient vectors $\vec{\mathbf{A}}$, $\vec{\mathbf{C}}$, $\vec{\mathbf{D}}$ represent the learning parameters, which are random values that govern the hunting process and determine whether (w_ω) approaches or run away the superior wolves (w_α, w_β, *and*, w_δ). The vectors $\vec{\mathbf{A}}$, and $\vec{\mathbf{C}}$ are updated during the iterations using (3, 4), where $\vec{\mathbf{r_1}}$, $\vec{\mathbf{r_2}}$ are random vectors in [0,1]. $\vec{\mathbf{a}}$ represents the exploration-exploitation tradeoff parameters, which are updated linearly from 2 to 0 using (5):

$$\vec{\mathbf{A}} = 2\vec{\mathbf{a}} \cdot \vec{\mathbf{r_1}} - \vec{\mathbf{a}} \tag{3}$$

$$\vec{\mathbf{C}} = 2\vec{\mathbf{r_2}} \tag{4}$$

$$\vec{\mathbf{a}} = 2 - t \cdot \frac{2}{Max_{iter}} \tag{5}$$

b: HUNTING

After encircling the prey, the most crucial phase of the GWO algorithm is to initiate the hunting process (exploitation). The alpha (w_α) leads the hunt while the other members of the pack comply with his command. The exact position of the prey is unknown to the other wolves but the (w_α) is deemed the optimal solution, and its two subordinates (w_β, and w_δ) have a better knowledge of its location. The behaviour of the three highly ranked wolves that leads the pack members is modelled using (6,7) that show how to update the wolves' location and the optimal location is determined using (8).

$$\vec{\mathbf{D}_\alpha} = |\vec{\mathbf{C_1}} \cdot \vec{\mathbf{X}}_\alpha - \vec{\mathbf{X}}|,$$

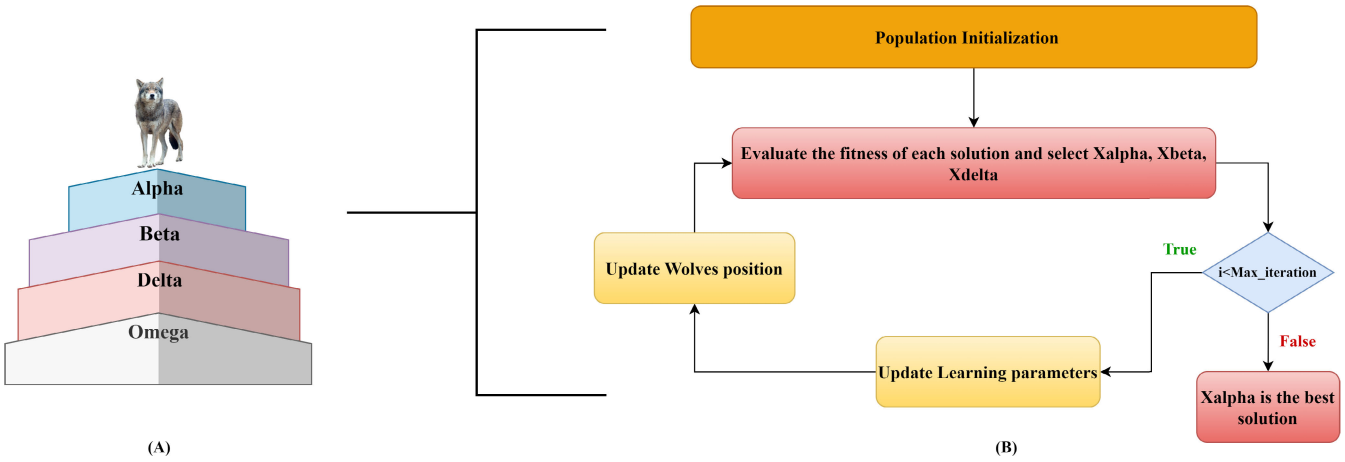


FIGURE 2. Grey wolf optimization hierarchy and inspired algorithm; (A): Social pack hierarchy; (B): Hunting behaviour inspired algorithm.

$$\begin{aligned} \vec{\mathbf{D}}_\beta &= |\vec{\mathbf{C}}_2 \cdot \vec{\mathbf{X}}_\beta - \vec{\mathbf{X}}|, \\ \vec{\mathbf{D}}_\delta &= |\vec{\mathbf{C}}_3 \cdot \vec{\mathbf{X}}_\delta - \vec{\mathbf{X}}|. \end{aligned} \tag{6}$$

$$\begin{aligned} \vec{\mathbf{X}}_1 &= |\vec{\mathbf{X}}_\alpha - \vec{\mathbf{A}}_1 \cdot \vec{\mathbf{D}}_\alpha|, \\ \vec{\mathbf{X}}_2 &= |\vec{\mathbf{X}}_\beta - \vec{\mathbf{A}}_2 \cdot \vec{\mathbf{D}}_\beta|, \\ \vec{\mathbf{X}}_3 &= |\vec{\mathbf{X}}_\delta - \vec{\mathbf{A}}_3 \cdot \vec{\mathbf{D}}_\delta|. \end{aligned} \tag{7}$$

$$\vec{\mathbf{X}}(t+1) = \frac{\vec{\mathbf{X}}_1 + \vec{\mathbf{X}}_2 + \vec{\mathbf{X}}_3}{3} \tag{8}$$

c: ATTACKING PREY

Once the prey has stopped moving, the grey wolves will attack it to end the hunt (exploitation). The narrowing of the gap between $\vec{\mathbf{a}}$ and $\vec{\mathbf{A}}$ governs the search. Where $|\vec{\mathbf{a}}|$ falls from 2 to 0, bringing the prey closer than in previous iterations, and when, $|\vec{\mathbf{A}}| < 1$ all the wolves are forced to attack the prey.

d: SEARCH FOR PREY (EXPLORATION)

In accordance with the positions of alpha, beta, and delta, the pack members disperse from one another to forage the prey and converge to attack it. From a mathematical modelling side, the vector $\vec{\mathbf{A}}$ governs the behaviour of agents in this phase; when $|\vec{\mathbf{A}}| > 1$, all search agents are forced to diverge from the current prey in order to forage a fitter one.. This also reinforces exploration and allows for global GWO search.

3) GWO FOR FEATURE SELECTION

The problem of feature selection can be framed as a multi objective optimization problem in which the optimizer seeks to identify the smallest subset of features in the solution space that achieves the highest prediction performance in classification or regression. Performing feature selection using GWO has been introduced in [32] as a wrapper model. The author used a KNN classifier [33] to determine the classification accuracy of each subset chosen by the GWO optimizer. In each iteration, each subset of features represents a wolf location, and the selected subset is the wolf's location as a result of the algorithm. The objective function that we consider in our work, as shown in (9), is an aggregation of the two aforementioned objectives, namely the number of features and prediction performance.

$$F_n = \sigma \cdot E(X) - \theta \cdot \frac{size_{SF}}{size_{AF}} \tag{9}$$

where $E(X)$ represents the classification error rate, $size_{SF}$ represents the number of selected features in each solution subset, and $size_{AF}$ represents the total number of input features The parameters $\sigma \in [0, 1]$ and $\theta = 1-\sigma$ are used to fine-tune and balance the importance of the number of features chosen and the classification error. A potential solution is encoded as a binary vector with the same size as the number of features in the dataset. If a feature is selected, it is assigned a value of 1; otherwise, it is assigned a value of 0. A threshold is used to adapt the GWO to deal with binary encoded solutions, with values of wolf location above the threshold set to 1 and values below the threshold set to 0.

4) THE PROPOSED GWO ALGORITHM FOR FEATURE SELECTION

The main idea behind the proposed GWO algorithm for selecting features is that the initial population is generated using clustering. Therefore, we refer to our proposed Clusterin Based GWO algorithm as CB-GWO. The primary objective of the proposed optimization method is to reduce the dimensionality of the PPEM as well as to identify the subset of proteins with the highest relevance, which implies removing features with negative impact on training machine learning models that aims to either predict a clinical outcome or to define a certain gene biomarkers. The proposed CB-GWO algorithm1 takes the PPEM and PPCM as inputs and functions in two phases. First, an initialization procedure is designed to generate initial positions or potential solutions. Then, this set of solutions will undergo an iterative process governed by the previously described dynamics of the GWO algorithm. The output of the optimizer is a subset of selected

features (proteins or biomarkers) determined by the position of the Alpha individual. Apart from its intriguing and promising search ability, the GWO was chosen because of its hierarchical social behavior, which allows us to update the population initialization procedure by injecting a new graph base data set (PPI network), which will be clustered into feature subsets and ranked according to the GWO hierarchy. Using a clustering algorithm and information provided by the

Algorithm 1 Clustering Base Grey Wolf Optimisation for Feature Selection Algorithm

Data: PPEM, PPCM, Max_{iter}
Result: K best features
`Population initialization(`$PPCM$`)` ;
Evaluate the fitness of each solution X using (9);
Select $X_\alpha, X_\beta, X_\delta$;
while $t <= Max_{iter}$ **do**
  Update Wolves' position using (6, 7, 8);
  Update $\vec{\mathbf{a}}$ using (5);
  Update learning parameters using (3, 4);
  Evaluate the fitness of each solution X using (9);
  Update $X_\alpha, X_\beta, X_\delta$;
  t=t+1;
Procedure `Population initialization(`$PPCM$`)`
  Result: P: Population size, X: initial wolves position
  Define optimal P number of cluster using silhouette average ;
  Cluster the PPCM input into P cluster using K-means;
  Calculate the position Matrix X of each wolf based on the distance between the centroid of $Cluster_i$ and $Protein_{ji}$.

PPCM, proteins are grouped into clusters during initialization. Therefore, we must determine the clustering algorithm and the number of clusters denoted by P. Since a natural wolf pack has five to twelve members, we used this range to determine the value that gives the best silhouette index value when using the Kmeans algorithm. After determining the optimal number of clusters P, we consider the P clusters produced by the K-means algorithm as the total number of the pack. As illustrated in Fig. 3, and in order to determine the initial positions of wolves, we proceed as follows. First, we calculate the distance between each protein and each centroid to construct the protein-to-centroid distance matrix $X(N \times P)$, in which distance values are discretized using a threshold in order to derive binary values. At the end of this process, initial potential solutions are generated.

In the second phase of the optimization procedure, the set of initial positions X is modified iteratively using GWO equations. As in the original work [32], the evaluation of solutions or calculation of fitness values is performed at each iteration using the KNN classifier on the PPEM. KNN has been considered for its simplicity as it is a lazy classification model that does not require intensive model training. Furthermore, to ensure fair comparison with the original work. To achieve this, a hold-out sampling evaluation design is implemented, with 70% of data used for training and 30% for testing. The best alpha (w_α), beta (w_β), and delta (w_δ) are selected based on fitness values, and the pack position is updated. The process continues in this manner till a termination criterion (Max iteration reached) is met. At the end of the procedure, the fittest solution, in this case the selected biomarkers, is the alpha wolf. Consequently, the output of step 2 is a set of the K best-selected proteins, which we use to reduce the dimensionality of the PPEM from (N, M) to (N, K), where $K < M$.

The time complexity of the proposed CB-GWO depends on the initialization and the iterative optimization phases of the algorithm. Let's adopt the notation shown below:

- N : the number of samples in the dataset,
- d: the dimension of the problem i.e. the number of features.
- Max_iter: the maximum number of iterations of the optimization process.

The K-means algorithm and the setting of initial positions for all individuals have the largest impact on the time complexity of the initialization phase. Consequently, the time complexity (10) of this phase is :

$$O(N^2) + O(P \cdot d) \tag{10}$$

The time complexity ((11, 12)) of the optimization phase is mostly determined by the update of individual positions and the fitness function calculation. Therefore, the time of this phase is as follows:

$$O(P \cdot d \cdot Max_{iter}) + O(P \cdot N \cdot d \cdot Max_{iter}) \tag{11}$$

As a result, the time complexity of the CB-GWO algorithm is:

$$O(N^2) + O(P \cdot N \cdot d \cdot Max_{iter}) \tag{12}$$

The space complexity is governed by the size of the data set. Hence the complexity is $O(N \cdot d)$. It is worth noting that even the time complexity is quadratic, the algorithm is used once as a preprocessing step before training and using the proposed trained classification model. The complexity of this algorithm has no impact on the prediction phase of the proposed classification model.

C. COLLECTION AND PREPARATION OF M-RNA AND MI-RNA DATA SET

Following the construction of the final PPEM by leveraging the K best features that were chosen, we proceed on to the phase of collecting and filtering the instances of gene expression and miRNA (Fig. 1.(C)). For the purpose of constructing the patient-mRNA expression matrix (PmREM) and the patient-miRNA expression matrix (PmiREM), we made use of the TCGAbiolinks to collect the mRNA and the miRNA data set of patients that were contained within the PPEM

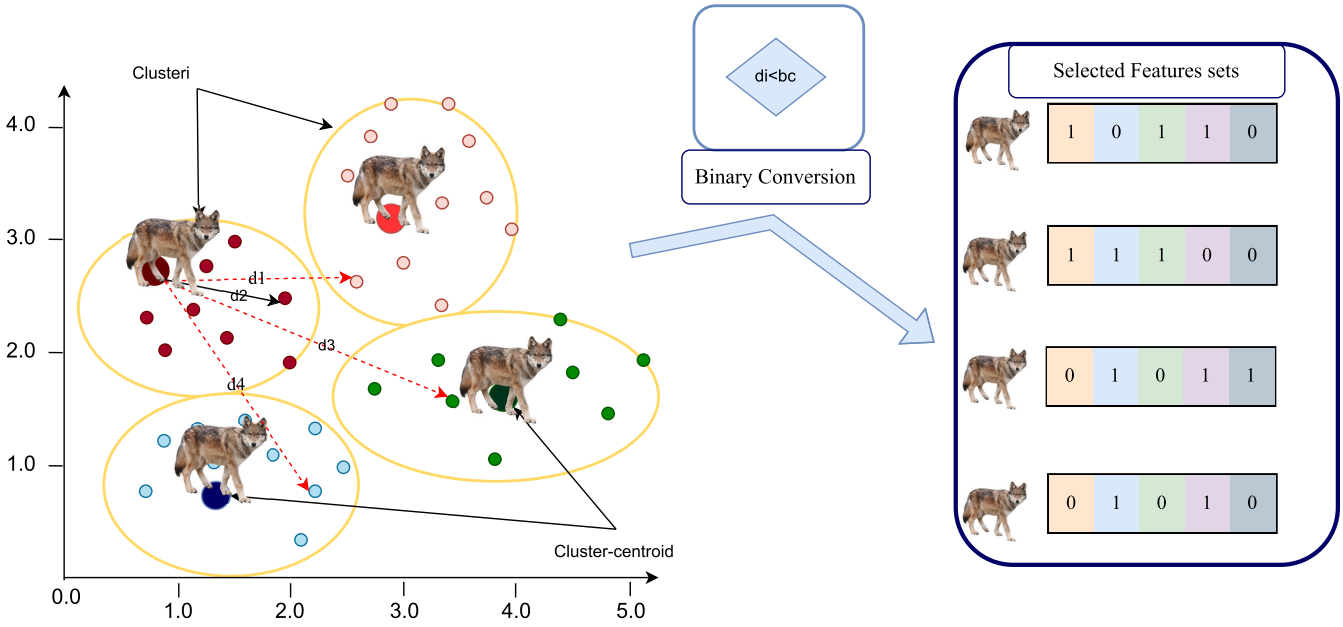


FIGURE 3. Proposed population initial distribution in CB-GWO; bc is a binary conversion threshold, bc= 0.5, and d_i is the distance between wolf i and protein j.

matrix. The number of patients in the collected data sets may vary because some patient records may miss tests on their mRNA or miRNA instances; therefore, from the three views (PPEM, PmREM, and PmiRNA), we selected the patients that appear across all the omics levels, so that the three matrices have the shape PPEM($M \times K$), PmREM($M \times H$), and PmiREM($M \times Z$). Then, the expression of genes responsible of coding the selected proteins are extracted and the final patient-gene expression matrix PmREM($M \times L$), where $L < H$ is constructed.

The subsequent step in the data integration proposal is to incorporate the biological knowledge of the miRNA-mRNA target network, where we download the miRNA potential gene targets of the collected miRNA from the *miRNet (mirnet.ca/)* database. Then, we filtered the PmiREM by retaining only the miRNAs that have targets among the PmREM genes. We ultimately construct the patient-miRNA matrix PmiREM($M \times R$), where $R < Z$. The three views are then z-score normalized to have a mean of zero and a standard deviation of one.

D. PREDICTING CLINICAL ENDPOINTS USING MACHINE LEARNING MODELS AND SELECTED BIOMARKERS

Various machine learning models were used to predict clinical endpoints from the integration of transcriptomics and proteomics views in order to test the consistency of the selected features. In addition, we proposed a deep learning architecture based on auto-encoders using unsupervised canonical correlation analysis(CCA) and supervised learning(Fig. 1.(D)). The models under consideration are classified into three types: (1) traditional supervised machine learning models, (2) deep learning models, and (3) deep features learning models. k-nearest neighbors (KNN) [33], support vector machines (SVM) [34], random forest (RF) [35], naive bayes (NB) [36], decision trees (DT) [37], and AdaBoost [38] are the models of the first class. We used convolutional neural network (CNN) [39], shallow neural networks (SN) [40], and deep neural networks (DNN) [41] for the second class. The ultimate goal of deep features learning is to train unsupervised deep autoencoders, such as the Maximum mean discrepancy variational autoencoder (mmdVAE) [42] and the deep canonical correlation autoencoder (DCCAE) [43], to learn a latent features representation that holds information from both perspectives. These models' latent spaces were used to train a supervised Adaboost classifier for endpoint prediction. The final deep features learning model is the deep supervised canonical correlation analysis autoencoder (DSCC-AE). The model's goal is to detect the occurrence of events based on the level of expression of omic profiles in patients. The proposed DSCC-AE model is described in the following section.

1) PROPOSED DEEP SUPERVISED CANONICAL CORRELATION ANALYSIS AUTO-ENCODER

Galen et al. [44] proposed the first application of deep CCA (DCCA), in which each view represents an encoding phase of a multi-view autoencoder. The DCCA takes each view separately and progresses it through a series of hidden neural layers until the final layer of each view represents the new latent feature space. These features are fused in one layer and a canonical correlation analysis is applied to transform these views into more correlated representations. Wang et al. wang2015 extended the DCCA architecture by adding two decoding networks that are built symmetrically to the deep

network of each view. These decoders attempt to reconstruct the inputs (x, y) as $(x\prime, y\prime)$ from features that passed through a canonical correlation analysis, so that $(x, y) \approx (x\prime, y\prime)$. The goal of combining autoencoder and CCA as deep canonical correlation autoencoder (DCCAE) is to learn correlated transformed features that can reproduce the same input with as little loss as possible.

Inspired by the DCCAE, we changed the objective function to learn a feature representation that is oriented to predict PFI and OS endpoints. To this end, we incorporated a deep neural network classifier to obtain a deep supervised canonical correlation autoencoder (DSCC-AE) architecture (Fig. 4). The proposed DSCC-AE uses the learned features space generated by the DCCAE' encoders to predict the corresponding class of the input. The loss error, given by the difference between the original class and the predicted one, that governs the training of the deep classifier is added to the objective function of the DCCAE, which will further tune the learned features space to be more correlated based on their corresponding classes distribution.

Therefore, as shown in Fig. 4, we took the outputs of phase (C) (PmRmiREM($M \times (L + R)$) as a transcriptomics view and PPE$M(M \times K)$ as a proteomics view; then we built two encoders accordingly. The dataset of each encoder is then propagated through its hidden layers towards the last layer of the encoder. Then the two outputs are merged into a single fusion layer that represents the latent space layer. This layer serves as input to four distinct models. Two decoders ϕ_1 and ϕ_2 that are symmetrically built to the encoders in order to reconstruct the inputs from the learnt latent space, a CCA layer to apply canonical correlation transformation, and an additional supervised multi-layer perceptrone (C) to predict a clinical endpoint based on the learned features and the target class. These four output models will train the deep architecture deep supervised canonical correlation autoencoder (DSCC-AE) cooperatively by adjusting its weight to learn a more accurate representation of correlated features representation targeting a specific clinical endpoint. The architecture is trained to minimize the set of functions depicted by (13).

$$\min_{w_f, w_g, w_{\Phi_1}, w_{\Phi_2}, w_C, U, V} [-\frac{1}{N} tr(U^T f(x) g(y)^T V), \frac{\lambda}{N} \cdot \sum_{i=1}^{N} (|x_i - \Phi_1(f(x)|^2), \frac{\lambda}{N} \cdot \sum_{i=1}^{N} (|y_i - \Phi_2(g(y_i)|^2), \lambda(\Psi(T, C(f(x)g(y))))]. \quad (13)$$

where, $(x, y\ )$ respectively represent the input views, $-\frac{1}{N} tr(U^T f(x) g(y)^T V)$, is the CCA layer loss that aims to maximize the correlation between the outputs of the two encoders g, f, $[\frac{\lambda}{N} \cdot \sum_{i=1}^{N} (|x_i - \Phi_1(f(x)|^2), \frac{\lambda}{N} \cdot \sum_{i=1}^{N} (|y_i - \Phi_2(g(y_i)|^2)]$, are the mean square reconstruction errors of the loss value between x, y and the outputs generated by the decoder Φ_1, and Φ_2. $\Psi(T, C(f(x)g(y)))$ represents the classification loss(e.g. crossentropy loss) and λ is a lasso regularisation score, to avoid the overfitting of the model. Similarly to DCCA [44] and DCCAE [43], we apply stochastic optimization and tanh activation function of the DCCAE parts to update the architecture weights and its objective.

IV. EXPERIMENTAL STUDY, RESULTS AND DISCUSSION

To evaluate the performance of the integrative framework, we collected four distinct cancer data sets: breast cancer (BRCA), colon cancer (COAD), rectum cancer (READ), and squamous lung cancer (LUSC) with two clinical endpoints from the TCGA data portal [29]. Python was used to implement the proposed CB-GWO, feature selection, and traditional machine learning models. The deep learning models and DSCC-AE were implemented using the Python machine learning keras package with tensorflow backend. Two comparative studies were conducted to demonstrate the efficiency of the proposed CB-GWO for feature selection. On the one hand, step B results were compared to state-of-the-art bioinspired and evolutionary-based optimization algorithms, such as grey wolf optimization (GWO) [32], whale optimization algorithm (WOA) [45], cuckoo search (CS) [46], bat algorithm (BA) [47], and differential evolution algorithm (DE) [48]. CB-GWO performance as a feature selection algorithm, on the other hand, has been compared to statistical and machine learning based feature selection models, specifically univariate feature selection models such as correlation based feature selection [49], and chi-square [50]. Feature importance based models such as logistic regression (L2) [51], and random forest (RF) [52], and recursive features elimination RFE [53]. Besides, we used MRMD3.0 to compare the performance of CB-GWO with an ensemble features selection method that combines features ranking methods and link analysis algorithm to finally identify a reduced data representation [54]. In phase D, the various machine learning models used in this study were compared based on their MCC and AUC score.

A. DATASET DESCRIPTION

Normalized RPPA data set and protein list, along with the clinical follow-up patient data were downloaded from the TCGA and TCPA data portals(https://portal.gdc.cancer.gov/, https://www.tcpaportal.org/), for breast cancer(BRCA), colon cancer (COAD), rectum cancer(READ), and squamous cell lung cancer(LUSC). As for the transcriptomics data we used the R package TCGABiolink to retrieve the mRNA and miRNA expression data based on the patient bar-code in the patient protein data set.

As cancer targets, we selected both overall survival (OS) and progression-free interval (PFI) as clinical endpoints (PFI). The former is the survival of a group of patients after a cancer diagnosis or the initiation of a specific treatment. The latter is the length of time a patient with cancer lives without its progression or any recorded event. The binary nature of the two endpoints results in a binary classification. The description of the data sets will be presented in

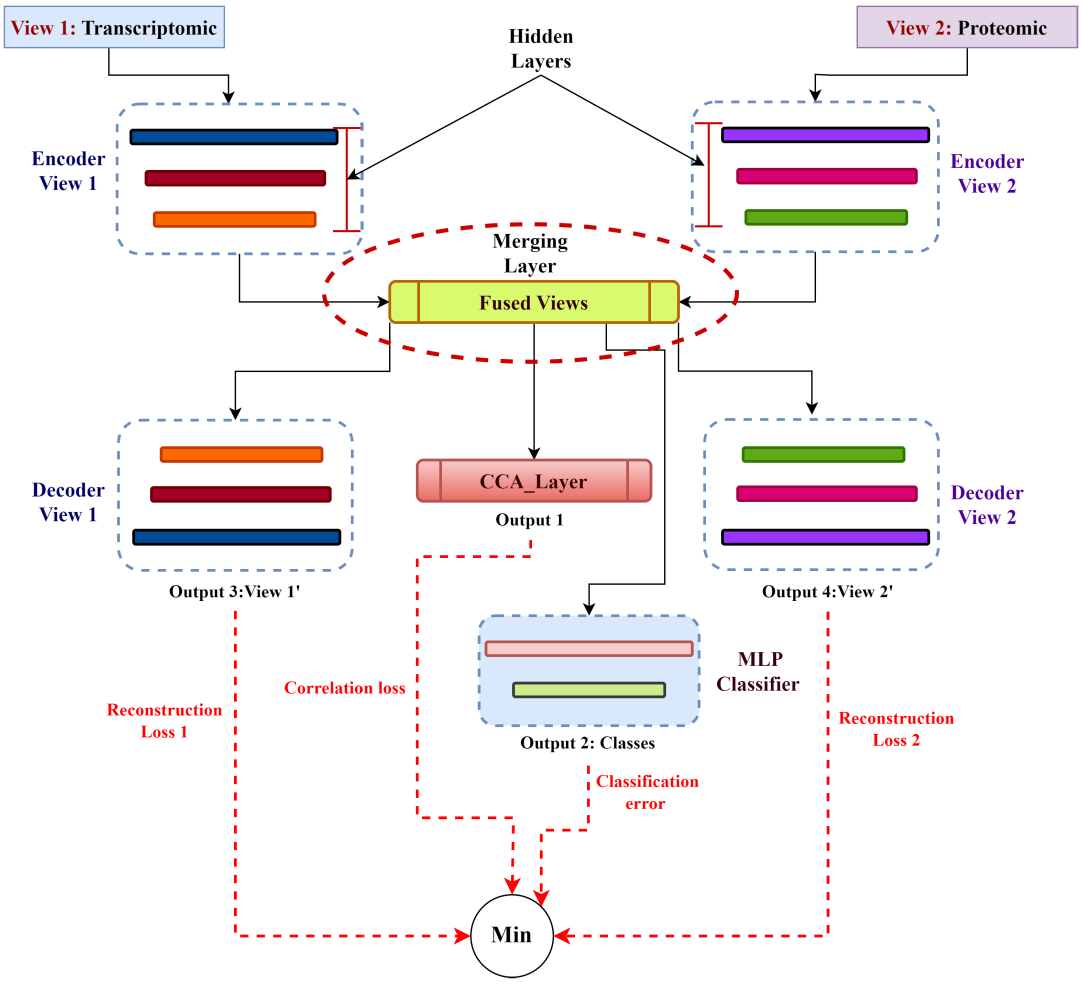


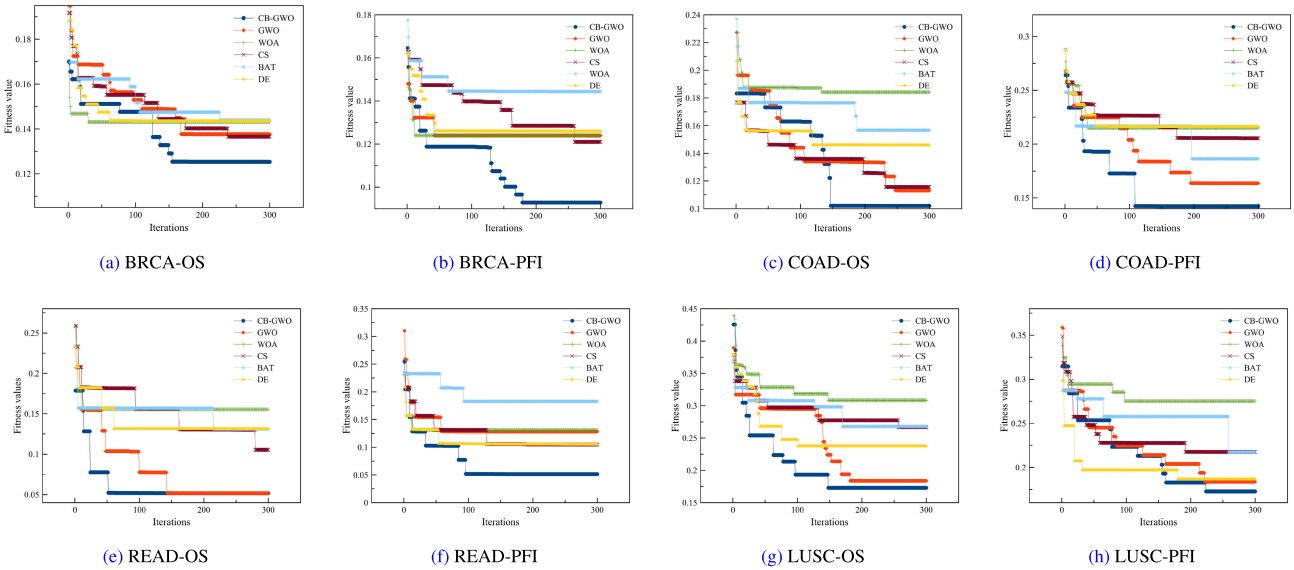
FIGURE 4. Architecture of the multiview supervised cannonical correlation analysis integrative model.

the remaining sections of the paper in the order of their application.

B. CB-GWO VS OTHER FEATURE SELECTION METHODS

Table 1 shows the dimensionality of the RPPA data set for each type of cancer and each clinical endpoint used to identify cancer biomarkers. It also shows the initial accuracy of the KNN classifier when the entire data set (including all features) is considered for each cancer type and target. As can be seen, the classifier achieved the highest accuracy, 83 %, with BRCA dataset and PFI target while it achieved lowest accuracy, 54%, with the LUSC dataset and OS target. For comparison, the accuracy of the KNN classifier was computed for the reduced dataset using only the features identified by the proposed CB-GWO and other cutting-edge bio-inspired algorithms. To ensure a fair comparison, all algorithms used the same population size determined by the initialization procedure, as well as the same number of iterations which was set to 300. To monitor the convergence of the bio-inspired algorithms, we recorded the objective function values across iterations. Fig. 5 depicts the behaviour of the algorithms during the minimization process. All plots clearly show that the CB-GWO surpassed all other bioinspired and evolutionary-based algorithms in terms of fitness. The performance of the algorithms can be further analyzed over three iteration ranges [0-100], [100-200], [200-300] as follows:

- In the first 100 iterations, CB-GWO achieved the best fitness value in, BRCA-PFI, COAD-PFI, READ-OS, and LUSC-OS, whereas, for READ-PFI, CB-GWO, and DE were able to achieve approximately the same fitness value with a superiority of ≈ 0.01 for CB-GWO. As for the rest of datasets, CS converged first to the best solution in COAD-OS, while DE converged rapidly towards the best fitness value in LUSC-PFI. As for BRCA-OS, WOA, and DE achieved the best fitness value compared to the rest of algorithms.
- In the following 100 terations CB-GWO was able to outperform CS, WOA, and DE in BRCA-OS, and COAD-OS. In LUSC-PFI, CB-GWO, and DE achieved approximately the same best fitness value by the iteration 160,180 respectively with a superiority of ≈ 0.01 to CB-GWO.



(a) BRCA-OS (b) BRCA-PFI (c) COAD-OS (d) COAD-PFI

(e) READ-OS (f) READ-PFI (g) LUSC-OS (h) LUSC-PFI

FIGURE 5. Convergence Performance comparison on RPPA Data.

TABLE 1. Description of dataset size, population size, and initial accuracy.

		Initial Data size	Accuracy	Population Size
BRCA	OS	(900 × 176)	81.48	6
	PFI		82.59	
READ	OS	(128 × 176)	76.92	11
	PFI		66.66	
COAD	OS	(327 × 176)	71.42	6
	PFI		72.44	5
LUSC	OS	(325 × 184)	54.08	5
	PFI		57.14	

- At the last set of iterations, CB-GWO was able to achieve a better solution regarding LUSC-PFI by the iteration 220, while DE stabilized at the same fittest value scored in iteration ≈ 180.

The objective function values attained after 300 iterations by all bio-inspired algorithms applied o each data set and target are detailed in table (2). As we seek to minimize the objective function, it is evident that the proposed CB-GWO produced the best results in all cases except for the READ-OS dataset, where GWO-like results were obtained. In addition, the algorithm's performance was evaluated based on its ability to select a relevant and concise minimum set of features and compared to the other algorithms. Tables 3, and 4 show that, across all test cases, the proposed CB-GWO achieved the highest accuracy with a remarkably smaller set of features. Both GWO and CB-GWO achieved the same classification accuracy and fitness value on the READ-OS dataset, but the CB-GWO identified a smaller set of features.

As for the statistical and machine learning-based feature selection models, the selected models have been trained to acquire the same number of features subset chosen by the CB-GWO algorithm using the default parameters defined by the standard representation of the models. The learned features have been used to train a KNN model to predict PFI and OS. Tables 5, and 6 compare the performance of CB-GWO to other models; in all cases, CB-GWO achieved the highest accuracy score.

When comparing the results of table1, and tables 3, 4, 5, and 6, it is evident that selecting features is preferable to taking the entire set of features. The classifier performed many orders of magnitude better with the selected features, particularly with the proposed CB-GWO where an 18% improvement was achieved for the READ-OS dataset. In comparison to black box dimensionality reduction with feature transformation, the use of feature selection methods guarantees the interpretability of features, which is crucial for biomarker discovery and biological interpretation. Furthermore, tables 3, and 4 show that GWO outperforms all bioinspired algorithms except CB-GWO. This justifies our selection of GWO, which we enhanced by incorporating a suitable initialization procedure based on biological protein-protein correlation that captures the proteins with potential signalling pathway implication to produce CB-GWO. This incorporation of biological knowledge improved the filtering and selection of features, as evidenced by the results of READ-PFI, where only 13 features were selected using the CB-GWO and the KNN model was trained to classify patients according to their PFI class using these features. The same result can be drawn from examining tables 5 and 6, even when using robust tools like MRMD3.0, the integration of biological knowledge data significantly improved GWO performance. As a result, the investigation into combining this type of data with various features selection methods and tools has the potential to be revolutionary.

At the end of this phase (Step B in the proposed framework shown in Fig. 1), we construct the final $PPEM(M - K)$ for each data set and for each target. The initial sizes

TABLE 2. Objective function value comparison between the proposed CB-GWO against bio-inspired and evolutionary based algorithms.

		CB-GWO	GWO	WOA	CS	BA	DE
OS	BRCA	**0.125**	0.137	0.143	0.136	0.143	0.143
	CODA	**0.101**	0.113	0.184	0.115	0.156	0.145
	READ	**0.051**	**0.051**	0.155	0.105	0.131	0.131
	LUSC	**0.172**	0.183	0.308	0.267	0.267	0.237
PFI	BRCA	**0.092**	0.123	0.123	0.121	0.144	0.125
	CODA	**0.142**	0.163	0.215	0.205	0.186	0.216
	READ	**0.051**	0.128	0.131	0.104	0.182	0.104
	LUSC	**0.172**	0.183	0.275	0.217	0.217	0.186

TABLE 3. Performance of CB-GWO against bioinspired and evolutionary based algorithms in OS; SF:Selected Features.

	BRCA	COAD	READ	LUSC	
CB-GWO	**12**	**17**	**16**	**23**	N° SF
GWO	35	36	19	31	
WOA	66	42	54	98	
CS	80	79	70	80	
BA	81	91	76	95	
DE	72	76	75	92	
CB-GWO	**87.4**	**89.79**	**94.87**	**82.65**	Accuracy
GWO	86.29	88.77	**94.87**	81.63	
WOA	85.92	81.63	84.61	69.38	
CS	86.66	88.77	89.74	73.46	
BA	85.92	84.69	87.17	73.46	
DE	85.9	85.71	87.17	76.53	

TABLE 4. Performance of CB-GWO against bioinspired and evolutionary based algorithms in PFI; SF:Selected Features.

	BRCA	COAD	READ	LUSC	
CB-GWO	**21**	**15**	**13**	**18**	N° SF
GWO	52	34	24	34	
WOA	52	51	74	45	
CS	65	60	60	99	
BA	89	79	88	95	
DE	86	71	57	86	
CB-GWO	**90.74**	**85.71**	**94.78**	**82.65**	Accuracy
GWO	87.77	83.67	87.17	81.63	
WOA	87.77	78.57	87.17	72.44	
CS	88.9	79.59	89.74	78.57	
BA	85.92	81.63	82.05	78.57	
DE	87.77	78.57	89.74	81.63	

TABLE 5. Accuracy performance of CB-GWO against feature selection models in OS.

	BRCA	COAD	READ	LUSC
N° Selected features	12	17	16	23
CB-GWO	**87.4**	**89.79**	**94.87**	**82.65**
Correlation	83.7	78.57	71.79	59.18
Lasso	83.33	81.63	79.48	61.22
RF	84.81	79.59	79.48	65.3
chi-square	82.96	75.51	71.79	59.18
RFE	84.07	81.63	82.05	66.32
MRMD3.0	0.85	0.78	0.80	0.66

of feature sets for mRNA-seq data and miRNA data are 19947 and 1881 respectively. We construct the final mRNA and miRNA expression matrix and build the transcriptomics view $PmRmiREM(M \times (R + L))$ using the data collection and filtration techniques described in phase 3 (section III-C).

TABLE 6. Accuracy performance of CB-GWO against feature selection models in PFI.

	BRCA	COAD	READ	LUSC
N° Selected features	21	15	13	18
CB-GWO	**90.74**	**85.71**	**94.78**	**82.65**
Correlation	85.56	69.38	74.36	63.26
Lasso	83.7	73.46	66.66	64.28
RF	87.03	67.34	79.49	72.45
chi-square	87.03	72.45	76.92	64.28
RFE	83.7	64.29	74.36	73.46
MRMD3.0	0.86	0.74	0.78	0.63

In table 7, which provides a description of the data set, we can see that the number of genes and miRNA has been reduced from thousands and hundreds considerably. The number of selected miRNA instances for colon and rectum cancer is significantly higher than for other cancer types, as these two cancer types are classified as colorectal cancer (CRC). The role of miRNAs in CRC is described in section VI.

C. RESULTS OF MACHINE LEARNING MODELS AND DATA INTEGRATION FOR CLINICAL ENDPOINT PREDICTION

1) DSCC-AE IMPLEMENTATION AND RANDOM CROSS-VALIDATION EVALUATION

We used both hold-out validation and K-fold cross validation to evaluate the proposed model's performance. We were inspired by the work of Chuang et al. [18] to create a variety of training and test sets. The scenarios listed below have been considered.

- 15 scenarios of K-fold cross validation were generated to test the performance of the models, with BRCA, COAD, and LUSC evaluated using 10-fold cross validation, implying 150 randomly generated training/testing datasets, and READ evaluated using 4-fold cross validation due to its small size and to ensure the presence of the minority class in the testing sets.
- 50 randomly generated data sets, 80% training and 20% testing, to be used for hold-out validation across all datasets and situations.

The following architecture has been adopted for the proposed DSCC-AE after a series of experiments.

- Two encoders for transcriptomics and proteomics data with two hidden layers and bottelneck layers (100,100,10)
- A layer with 20 nodes (merging layer) that combines the outputs of the two encoders.
- Symmetrically to the two encoders, two decoders are employed to reconstruct the inputs using two hidden layers and an output layer: (10,100,100).
- A CCA layer to adjust the correlation between the inputs from the merging layer
- A deep forward classifier with a batch normalization layer to normalize the learned features in the fused layer, two hidden layers (100,50) and an output layer to predict the patient class.

TABLE 7. Integration data description.

		N°:RPPA	N°: mRNA	N°:miRNA	N°:samples	Classes	
						0	1
BRCA	OS	12	12	2	856	738	118
	PFI	21	21	5		746	110
COA D	OS	17	17	33	216	168	48
	PFI	15	15	27		159	57
READ	OS	10	10	29	73	56	17
	PFI	13	17	16			
LUSC	OS	23	23	5	309	182	127
	PFI	18	18	2		206	103

The architecture was trained using mini-batch training, with batch sizes ranging from 65 to 250 depending on the size of the data set and a learning rate ranging from $[10^{-9} - 10^{-7}]$ to 200 epochs. For the encoder, decoder, and classifier's hidden layers, the non linear "*tanh*" activation function was used, while "*softmax*" was used for the output layer. In addition to DSCC-AE, all machine learning, feature learning, and deep learning models were trained and tested on various randomly generated training/test datasets in order to assess the relevance of the selected features in previous phases on the one hand, and compare their performance to the designed DSCC-AE on the other.

2) RESULTS AND DISCUSSION

To evaluate the performance of the proposed model and other machine learning models, we used two performance measures: AUC and Mathew Correlation Coefficient (MCC). The results are presented in the form of mean values and Box-and-Whiskers plots (boxplots) of the AUC and Mathew correlation coefficient (MCC). The two metrics were chosen based on recent research presented by Chicco et al. [55] demonstrating that when dealing with imbalanced data, measures such as accuracy and F1-measure are misleading when the minority class represents the true negative (TN), as in our case. The MCC can be seen as the natural extension of the phi coefficient, which was first introduced by Udny Yule [56]. MCC is frequently used in the fields of bioinformatics and Machine learning. The MCC value for a classification model can be computed using the confusion matrix as follows:

$$MCC = \frac{TP \cdot (TN - FP) \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

MCC has been shown to be a more reliable measure because it depends on all confusion matrix elements (true positive (TP), true negative (TN), false positive (FP), and false negative (FN) (14)) and only performs well if the majority of TP and TN are correctly predicted [57], [58]. Whardani et al. [59], on the other hand, demonstrated AUC score as the most recent robust metric to measure imbalanced data because it asserts how accurately the majority and minority classes are predicted. Tables 8, and 9 show the results obtained for all models using K-fold cross validation and hold-out validation, respectively. Comparing the tables, it is evident that the results from k-fold cross-validation are significantly superior to those from hold-out validation, except for the READ data set, where hold-out validation results exceeded k-fold cross validation results by +0.01 for AUC and MCC scores.

As depicted in table 8, DSCC-AE achieved the best possible AUC performance across all of the tested data sets, with scores of 0.8 for the COAD, READ, and LUSC datasets and 0.75 for the BRCA dataset. On the READ dataset, shallow and convolutional networks, together with mmdVAE, achieved a performance comparable to DSCC-AE. In terms of MCC, DSCC-AE was able to achieve an MCC $>$ 0.5 for the COAD, READ, and LUSC datasets, but failed in the case of BRCA datasets, where SN and DCCAE achieved the best possible MCC value of 0.44 in BRCA-OS and 0.48 in BRCA-PFI, respectively. Similarly to the AUC score, some models outperformed others, such as the LUSC-PFI by DCCAE and SN.

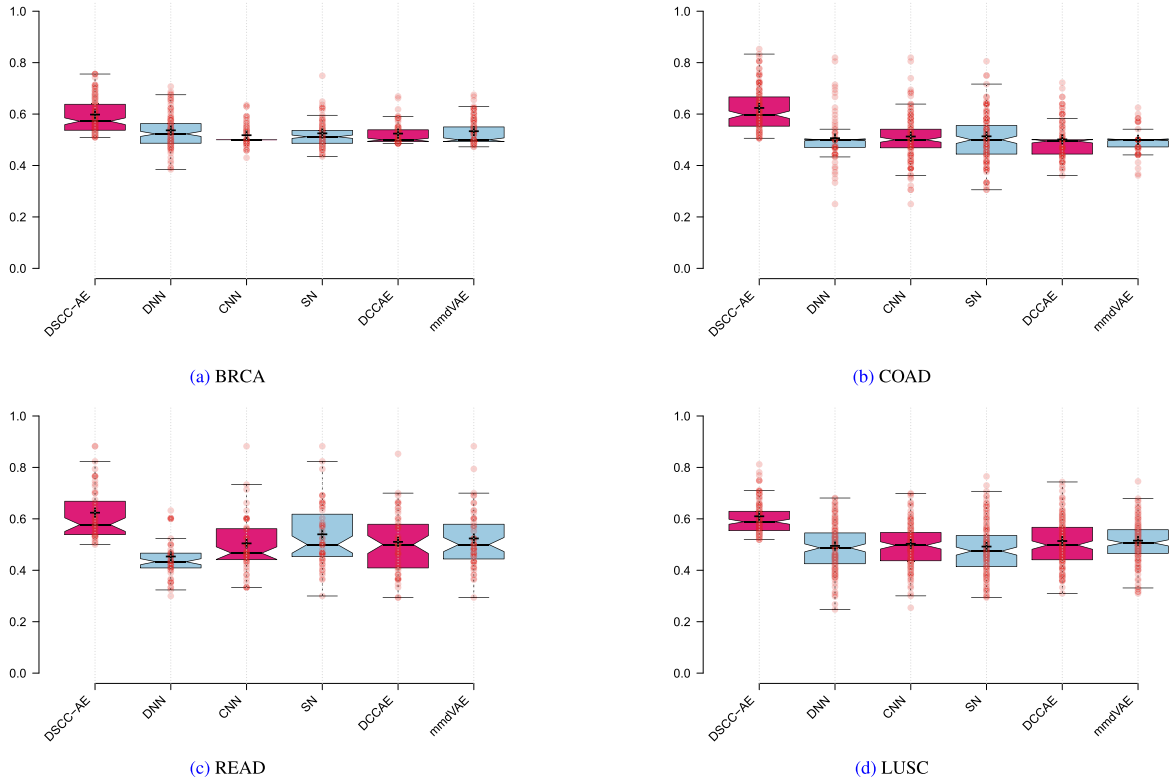
To assess deeply and better understand the performance of the various models and also to ensure more credibility and fairness of the evaluation process, we reported the results across the 15 scenarios using K-fold cross validation in the form of box-plots. There are a total of 150 tests conducted on the BRCA, COAD, and LUSC datasets, and 60 tests conducted on the READ dataset, and their respective boxplots are shown in Figs.6, and 7. When comparing the boxplots from other models, the DSCC-AE model consistently produces the best resultsin terms of the five summary statistics that we can get from the boxplots. The vast majority of observations for distributions like BRCA-OS, COAD-OS, READ-OS, and COAD-PFI cluster around the middle and first quartile.

In addition, we reported the mean values of AUC and MCC across the 15 scenarios for the proposed DSCC-AE, as well as the other machine learning models, via barplots depicted in Figs.8, and 9 respectively. The obtained results indicate that the performance of the DSCC-AE during the randomly generated training/testing was the best for both measures.

The random generation of different training/testing data sets enables the selection of a combination containing the most informative samples, thereby enabling the model to achieve a very promising performance.

TABLE 8. K-fold cross-validation best performance results, across 15 scenarios, of the proposed DSCC-AE and the other machine learning models using the selected.

		KNN	SVM	RF	NB	DT	AdaBoost	mmdVAE	DCCAE	CNN	SN	DNN	DSCC-AE	
BRCA	AUC	0.55	0.52	0.5	0.72	0.69	0.61	0.6	0.6	0.62	0.73	0.68	**0.78**	Overall Survival
	MCC	0.3	0.18	0	0.3	0.41	0.43	0.43	**0.44**	0.38	**0.44**	0.35	0.33	
COAD	AUC	0.55	0.69	0.6	0.63	0.81	0.58	0.67	0.84	0.68	0.72	0.8	**0.91**	
	MCC	0.16	0.29	0.4	0.46	0.55	0.14	0.42	0.64	0.32	0.51	0.64	**0.71**	
READ	AUC	0.5	0.7	0.5	0.55	0.6	0.6	0.84	0.88	**0.91**	**0.91**	0.84	**0.91**	
	MCC	0	0.57	0	0.12	0.25	0.21	0.51	0.5	0.57	0.57	0.44	**0.64**	
LUSC	AUC	0.6	0.59	0.68	0.45	0.75	0.66	0.69	0.75	0.76	0.75	0.76	**0.8**	
	MCC	0.31	0.27	0.48	0.08	0.5	0.36	0.38	**0.62**	0.52	**0.62**	0.5	**0.62**	
BRCA	AUC	0.68	0.52	0.56	0.63	0.7	0.61	0.68	0.69	0.63	0.74	0.71	**0.76**	Progression Free
	MCC	**0.6**	0.19	0.32	0.33	0.33	0.5	0.5	0.44	0.18	0.35	**0.48**	0.35	
COAD	AUC	0.67	0.5	0.5	0.78	0.71	0.63	0.68	0.72	0.82	0.75	0.81	**0.85**	
	MCC	0.42	0	0	0.49	0.41	0.46	0.5	0.5	0.58	0.38	0.58	**0.59**	
READ	AUC	0.5	0.69	0.66	0.85	0.85	0.61	**0.88**	0.85	**0.88**	**0.88**	0.63	**0.88**	
	MCC	0	0.32	0.24	0.45	0.45	0.13	0.5	0.44	0.5	0.5	0.18	**0.51**	
LUSC	AUC	0.6	0.55	0.61	0.54	0.76	0.56	0.75	0.74	0.7	0.77	0.68	**0.81**	
	MCC	0.23	0.25	0.21	0.21	0.59	0.15	0.49	0.48	0.4	0.53	0.38	**0.61**	



(a) BRCA (b) COAD

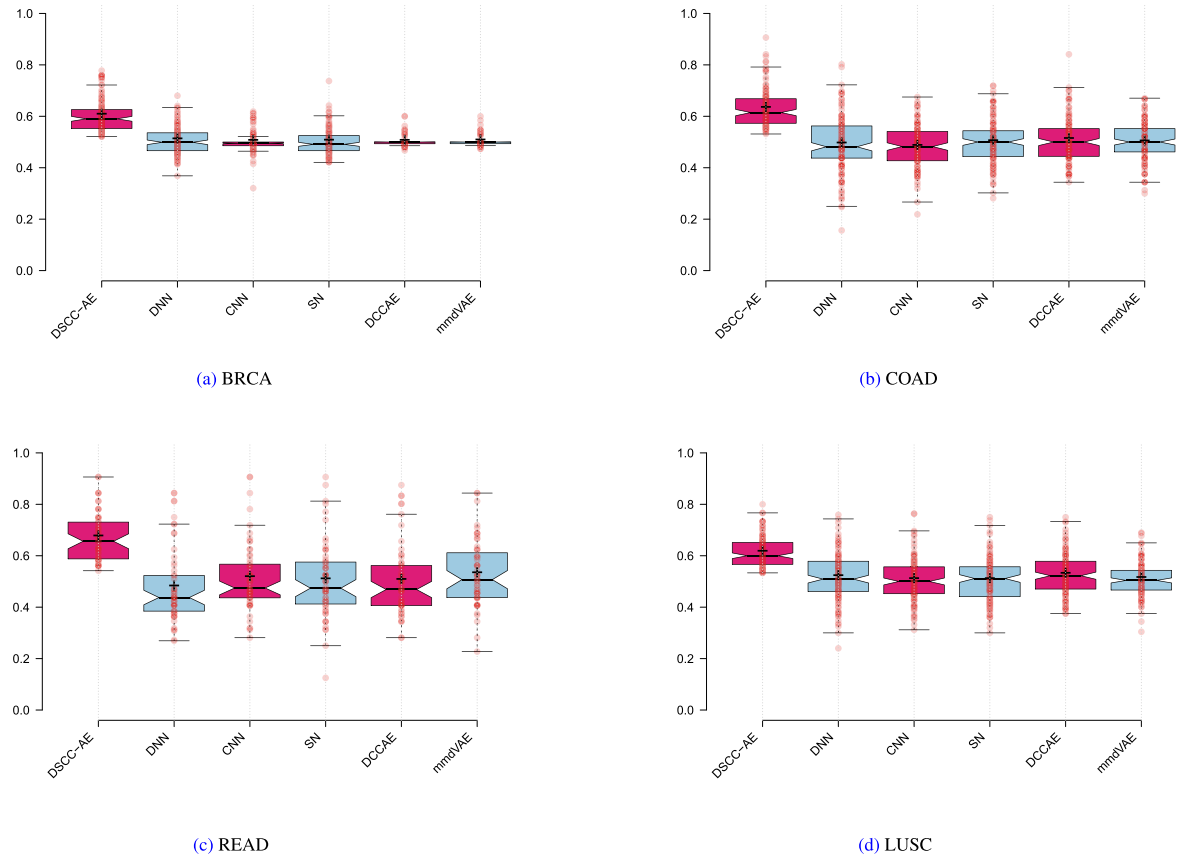
(c) READ (d) LUSC

FIGURE 6. K-fold AUC boxplot performance of DSCC-AE, and deep learning models in predicting OS across 15 testing scenario.

V. FUNCTIONAL AND ENRICHMENT ANALYSIS

In order to evaluate and validate the discovered biomarkers we performed a functional analysis for the selected genes (Fig. 1.(E)), we queried them in Gene Ontology [60], ShinyGo [61] to list the biological process of the genes, and to visualise the statistics of the genes distribution on different biological processes through fold enrichement based on False discovery rate (FDR) and to define the most expressed signalling pathways, and their interaction network. We also queried the list of biomarkers of each cancer-endpoint in cytoscape [62], using String package to visualise the protein-protein interaction network and to extract some of the most cancer-affecting gene ontology biological process terms and KEGG pathway. Every GO term and KEGG pathway with an FDR< 0.05 was considered significant.

We used IDEP.93 [63] to visualize the correlation heatmap of the expression of each selected mRNA data set by querying the genes expression matrices of the selected biomarkers for each cancer type and clinical endpoint. To select hub genes, or rather address them as the most altered genes, from the set of discovered biomarkers, we queried the genes using CBioportal [64], to analyze the performance of genes on



(a) BRCA (b) COAD

(c) READ (d) LUSC

FIGURE 7. K-fold AUC boxplot performance of DSCC-AE, and deep learning models in predicting PFI across 15 testing scenario.

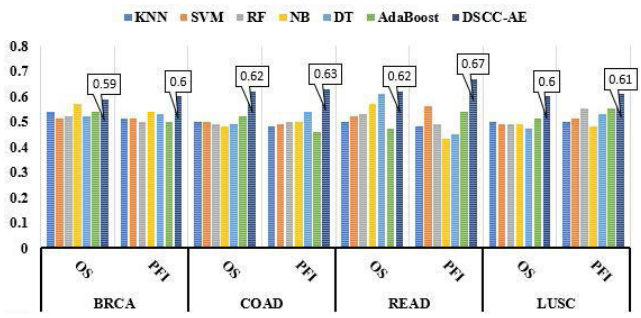


FIGURE 8. Mean AUC value of DSCC-AE and traditional machine learning performance across 15 testing scenarios through K-fold cross validation.

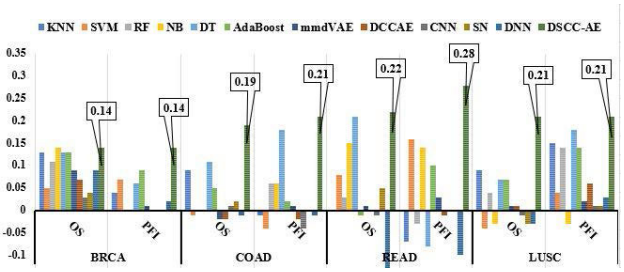


FIGURE 9. Mean MCC value of DSCC-AE and machine/deep learning models across 15 testing scenarios through K-fold cross validation.

larger cancer data bases, where we queried the genes set of each cancer type independently(BRCA/ CRC/ LUSC), then we integrated the samples of the three cancer types and queried the list of shared genes between the three types of cancer to visualise the overall survival and progression free interval prognosis as well as the statistics of genes distribution on patient samples. We selected a set of hub genes and frequent mutation variants from the CBioportal study, and then used Varsome [65] and COSMIC [66] to interpret the clinical significance of these variants. Finally, we examined the functionality of the selected miRNA in COAD and READ cancers using miRNET2.0 [67], miRSystem [68], and DIANATools-miRPath [69]. Because colorectal cancer (CRC) is the most used term in scientific reports, we refer to COAD and READ cancers as colorectal cancer (CRC) in the rest of the paper.

A. THE 77 DISCOVERED BIOMARKER GENES SHOW IMPORTANT IMPLICATION, IN NEGATIVE REGULATION BIOLOGICAL PROCESS AND HIGH CO-AFFECT OF KEGG LEADING CANCER PATHWAYS

We collected the set of functional analysis results illustrated in Fig. 10 of the 77 selected genes (table 1 in appendix A) from the experimentation on cancers and clinical endpoints in the second step of the integrative framework using shinyGo

TABLE 9. Hold-out (80%-20%) validation best performance results, across the 50 scenarios, of DSCC-AE and the other machine learning models using selected biomarkers.

		KNN	SVM	RF	NB	DT	AdaBoost	mmdVAE	DCCAE	CNN	SN	DNN	DSCC-AE	
BRCA	AUC	0.55	0.52	0.6	0.58	**0.64**	0.53	0.59	0.58	0.59	0.61	0.58	**0.64**	**Overall Survival**
	MCC	0.16	0.18	0.3	0.16	0.25	0.12	**0.36**	0.24	0.24	0.19	0.16	0.21	
COAD	AUC	0.55	0.6	0.56	0.57	0.65	0.58	0.62	0.64	**0.68**	0.65	0.62	**0.68**	
	MCC	0.1	0.26	**0.31**	0.17	0.27	0.16	**0.31**	0.3	0.28	0.27	0.24	**0.31**	
READ	AUC	0.5	0.88	0.85	0.85	0.88	0.71	0.65	0.71	**0.92**	0.88	0.75	**0.92**	
	MCC	0	0.55	0.48	0.45	0.55	0.42	**0.68**	0.42	0.65	0.55	**0.68**	0.65	
LUSC	AUC	0.53	0.52	0.58	0.58	0.66	0.58	0.63	0.64	0.59	0.66	0.64	**0.67**	
	MCC	0.1	0.13	0.2	0.17	0.32	0.19	0.27	0.3	0.19	0.33	0.3	**0.34**	
BRCA	AUC	0.55	0.52	0.52	0.52	0.63	0.54	0.53	0.53	0.59	0.63	0.60	**0.65**	**Progression Free**
	MCC	0.19	0.19	0.19	0.04	0.25	0.1	0.13	0.23	**0.29**	0.21	0.20	0.21	
COAD	AUC	0.57	0.5	0.5	0.61	0.67	0.55	0.6	0.62	0.64	0.61	0.62	**0.69**	
	MCC	0.17	0.01	0	0.34	0.3	0.11	0.26	0.32	0.31	0.19	0.21	**0.36**	
READ	AUC	0.5	0.5	0.71	0.55	0.88	0.46	0.75	0.88	0.71	0.63	0.71	**0.92**	
	MCC	0	0	0.43	0.12	0.55	-0.1	0.65	0.55	0.42	0.21	0.42	**0.65**	
LUSC	AUC	0.51	0.52	0.56	0.54	0.68	0.62	0.61	0.63	0.56	0.61	0.66	**0.71**	
	MCC	0.06	0.11	0.22	0.08	0.38	0.28	0.26	0.33	0.22	0.26	0.35	**0.42**	

and gene ontology. Figs 10.(A), and (B) visualize the gene ontology biological process of the selected genes, in which the genes are primarily involved in cellular, and developmental process, biological regulation, response to stimulus, and signalling process as regulatory genes or suppressing and communication genes. The fold enrichment demonstrates that some of the selected genes have a negative regulation of apoptotic process and cell death (table 3 in appendix A), which has a direct impact on any update or sudden events on cancer patients which may lead to recording a new event on the tumor state that affects directly the progression free interval and the overall survival of the patient. Figs. (10.C, 10.D, 10.E, 10.F) exhibit the characteristics of the selected genes in comparison to the whole genome,revealing that all of the selected genes are protein coding genes that are randomly distributed on all chromosomes with a 0.72 Chi-squared test P-value, though it have higher transcript isoforms per coding genes (P=0.0079), and higher GC density (P=0.2).

Fig.11 depicts the discovered pathway-pathway interaction network, for each cancer type, with an edge connecting two pathways if they share more than 20% of the queried genes. The pathway-pathway interaction network, reveals that the generated enrichment pathways are primarily the most relevant signal transduction pathways responsible of regulating gene activities, and signalling, besides these pathways are primarily the most important signal transduction pathways in charge of regulating gene activity. The experiment shows that the PI3KAKT signaling pathway is linked to the three cancer types (table 4 in appendix A), with 9 genes for CRC and BRCA and 14 genes for LUSC having a direct implication in the mTOR signalling pathway. Inhibitors of the PIK3/AKT/mTOR pathways have been shown to be effective treatment targets for solid cancers [70], [71]. The implication of microRNAs in solid cancer can be visualised through the interaction between pathways and on the number of implicated selected genes (>10 genes, table 4 in appendix A), microRNA have a significant association with the negative regulation of oncogene, regulator and suppressor genes that may affect the behaviour of tumour cells [72], [73]. The interaction between colorectal cancer pathway and the AGE-RAGE signalling pathway diabetic complication comes from the fact that RAGE have been implicated in the pathogenesis of several diseases including colorectal cancer [74]. Furthermore, recent studies implies that type 2 diabetes patient are subjected to develop colorectal cancer, as well as the RAGEs circulating is a potential CRC risk factor with relation to type 2 diabetes inflammation [75], [76].

The findings depicted in the pathway-pathway interaction networks visualize the relevance of the selected gene signatures based on proteomic, PPI networks, and pathway analysis in the first and second phase of our integrative model (Fig. 1.(A)/(B)), where, the selected genes are highly connected to the most important signaling pathways regarding cancer research. Also the interconnection between the pathways through shared genes can be a targeting study for metastasis cancer behaviours, and therapeutic targets, on an in-silico and in-vitro level.

Using IDEP, we built the heatmap matrices of each cancer type and its corresponding clinical endpoint, with a cut-off z-score =3, and correlation based distance and average linkage. The eight matrices are shown in Fig.12, where we can see that the patients can be clustered into two major groups, in our case PFI class 0 or 1, OS class 0 or 1. The impact of an imbalanced data set can be seen on the heatmap matrices, where we notice that the expression of mRNA abundance is unequally distributed between the samples.

B. SELECTING HUB GENES WITH HIGHEST ALTERATION FOR EACH CANCER TYPES

To determine the most altered genes for each cancer type, we queried the set of genes for each cancer type independently in CBioportal using all available related cancer studies, where as listed in (table 3 in appendix A), we queried the BRCA genes on 17 related studies with a total of 10811 samples, the LUSC genes on 4 studies with a total of 1256 samples, and the CRC genes on 13 related studies with a

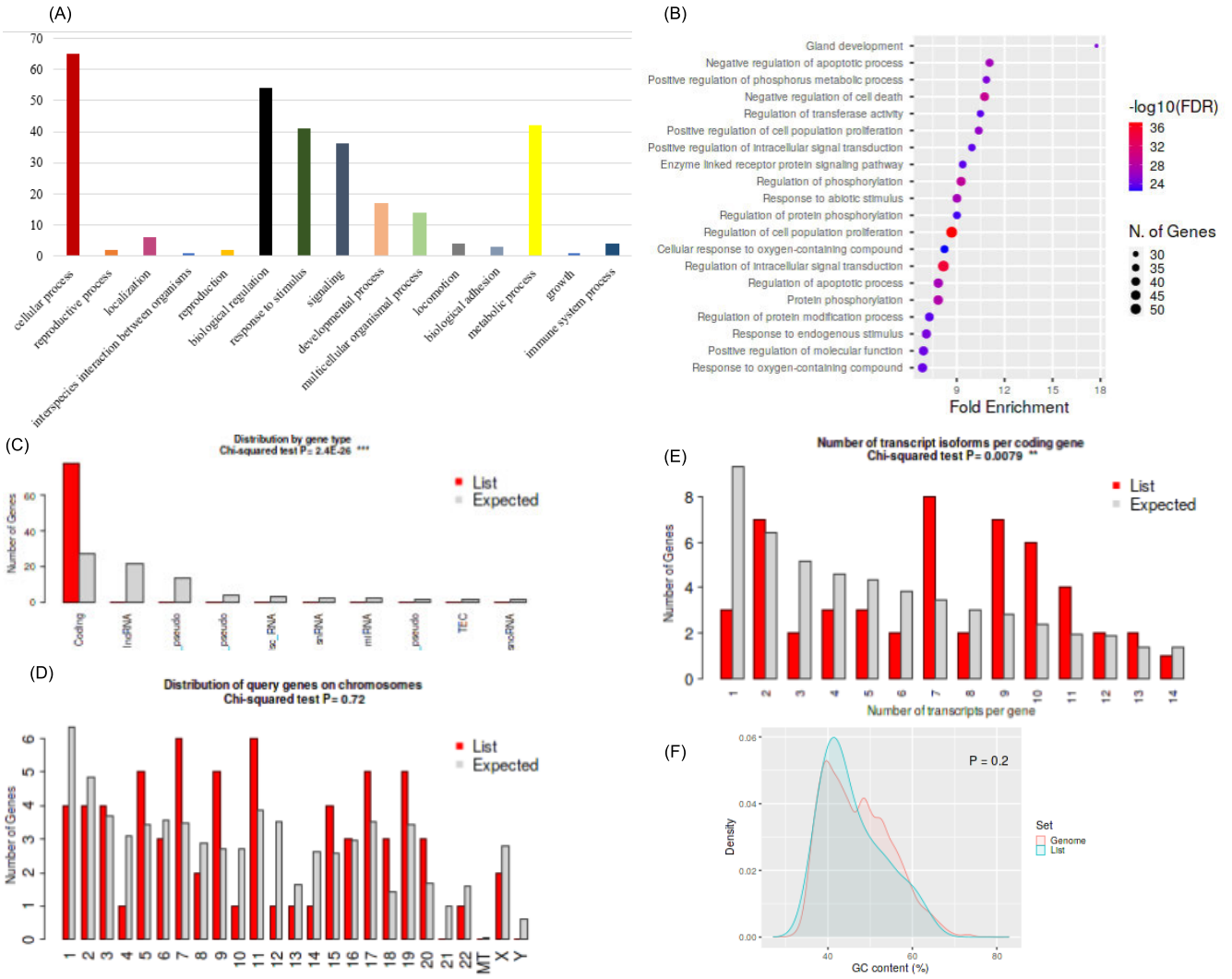


FIGURE 10. Enrichment analysis of all selected genes; (A): Barplot illustrates the gene ontology biological process of the selected genes;(B): Fold enrichment dotplot of the selected genes based on the -log10 FDR value;(C): Genes type distribution; (D): Distribution of genes on chromosomes barplot; (E): Barplot visualize the distribution of transcript isoforms per coding genes; (F):GC density of the selected genes to genome comparison.

total 4535 samples. Fig.13, summarizes the queries results of the most altered genes of each cancer type (Fig.13), where TP53, SMAD4, PTEN, ATM, CTNNB1, ERBB2, MTOR, and EGFR are the most altered genes with TP53 being altered in 58% of samples. As for BRCA ERBB2, PTEN, RB1, MTOR, SMAD4, MAP2K1 are the most altered genes across samples,with ERBB2 being altered in 13% of samples. With 37 % CDKN2A altered samples and 45 % PIK3CA altered samples, CDKN2A, PTEN, PICK3CA, EGFR, ERBB2, and MTOR have higher alteration percentages across LUSC samples.

Fig.13.B depicts the kelpen-meier plots of overall survival and progression free interval through time(months), where we selected two subsets from the selected hub genes set of each cancer: Those that appear in the OS experiment and those that appear in the PFI experiment. SMAD4, PTEN, CTNNB1, MTOR, and EGFR are the progression free selected hub genes in CRC, where, the plots shows a very poor PFI prognosis, in which population with CTNNB1 alteration dropping rapidly bellow 25% in a range of 15 months (2 years). In terms of overall survival, the selected hub genes are TP53, ATM, CTNNB1, ERBB2, and MTOR, as with PFI, the studied population shows a poor prognosis with the worst results for patients with CTNNB1, and ERBB2 alterations. BRCA patients with altered, RB1, SMAD4, MAP2K1, and/or ERBB2, have a good progression free prognosis that is stable over 50% across the time progression of the study. PTEN, RB1, MTOR, and MAP2K1 are the OS hub genes in breast cancer, where the population with altered MTOR genes have a better overall survival than those with alterations in the other genes. The studies on LUSC data sets define MTOR, PICK3CA, and PTEN as the PFI hub genes, where patients with altered PICK3CA have the poorest scored prognosis, the size sample of the MTOR alteration affects

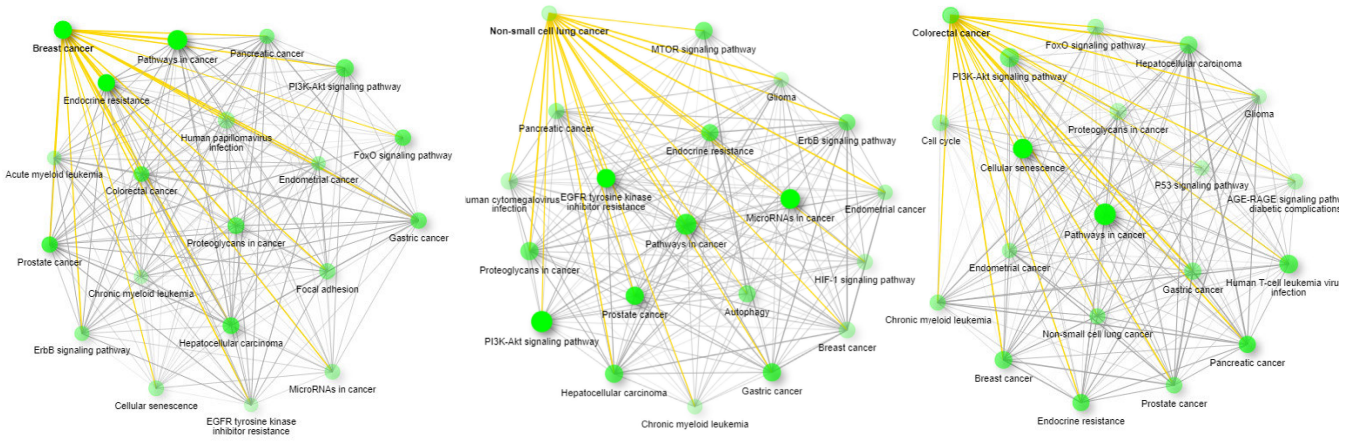


FIGURE 11. Selected biomarkers involvement in KEGG pathway in cancer, the selected biomarkers in these study are highlighted in yellow.

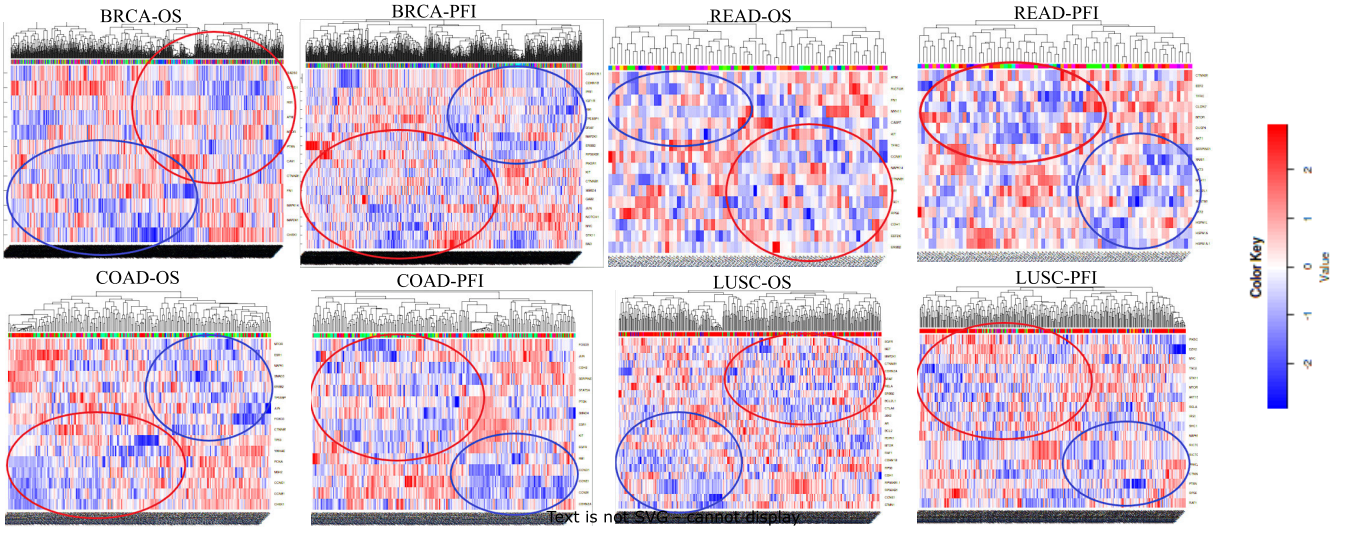


FIGURE 12. Heatmap illustrates the expression of genes to their corresponding clinical endpoint cancer type.

the observation of the overall PFI performance of patients with altered MTOR, yet as a shallow observation, we can assume that patients with altered MTOR may have a good progression free interval. CDNK2A, EGFR, ERBB2, and MTOR represent the hub genes in OS prognosis, where patients with ERBB2 and EGFR have a poor overall survival, followed by those with CDNK2A alteration, the results obtained from the patients with MTOR alteration may assess the prior hypothesis observed on PFI that squamous cell lung cancer with MTOR alteration may have a good PFI/OS prognosis.

Fig.13.C is the protein-protein interaction network of the selected hub genes for each cancer type and clinical endpoint, the networks show that in each case the selected genes are fully connected which implies the presence of interaction and a biological communication between them(a PPI network of all selected genes in each cancer is available in (fig 2 in appendix A).

C. MTOR/CTNNB/ERBB2/MAPK14/PTEN AS CANCER BIOMARKERS IN BRCA,LUSC, AND CRC

To visualise the shared genes between the cancer studied, we construct a venn diagram as shown in Fig.14, where the first venn diagram(a), extracts the shared genes between LUSC, BRCA, READ and COAD cancer, yielding three biomarker genes, ERBB2, MTOR, and CTNNB1. The second venn diagram merged the results of COAD and READ as CRC, resulting in two additional genes, PTEN and MAPK14. To better understand the impact of the selected biomarkers, we conducted an integrative study using Cbioportal, that merge the available studies on breast cancer, squamous cell lung cancer and colorectal cancer. The integration of the selected studies yielded a 15797 patient in which 22% of the population have an alteration in at least one of the shared genes (ERBB2, MTOR, CTNNB1, PTEN, and MAPK14). Fig.15 depicts the findings of the integrative study. The shared genes oncoprint represents the genomics alterations

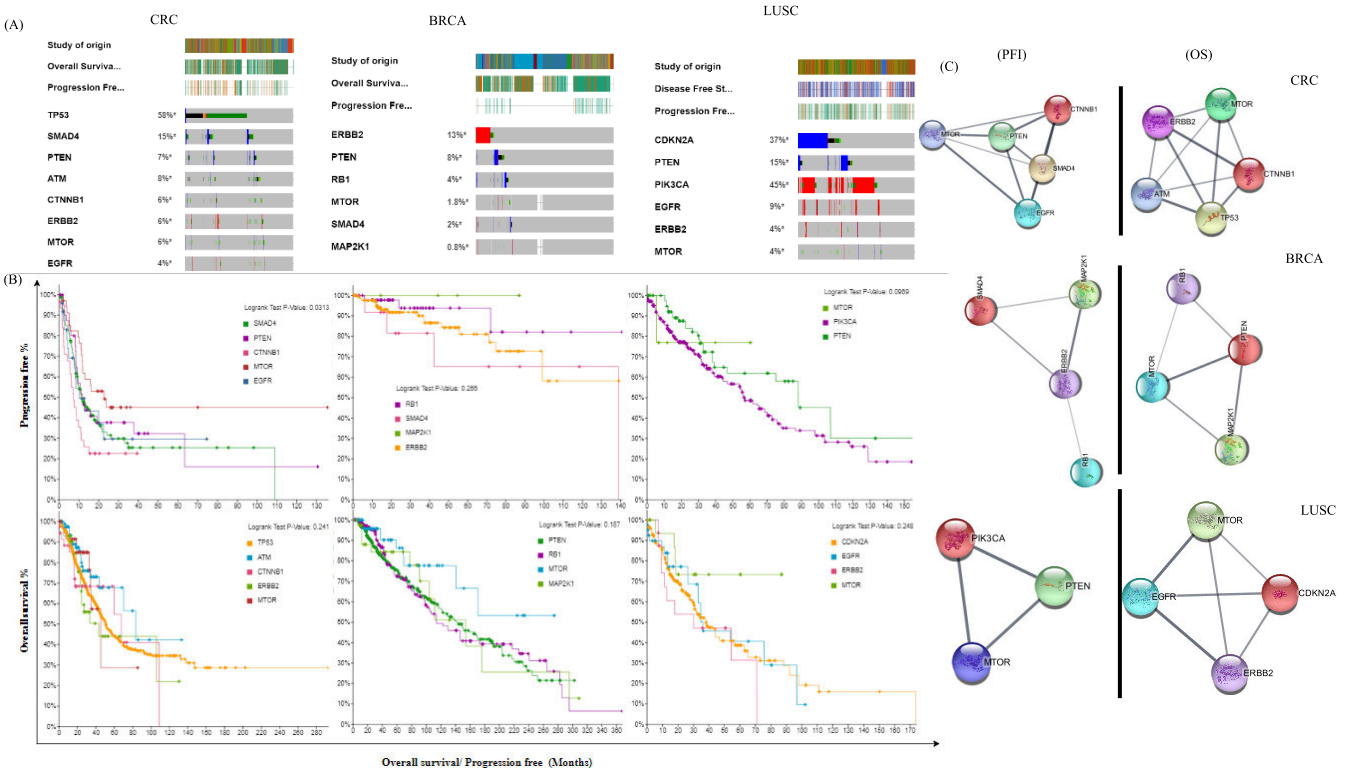


FIGURE 13. Define the cancer hub genes from the selected biomarkers, (A): Oncoprint of the most altered genes in CRC, BRCA, and LUSC; (B): progression free, and overall survival progress illustrating recoded events through months of samples with altered hub genes; (C): Protein-Protein interaction network of hub genes.

of the selected studies, with OS, and PFI track. ERBB2, and PTEN represents the most altered genes with a population frequency of 10%, 8% respectively. Fig.15.B shows the type of genomics alteration of the queried genes across all the cancer type, where mutations are the most present alteration types. The upset plot in Fig.15.C demonstrates the number of patients with altered genes in one or a set of the shared genes, where 74 patients having an alteration in PTEN, and ERBB, 57 in MTOR, and PTEN, 38 in CTNBB1, and PTEN, and 31 in ERBB2, and MTOR, while the other set have fewer than 30 patients. Figs.(15.(F), and 15.(G)), visualise the distribution of the altered genes across the queried cancer types and the TCGA cancer types used in the experiments. ERBB2, PTEN, and MTOR are the most observed altered genes in all cancer types, where 87% of the altered genes have a positive somatic status. The kaplan-meier plots illustrated in Figs. 15.(D), and 15.(E), where we notice that patients with CTNNB1 alteration have a considerably poor prognosis in both PFI, and OS, whereas, MTOR patients have a poor progression free, but a better overall survival. Patients with MAPK14 alteration have a good overall survival and progression free prognosis.

D. UNCOMMON CANCER VARIANTS WITH IMPORTANT CLINICAL IMPACT ARE SELECTED FROM THE HUB GENES

From the queried genes and cancer studies we collected the mutations profile statistics of the selected hub genes and the extracted shared genes, for each sample, where the set of mutations are classified as putative driver or variants with uncertain significance(VUS). Table 10, displays the number of mutations of each cancer collected only from the hub genes. Figs. (3,4,5,6,7 in appendix A) depict the lollipop chart of the distribution of the mutation of the shared genes on CRC, BRCA, and LUSC patients, from each lollipop chart we collected the variants with higher frequency on the study population. The analysis of the lollipop charts resulted in unveiling a set of variants that can be addressed as uncommon variants in CRC, BRCA, and LUSC cancer. We used Varsome and COSMIC to collect the interpretation of the selected variants focusing on their pathogenicity score and the ACMG classification (pathogenic, benign, uncertain significance), Table. 10 summarizes the analysis of the variants of each hub genes and the classification of the selected variants.

As shown in Table. 10, the genes with the highest frequency of pathogenic variants are TP53, ATM, and PTEN. Pathogenic variants can be targeted for cancer monitoring and a potential therapy. In our study, 15 of the 20 extracted variants were classified as pathogenic, two as likely pathogenic, and four as variants with uncertain significance that require careful monitoring. We conducted a selective literature review on published reviews, case reports, clinical and experimental publications to investigate the selected genes and the revealed mutation.

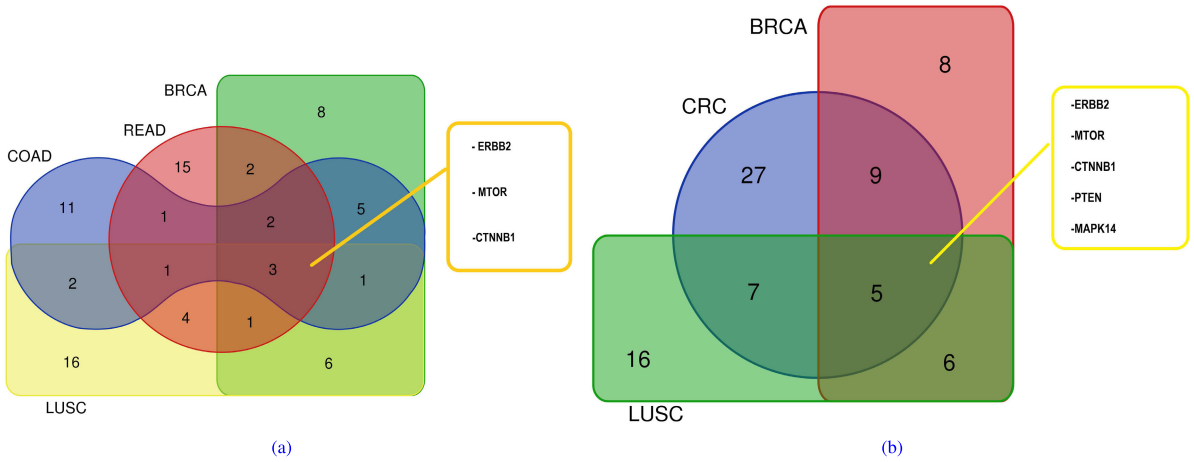


FIGURE 14. Venn diagram of shared hub genes in BRCA/COAD+READ(CRC)/LUSC.

TABLE 10. Selected variants.

Gene	Pathogenic	Uncertain significance	Benign	Common variants	
				ID	ACMG Classification
TP53	1221	2513	476	-p.R175H/G/C -p.R273H/C	-Pathogenic
SMAD4	178	545	314	- p.R361H/L/C	-Pathogenic
PIK3CA	355	312	142	-p.E545Q/K -p.E542V/K	-Pathogenic
CDKN2A	158	627	164	-p.D84Y/N/G -p.D108Y/N/G	-Pathogenic -Likely pathogenic
PTEN	734	690	312	-p.DK267R -p.R130Q -p.Q245*	-Pathogenic
CTNNB1	217	113	39	-p.S45P/F -p.S37C	-Pathogenic
ERBB2	70	111	100	-p.R678Q -p.L755S/W -p.G222C	-Uncertain significance -Pathogenic -Uncertain significance
MTOR	80	391	483	-p.S2215F/Y -p.I1973F -p.A2210P	-Pathogenic -Uncertain significance -Likely pathogenic
EGFR	191	857	403	-p.L861Q	-Pathogenic
ATM	2108	6043	2152	-p.R337H/C p.R3008H/C	-Uncertain significance -pathogenic

Among the set of extracted genes in this study, TP53 is the most altered gene in colorectal cancer. TP53 is referred to as the guardian of the genome, where its coded proteins function as tumour suppressor responsible of regulating cell division. TP53 interacts heavily with other genes by activating their role in fixing DNA damages, or by forcing cells with abnormal activities to undergo apoptosis [77]. According to Kim et al. [78], TP53 is the emost frequently expressed altered gene in patients with early onset coloroctal cancer (age <50 years), where the early functional loss of TP53, resulted to whole genome doubling and focal oncogene amplification. The most expressed along the TP53 altered samples The R175 H/C/G, and R273 H/C. According to Huang et al. [79] a co-mutation in patient with colorectal cancer was noticed between the IDH1/2 and TP53. Where a case was observe with (IDH1, TP53) mutations, on variants(p.R132C, p.R273C) with allele frequency (63%, 43%). the others reported another case with co-mutation of (IDH2, TP53), on variants (p.R140Q, p.R175H) with allele frequency (4.2%, 32%). In the study of Fassan et al. [80] on stage III, and IV CRC patients, shows that TP53 is the most mutated gene, with 8 out of 15 cases harboring the p.R273H variant.

Along with TP53, SMAD4, was also defined as a hub gene in CRC cancer,as well as a driver gene in breast cancer, SMAD4 is responsible of the chemical signals transmitting from cell surface to the nucleus, besides, it is associated with cell growth and proliferation. According to the suggestions of Fang et al. [81] intensive research, SMAD 4 is associated with over-all survival, progression-free survival/ recurrence-free survival, and clinico-pathological parameters (tumour site, disease stage, RAS status, lymph node metastasis, and mucinous status) in a study that enrolled more than 4394 patients with colorectal cancer. Lanauze et al. [82], conducted a recent study to investigate the role of Smad4 R361

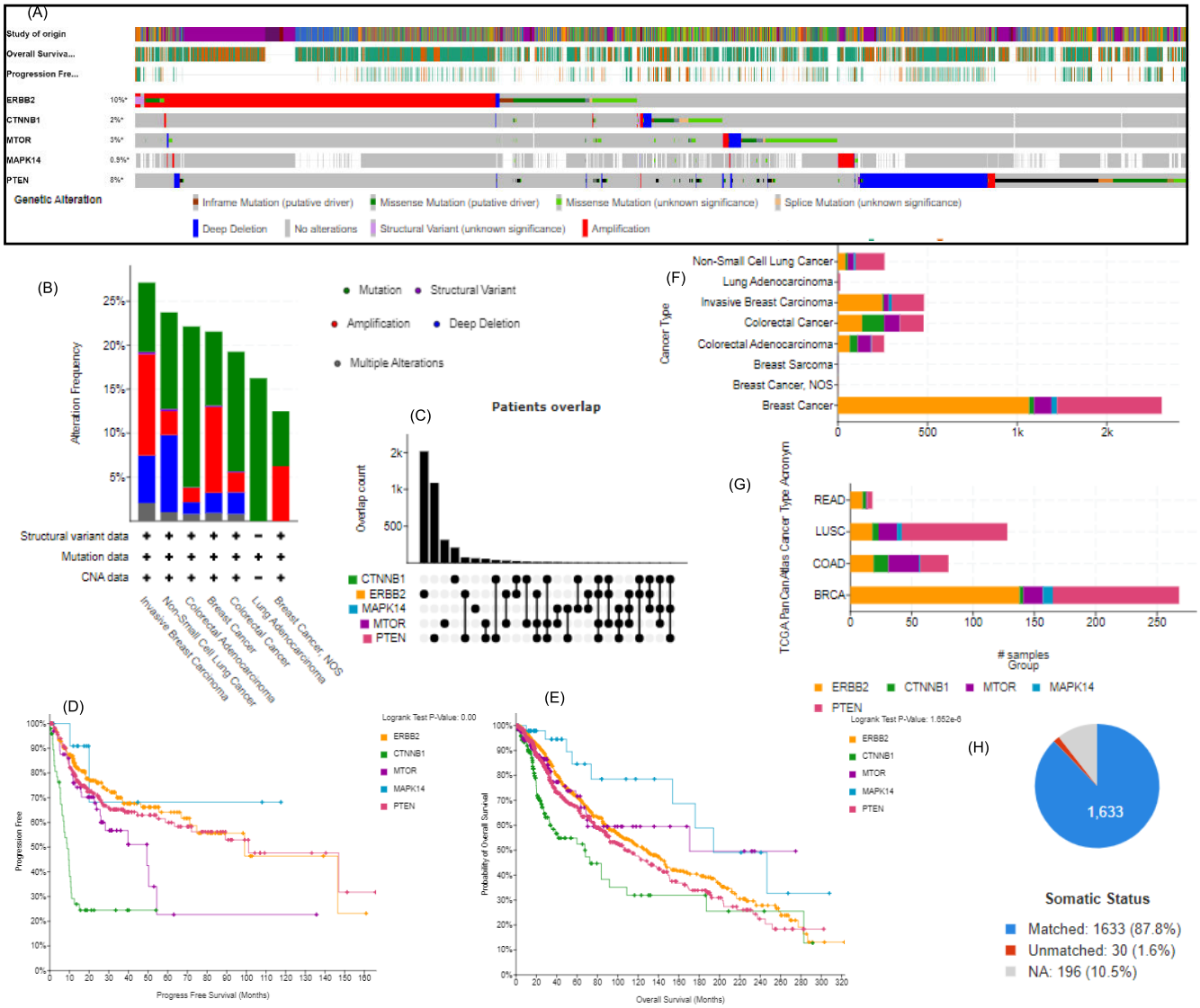


FIGURE 15. Evaluation of shared genes on cancer type Cbioportal resuts; (A): Oncoprint of the shared altered genes in BRCA, LUSC, and CRC; (B): Distribution of the type of the alterations across cancer studies;(C): Upset graph of the distribution of the shared genes mutations on patients; (D): Progression free survival of the hub genes on the cancer population;(E): Overall survival of the hub genes on the cancer population;(F): Distribution of shared genes alterations on cancer studies ;(G): Distribution of shared genes alterations on TCGAPan Cancer Atlas; (H): Somatic alteration ratio in selected patients.

variants in colorectal cancer by implanting two R361 variants in various Smad4 free CRC cell lines. The experimentation results unveiled that R361 missense mutation in CRC disturb the binding to endogenous p-smad2/smad-3, also the study suggest that SMAD4 variants are loss-of-function which may led to a poor prognosis in CRC patients (see Table. 13.(B); PFI). Though Frey et al. [83], adress SMAD4 mutation as a non preventer of epithelial-mesenchymal transition in CRC. Woo et al. [84], analyzed 250 patients with invasive ductal carcinoma using tissue microarray-based immunohistochemical assay, the experiments led to a significant observation of the low level of SMAD4 expression in early stage breast cancer and the promosing prospect for using SMAD4 as a targeted prognostic marker.

PTEN, is a suppressor gene, with a direct role in apoptosis, adhesion, migration, and angiogenesis, its alterations can lead to trigger, maintain, and mange different cancer types in various organisms [85], in this study PTEN have been selected as relevant cancer driver gene in the three cancer types. Salvatore et al. [85], reviewed research studies that investigate the potential role of PTEN as a predictive biomarker and/ or therapeutic target in colorectal cancer, citing that the loss of PTEN expression triggers the PI3K/AKt intracellular signalling, involved in CRC oncogenic mechanisms. Furthermore, the study addresses PTEN as the future clinical therapeutic target for CRC patients. According to Fusco et al. [86], PTEN:p.R130Q/G/* is the frequent expressed mutation in colorectal cancer, and PTEN

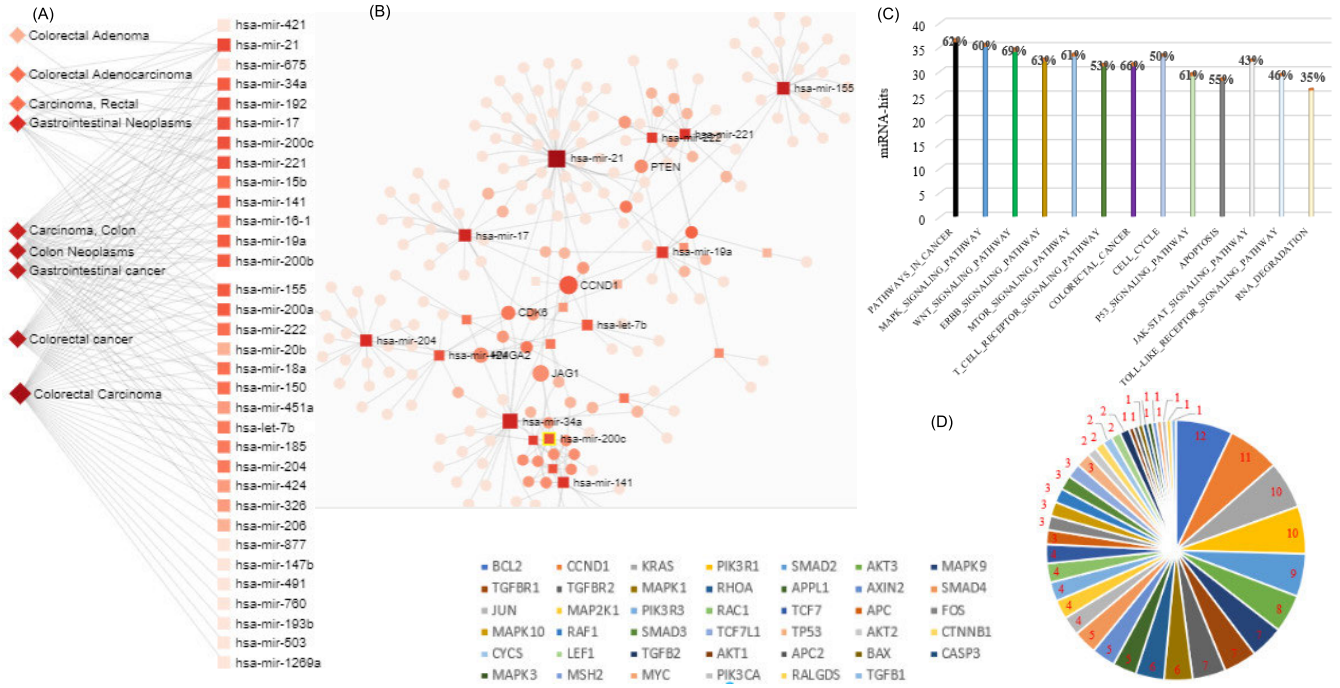


FIGURE 16. Analysis of the selected CRC miRNA; (A): Association of selected miRNA with CRC and gastrointestinal cancers; (B): miRNA-gene target network; (C): Frequency of the targeted genes in signalling pathways by the selected miRNA-hits; (D): Pie-chart visualisation of the targeted genes of CRC pathway.

may be of great help in clinical observation and monitoring in CRC patients. Carbognin et al. [87], comments on the clinical use of PTEN as an unfulfilled promise, yet it remains an interesting perspective, since PTEN alterations are clinically associated with breast cancer prognosis and treatment response. Kingston et al. [88], reported tow clinical case, one of a 50 years old women, with a history of cowden syndrome, diagnosed with a T3N3 breast cancer(ER+,PR-, and HER2-, gradeIII), the patient sequencing resulted to a p.R130Q PTEN heterozygous germline mutation.Whereas, the second case is a 37 years old women with clinically confirmed cowden syndrome, diagnosed with T3 breast cancer (ER+, PR+, HER2-, gradeII), the second case harbor a germline PTEN mutation p.T68G on chromosome 23X, and another somatic point mutation p.Y88X. The two cases where subjected to a AKT target therapy mainly based on capivasertib, the two patients with altered PTEN showed a dramatic response to the medicine and according to the authors' results and discussion. According to Gkountakos et al. [89], in lung cancer, the cases with PTEN protein loss showed the worst associated prognostic in terms of progression free and overall survival, also the genetic silencing of PTEN in in-vitro and in-vivo experimentation showed an upregulation of the EMT marker, while the induction of PTEN4 in an hypoxia-cultured cells in the experiment of Kohnoh et al. [90], induced the reversing of the EMT upregulation. Another shared gene is the ERBB2, the receptor tyrosine kinase 2 protein' coding gene, a member of the human epidermal growth factor family, for many years ERBB2 have been an active target in breast cancer and gastric cancer monitoring and therapy. However,recent emerging studies have addressed ERBB2 as an actionable target in solid tumours [91], including colorectal cancer [92], and lung cancer [93], [94]. A three-year study of a 469 Asian cancer cohort, reported in the manuscript of Lee et al. [95] demonstrates that ERBB2 alterations were found in 52 patients in which five samples are colorectal cancer patient and six are lung cancer.

CTNNB1 is the β-catenin coding gene responsible of cell addition, communication and co-activator in the oncogenic signaling Wnt pathway therefore CTNNB1 mutations posses an oncogenic role [96]. The hotspot mutation profile of CTNNB1 are expressed on Exon3, classified as tumorigenesis drivers [97]. In the recent studies CTNNB1 mutations have been associated with different cancer types like the p.SF45P/F in colorectal cancer, the p.S37C in lung cancer [98], [99] and the p.S37A in breast cancer [100]. Barggio et al. [101], demonstrated the apoptotic resistance of desmoid cells that harbor the p.SF45F β-catenin mutation with a downregulation of the RUNX3 protein expression. CTNNB1 mutations may result to rare cases, such as the case report of a never- smoked adolescent with a healthy family history diagnosed as the first reported pediatric case with lung aedenocarcinomas, where the molecular testing and sequencing panels found that CTNNB1 mutation is the only alteration in the 17 years old patient [102].

VI. ASSOCIATION OF SELECTED MI-RNAs WITH COLORECTAL CANCER (CRC)

In this section, we discuss the role of miRNAs as biomarkers in COAD and READ in light of the results obtained in step 3 and summarized in Table. 7. To that end, we performed an enrichment analysis and compiled a list of the most affected

pathways by the selected miRNA instances, as well as their associations with colorectal cancer.The graphics in Fig.16, are collected from miRNet data base, and mirPATH. where Fig.16.(A), was extracted from the miRNet networkviewer of the queried miRNAs that have a direct association with colorectal cancer or gastrointestinal cancer. The result of the extracted subnetwork mapped 33 out of 56 queried miRNAs, thus, the miRNAs with the highest interaction edges are hsa-mir-21, hsa-mir-129, hsa-mir,34a,hsa-mir-200a/b/c, hsa-mir17, and hsa-mir-221. Fig.16.(B), shows the targeted genes by the queried miRNAs, as post-transcriptional regulators, hsa-mir-21, hsa-mir-17, hsa-mir-155, hsa-mir-34a, and hsa-mir-221 are the the miRNAs instances with the highest gene interaction, with PTEN, CCND1, JAG1, and CDK5.

Fig.16.(C), shows the list of the affected pathways by the set of queried genes, where >30 miRNAs hits have a potential gene target implicated in colorectal cancer with more than 60% genes. The selected miRNAs targets the most cancer related pathways with a coverage of more than 50% coverage for the WNT _ signalling pathway, p53/mTOR/JAK-STAT signalling pahways. Fig.16.(D), depicts the distribution of the queried miRNAs and their potential targeted genes of the KEGG colorectal pathway, BCL2, CCND1, KRAS, and PIK3R1 are targeted by more than 10 miRNAs hits. fig 8 in appendix A, visualise the targeted genes in the KEGG colorectal cancer pathway, where we notice that mainly all the pathways that have critical biological process such as apoptosis are mainly affected by the selected miRNAs(The scheme was retrieved from the DIANA-mirPath results).

VII. CONCLUSION

The sciences of omics and system biology have revolutionized health and disease research; by incorporating artificial intelligence and bioinformatics tools, clinical applications of personalized medicine are becoming more and more prevalent. This study illustrates the significance of various artificial intelligence and computational methods and tools in omics-based cancer research. Incorporating biological knowledge databases, the paper proposes an upward integrative omics study using a new variant of the grey-wolf optimization algorithm and machine learning techniques, beginning with proteomics profiles and progressing to transcriptomics and micro-RNA profiles. The performance of the proposed CB-GWO is highly significant, and the use of a network of protein-protein interactions among the set of functional proteins significantly improved the selection of proteomics biomarkers, thereby increasing the classification rates using machine learning and deep learning models. This study identified MTOR, CTNNB1, ERBB2, MAPK14, and PTEN as hub genes with the potential occurence of alterations in colorectal cancer, breast cancer, and lung cancer. Hub genes along each cancer types and miRNAs in CRC, were defined and a set of variants with pathogenic classification were selected and defined as potential diagnostic and prognostic cancer biomarkers. As future work, we intend to investigate the potential therapeutic targets of the selected genes and hub genes, as well as extend the CB-GWO to be used on different biologically related feature selection problems and additional feature selection problems, as the item-item similarity matrix may replace the PPI network and the signalling pathway filtration. In the context of multi-disciplinary collaborative research, we intend to conduct an in-vitro evaluation of the selected shared gene and miRNA biomarkers and a potential in-vivo study.

APPENDIX A SUPPLEMENTARY MATERIALS

Table 1 lists the set of the selected genes using phase two and three of the proposal. fig 1, is the gene-gene interaction using GENEMANIA. fig 2 is the protein-protein interaction network of the selected proteins for each cancer types. Table 2 lists the statistics of the used cancer studies in the functional analysis section. Figs 3, 4, 5, 6, 7 are the lolipop charts of distribution and frequency of mutations of the selected shared hub genes. fig 8 highlights the targeted CRC colorectal cancer related genes by the selected miRNAs. Tables 3, and 4, depict the slected gene biomarkers role in gene ontology biological process and KEGG signalling pathways. All the supplementary figures and tables are available in the supplementary file.

ACKNOWLEDGMENT

The authors would like to express their appreciation to the University Constantine 2, Misc laboratory and CRBT members for their support. The authors would like to acknowledge the Princess Nourah Bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R196), Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia.

REFERENCES

[1] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Med.*, vol. 25, no. 1, pp. 44–56, 2019.
[2] P. Suwinski, C. Ong, M. H. T. Ling, Y. M. Poh, A. M. Khan, and H. S. Ong, "Advancing personalized medicine through the application of whole exome sequencing and big data analytics," *Frontiers Genet.*, vol. 10, p. 49, Feb. 2019.
[3] S. H. Shin, A. M. Bode, and Z. Dong, "Addressing the challenges of applying precision oncology," *NPJ Precis. Oncol.*, vol. 1, no. 1, pp. 1–10, Sep. 2017.
[4] C. Jean-Quartier, F. Jeanquartier, I. Jurisica, and A. Holzinger, "In silico cancer research towards 3R," *BMC Cancer*, vol. 18, no. 1, pp. 1–12, Dec. 2018.
[5] Q. Xiao, F. Zhang, L. Xu, L. Yue, O. L. Kon, Y. Zhu, and T. Guo, "High-throughput proteomics and AI for discovery," *Adv. Drug Del. Rev.*, vol. 176, Sep. 2021, Art. no. 113844.
[6] F. A. Castelli, G. Rosati, C. Moguet, C. Fuentes, J. Marrugo-Ramírez, T. Lefebvre, H. Volland, A. Merkoçi, S. Simon, F. Fenaille, and C. Junot, "Metabolomics for personalized medicine: The input of analytical chemistry from biomarker discovery to point-of-care tests," *Anal. Bioanal. Chem.*, vol. 414, no. 2, pp. 759–789, Jan. 2022.
[7] M. R. Karimi, A. H. Karimi, S. Abolmaali, M. Sadeghi, and U. Schmitz, "Prospects and challenges of cancer systems medicine: From genes to disease networks," *Briefings Bioinf.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab343.
[8] A. Statnikov, I. Tsamardinos, Y. Dosbayev, and C. F. Aliferis, "GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data," *Int. J. Med. Informat.*, vol. 74, nos. 7–8, pp. 491–503, Aug. 2005, doi: 10.1016/j.ijmedinf.2005.05.002.
[9] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015.
[10] A. Yuryev, "Gene expression profiling for targeted cancer treatment," *Expert Opinion Drug Discovery*, vol. 10, no. 1, pp. 91–99, Jan. 2015.

[11] M. Masuda and T. Yamada, "Signaling pathway profiling using reverse-phase protein array and its clinical applications," *Expert Rev. Proteomics*, vol. 14, no. 7, pp. 607–615, Jul. 2017.

[12] V. A. Hristova and D. W. Chan, "Cancer biomarker discovery and translation: Proteomics and beyond," *Expert Rev. Proteomics*, vol. 16, no. 2, pp. 93–103, Feb. 2019.

[13] M. Mann, C. Kumar, W.-F. Zeng, and M. T. Strauss, "Artificial intelligence for proteomics and biomarker discovery," *Cell Syst.*, vol. 12, no. 8, pp. 759–770, Aug. 2021.

[14] Y. Lu, S. Ling, A. M. Hegde, L. A. Byers, K. Coombes, G. B. Mills, and R. Akbani, "Using reverse-phase protein arrays as pharmacodynamic assays for functional proteomics, biomarker discovery, and drug development in cancer," *Seminars Oncol.*, vol. 43, no. 4, pp. 476–483, Aug. 2016.

[15] H. Rodriguez, J. C. Zenklusen, L. M. Staudt, J. H. Doroshow, and D. R. Lowy, "The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment," *Cell*, vol. 184, no. 7, pp. 1661–1670, Apr. 2021.

[16] F. Sanchez-Vega, "Oncogenic signaling pathways in the cancer genome atlas," *Cell*, vol. 173, no. 2, pp. 321–337, 2018.

[17] T. Ma and A. Zhang, "Integrate multi-omics data with biological interaction networks using multi-view factorization AutoEncoder (MAE)," *BMC Genomics*, vol. 20, no. S11, pp. 1–11, Dec. 2019.

[18] Y.-H. Chuang, S.-H. Huang, T.-M. Hung, X.-Y. Lin, J.-Y. Lee, W.-S. Lai, and J.-M. Yang, "Convolutional neural network for human cancer types prediction by integrating protein interaction networks and omics data," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, Oct. 2021.

[19] K. T. Chui, B. B. Gupta, H. R. Chi, V. Arya, W. Alhalabi, M. T. Ruiz, and C.-W. Shen, "Transfer learning-based multi-scale denoising convolutional neural network for prostate cancer detection," *Cancers*, vol. 14, no. 15, p. 3687, Jul. 2022.

[20] G. Nicora, F. Vitali, A. Dagliati, N. Geifman, and R. Bellazzi, "Integrated multi-omics analyses in oncology: A review of machine learning methods and tools," *Frontiers Oncol.*, vol. 10, p. 1030, Jun. 2020.

[21] N. Biswas and S. Chakrabarti, "Artificial intelligence (AI)-based systems biology approaches in multi-omics data analysis of cancer," *Frontiers Oncol.*, vol. 10, Oct. 2020, Art. no. 588221, doi: 10.3389/fonc.2020.588221.

[22] V. Coleto-Alcudia and M. A. Vega-Rodríguez, "Artificial bee colony algorithm based on dominance (ABCD) for a hybrid gene selection method," *Knowl.-Based Syst.*, vol. 205, Oct. 2020, Art. no. 106323.

[23] Y. Masoudi-Sobhanzadeh, H. Motieghader, Y. Omidi, and A. Masoudi-Nejad, "A machine learning method based on the genetic and world competitive contests algorithms for selecting genes or features in biological applications," *Sci. Rep.*, vol. 11, no. 1, pp. 1–19, Feb. 2021.

[24] A. K. Shukla, P. Singh, and M. Vardhan, "Gene selection for cancer types classification using novel hybrid metaheuristics approach," *Swarm Evol. Comput.*, vol. 54, May 2020, Art. no. 100661.

[25] S. Azadifar, M. Rostami, K. Berahmand, P. Moradi, and M. Oussalah, "Graph-based relevancy-redundancy gene selection method for cancer diagnosis," *Comput. Biol. Med.*, vol. 147, Aug. 2022, Art. no. 105766.

[26] S. Takahashi, K. Asada, K. Takasawa, R. Shimoyama, A. Sakai, A. Bolatkan, N. Shinkai, K. Kobayashi, M. Komatsu, S. Kaneko, J. Sese, and R. Hamamoto, "Predicting deep learning based multi-omics parallel integration survival subtypes in lung cancer using reverse phase protein array data," *Biomolecules*, vol. 10, no. 10, p. 1460, Oct. 2020.

[27] Z. Isik and M. E. Ercan, "Integration of RNA-Seq and RPPA data for survival time prediction in cancer patients," *Comput. Biol. Med.*, vol. 89, pp. 397–404, Oct. 2017.

[28] T. R. Kim, H.-H. Jeong, and K.-A. Sohn, "Topological integration of RPPA proteomic data with multi-omics data for survival prediction in breast cancer via pathway activity inference," *BMC Med. Genomics*, vol. 12, no. S5, pp. 1–14, Jul. 2019, doi: 10.1186/s12920-019-0511-x.

[29] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, Sep. 2013.

[30] J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J.-Y. Yang, B. M. Broom, R. G. W. Verhaak, D. W. Kane, C. Wakefield, J. N. Weinstein, G. B. Mills, and H. Liang, "TCPA: A resource for cancer functional proteomics data," *Nature Methods*, vol. 10, no. 11, pp. 1046–1047, Nov. 2013.

[31] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014.

[32] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, "Feature subset selection approach by gray-wolf optimization," in *Advances in Intelligent Systems and Computing*. Cham, Switzerland: Springer, 2015, pp. 1–13.

[33] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, "Feature subset selection approach by gray-wolf optimization," in *Advances in Intelligent Systems and Computing*. Cham, Switzerland: Springer, 2015, pp. 1–13.

[34] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[35] A. Parmar, R. Katariya, and V. Patel, "A review on random forest: An ensemble classifier," in *Proc. Int. Conf. Intell. Data Commun. Technol. Internet Things*. Cham, Switzerland: Springer, 2018, pp. 758–763.

[36] I. Rish, "An empirical study of the Naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, vol. 3, no. 22, 2001, pp. 41–46.

[37] P. H. Swain and H. Hauska, "The decision tree classifier: Design and potential," *IEEE Trans. Geosci. Electron.*, vol. GE-15, no. 3, pp. 142–147, Jul. 1977.

[38] R. E. Schapire, "Explaining adaboost," in *Empirical Inference*. Cham, Switzerland: Springer, 2013, pp. 37–52.

[39] Y. LeCun and Y. Bengio, *Convolutional Networks for Images, Speech, and Time Series*. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258.

[40] S. Judd, "On the complexity of loading shallow neural networks," *J. Complex.*, vol. 4, no. 3, pp. 177–192, Sep. 1988.

[41] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, Feb. 2015.

[42] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," 2017, *arXiv:1706.02262*.

[43] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[44] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, vol. 28, no. 3, Jun. 2013, pp. 1247–1255.

[45] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper feature selection," *Appl. Soft Comput.*, vol. 62, pp. 441–453, Jan. 2018.

[46] D. Rodrigues, L. A. M. Pereira, T. N. S. Almeida, J. P. Papa, A. N. Souza, C. C. O. Ramos, and X.-S. Yang, "BCS: A binary cuckoo search algorithm for feature selection," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2013, pp. 465–468.

[47] R. Y. M. Nakamura, L. A. M. Pereira, D. Rodrigues, K. A. P. Costa, J. P. Papa, and X.-S. Yang, "Binary bat algorithm for feature selection," in *Swarm Intelligence and Bio-Inspired Computation*, X.-S. Yang, Z. Cui, R. Xiao, A. H. Gandomi, and M. Karamanoglu, Eds. Oxford, U.K.: Elsevier, 2013, pp. 225–237.

[48] R. N. Khushaba, A. Al-Ani, and A. Al-Jumaily, "Differential evolution based feature subset selection," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[49] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1999.

[50] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles," in *Proc. Int. Workshop Data Mining Biomed. Appl.* Cham, Switzerland: Springer, 2006, pp. 106–115.

[51] A. Y. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 78.

[52] J. Rogers and S. Gunn, "Identifying feature relevance using a random forest," in *Subspace, Latent Structure and Feature Selection*, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Berlin, Germany: Springer, 2006, pp. 173–184.

[53] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sens. Actuators B, Chem.*, vol. 212, pp. 353–363, Jun. 2015.

[54] S. He, "MRMD2.0: A Python tool for machine learning with feature ranking and reduction," *Current Bioinf.*, vol. 15, no. 10, pp. 1213–1221, 2020.

[55] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[56] G. U. Yule, "On the methods of measuring association between two attributes," *J. Roy. Stat. Soc.*, vol. 75, no. 6, p. 579, May 1912.

[57] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, no. 1, pp. 1–22, Feb. 2021.
[58] D. Chicco, V. Starovoitov, and G. Jurman, "The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment," *IEEE Access*, vol. 9, pp. 47112–47124, 2021.
[59] N. W. S. Wardhani, M. Y. Rochayani, A. Iriany, A. D. Sulistyono, and P. Lestantyo, "Cross-validation metrics for evaluating classification performance on imbalanced data," in *Proc. Int. Conf. Comput., Control, Informat. Appl. (IC3INA)*, Oct. 2019, pp. 14–18.
[60] S. Carbon, "The gene ontology resource: Enriching a GOld mine," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, Dec. 2020.
[61] S. X. Ge, D. Jung, and R. Yao, "ShinyGO: A graphical gene-set enrichment tool for animals and plants," *Bioinformatics*, vol. 36, no. 8, pp. 2628–2629, Dec. 2019.
[62] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.
[63] S. X. Ge, E. W. Son, and R. Yao, "IDEP: An integrated web application for differential expression and pathway analysis of RNA-seq data," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–24, Dec. 2018.
[64] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signaling*, vol. 6, no. 269, pp. 1–20, Apr. 2013.
[65] C. Kopanos, V. Tsiolkas, A. Kouris, C. E. Chapple, M. A. Aguilera, R. Meyer, and A. Massouras, "VarSome: The human genomic variant search engine," *Bioinformatics*, vol. 35, no. 11, pp. 1978–1980, Oct. 2018.
[66] N. Bindal, S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, C. Kok, M. Jia, S. Bamford, C. Cole, S. Ward, J. Teague, M. R. Stratton, P. Campbell, and A. P. Futreal, "COSMIC: The catalogue of somatic mutations in cancer," *Genome Biol.*, vol. 12, no. 1, p. P3, 2011.
[67] L. Chang, G. Zhou, O. Soufan, and J. Xia, "MiRNet 2.0: Network-based visual analytics for miRNA functional analysis and systems biology," *Nucleic Acids Res.*, vol. 48, no. W1, pp. W244–W251, Jun. 2020.
[68] T.-P. Lu, C.-Y. Lee, M.-H. Tsai, Y.-C. Chiu, C. K. Hsiao, L.-C. Lai, and E. Y. Chuang, "MiRSystem: An integrated system for characterizing enriched functions and pathways of MicroRNA targets," *PLoS ONE*, vol. 7, no. 8, Aug. 2012, Art. no. e42390.
[69] I. S. Vlachos, K. Zagganas, M. D. Paraskevopoulou, G. Georgakilas, D. Karagkouni, T. Vergoulis, T. Dalamagas, and A. G. Hatzigeorgiou, "DIANA-miRPath v3.0: Deciphering microRNA function with experimental support," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W460–W466, May 2015, doi: 10.1093/nar/gkv403.
[70] T. Tian, X. Li, and J. Zhang, "MTOR signaling in cancer and mTOR inhibitors in solid tumor targeting therapy," *Int. J. Mol. Sci.*, vol. 20, no. 3, p. 755, Feb. 2019.
[71] X. Li, D. Dai, B. Chen, H. Tang, X. Xie, and W. Wei, "Efficacy of PI_3K/AKT/mTOR pathway inhibitors for the treatment of advanced solid cancers: A literature-based meta-analysis of 46 randomised control trials," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0192464.
[72] I. Vannini, F. Fanini, and M. Fabbri, "Emerging roles of microRNAs in cancer," *Current Opinion Genet. Develop.*, vol. 48, pp. 128–133, Feb. 2018.
[73] S. Shirjang, B. Mansoori, S. Asghari, P. H. G. Duijf, A. Mohammadi, M. Gjerstorff, and B. Baradaran, "MicroRNAs in cancer cell death pathways: Apoptosis and necroptosis," *Free Radical Biol. Med.*, vol. 139, pp. 1–15, Aug. 2019.
[74] H. J. Kim, M. S. Jeong, and S. B. Jang, "Molecular characteristics of RAGE and advances in small-molecule inhibitors," *Int. J. Mol. Sci.*, vol. 22, no. 13, p. 6904, Jun. 2021, doi: 10.3390/ijms22136904.
[75] A. Mollace, M. L. Coluccio, G. Donato, V. Mollace, and N. Malara, "Cross-talks in colon cancer between RAGE/AGEs axis and inflammation/immunotherapy," *Oncotarget*, vol. 12, no. 13, pp. 1281–1295, Jun. 2021.
[76] X. Zhou, N. Lin, M. Zhang, X. Wang, Y. An, Q. Su, P. Du, B. Li, and H. Chen, "Circulating soluble receptor for advanced glycation end products and other factors in type 2 diabetes patients with colorectal cancer," *BMC Endocrine Disorders*, vol. 20, no. 1, pp. 1–7, Nov. 2020.
[77] G. P. Zambetti, "The p53 mutation 'gradient effect' and its clinical implications," *J. Cellular Physiol.*, vol. 213, no. 2, pp. 370–373, 2007, doi: 10.1002/jcp.21217.
[78] J. E. Kim, J. Choi, C.-O. Sung, Y. S. Hong, S. Y. Kim, H. Lee, T. W. Kim, and J.-I. Kim, "High prevalence of TP53 loss and whole-genome doubling in early-onset colorectal cancer," *Experim. Mol. Med.*, vol. 53, no. 3, pp. 446–456, Mar. 2021.
[79] J. Huang, L.-H. Tseng, V. Parini, P. M. Lokhandwala, A. Pallavajjala, E. Rodriguez, R. Xian, L. Chen, C. D. Gocke, J. R. Eshleman, and M.-T. Lin, "IDH_1 and IDH_2 mutations in colorectal cancers," *Amer. J. Clin. Pathol.*, vol. 156, no. 5, pp. 777–786, Oct. 2021.
[80] M. Fassan, L. Vianello, D. Sacchi, G. N. Fanelli, G. Munari, M. Scarpa, R. Cappellesso, F. Loupakis, C. Lanza, R. Salmaso, C. Mescoli, N. Valeri, M. Agostini, E. D'Angelo, S. Lonardi, S. Pucciarelli, N. Veronese, C. Luchini, and M. Rugge, "Assessment of intratumor immune-microenvironment in colorectal cancers with extranodal extension of nodal metastases," *Cancer Cell Int.*, vol. 18, no. 1, pp. 1–9, Sep. 2018.
[81] T. Fang, T. Liang, Y. Wang, H. Wu, S. Liu, L. Xie, J. Liang, C. Wang, and Y. Tan, "Prognostic role and clinicopathological features of SMAD4 gene mutation in colorectal cancer: A systematic review and meta-analysis," *BMC Gastroenterol.*, vol. 21, no. 1, pp. 1–12, Dec. 2021.
[82] C. B. Lanauze, P. Sehgal, K. Hayer, M. Torres-Diz, J. A. Pippin, S. F. A. Grant, and A. Thomas-Tikhonenko, "Colorectal cancer-associated $Smad_4$ R361 hotspot mutations boost Wnt/β-catenin signaling through enhanced $Smad_4$–LEF_1 binding," *Mol. Cancer Res.*, vol. 19, no. 5, pp. 823–833, May 2021.
[83] P. Frey, A. Devisme, K. Rose, M. Schrempp, V. Freihen, G. Andrieux, M. Boerries, and A. Hecht, "$SMAD_4$ mutations do not preclude epithelial–mesenchymal transition in colorectal cancer," *Oncogene*, vol. 41, no. 6, pp. 824–837, Feb. 2022.
[84] J.-S. Woo, M. S. Chung, and S. S. Paik, "Clinicopathological significance of SMAD4 expression in breast cancer," *J. Breast Disease*, vol. 7, no. 2, pp. 52–58, Dec. 2019.
[85] L. Salvatore, M. A. Calegari, F. Loupakis, M. Fassan, B. D. Stefano, M. Bensi, E. Bria, and G. Tortora, "PTEN in colorectal cancer: Shedding light on its role as predictor and target," *Cancers*, vol. 11, no. 11, p. 1765, Nov. 2019.
[86] N. Fusco, E. Sajjadi, K. Venetis, G. Gaudioso, G. Lopez, C. Corti, E. G. Rocco, C. Criscitiello, U. Malapelle, and M. Invernizzi, "PTEN alterations and their role in cancer management: Are we making headway on precision medicine?" *Genes*, vol. 11, no. 7, p. 719, Jun. 2020.
[87] L. Carbognin, F. Miglietta, I. Paris, and M. V. Dieci, "Prognostic and predictive implications of PTEN in breast cancer: Unfulfilled promises but intriguing perspectives," *Cancers*, vol. 11, no. 9, p. 1401, Sep. 2019.
[88] B. Kingston, C. Bailleux, S. Delaloge, G. Schiavon, V. Scott, M. Lacroix-Triki, T. H. Carr, I. Kozarewa, H. Gevensleben, Z. Kemp, A. Pearson, N. Turner, and F. André, "Exceptional response to AKT inhibition in patients with breast cancer and germline PTEN mutations," *JCO Precis. Oncol.*, no. 3, pp. 1–7, Dec. 2019, doi: 10.1200/po.19.00130.
[89] A. Gkountakos, G. Sartori, I. Falcone, G. Piro, L. Ciuffreda, C. Carbone, G. Tortora, A. Scarpa, E. Bria, M. Milella, R. Rosell, V. Corbo, and S. Pilotto, "PTEN in lung cancer: Dealing with the problem, building on new knowledge and turning the game around," *Cancers*, vol. 11, no. 8, p. 1141, Aug. 2019.
[90] T. Kohnoh, N. Hashimoto, A. Ando, K. Sakamoto, S. Miyazaki, D. Aoyama, M. Kusunose, M. Kimura, N. Omote, K. Imaizumi, T. Kawabe, and Y. Hasegawa, "Hypoxia-induced modulation of PTEN activity and EMT phenotypes in lung cancers," *Cancer Cell Int.*, vol. 16, no. 1, p. 33, Apr. 2016.
[91] J. Subramanian, A. Katta, A. Masood, D. R. Vudem, and R. K. Kancha, "Emergence of $ERBB_2$ mutation as a biomarker and an actionable target in solid cancers," *Oncologist*, vol. 24, no. 12, pp. e1303–e1314, Dec. 2019.
[92] N. Tavberidze and W. Zhang, "HER_2 ($ERBB_2$) alterations in colorectal cancers," *Human Pathol. Rep.*, vol. 28, Jun. 2022, Art. no. 300628.
[93] X. Wei, X. Gao, X. Zhang, J. Yang, Z. Chen, Y. Wu, and Q. Zhou, "Mutational landscape and characteristics of $ERBB_2$ in non-small cell lung cancer," *Thoracic Cancer*, vol. 11, no. 6, pp. 1512–1521, Jun. 2020.
[94] J. Ni, X. Si, and L. Zhang, "Non-small-cell lung cancer with $ERBB_2$ mutation in non-tyrosine kinase domain benefits from pyrotinib: A case report," *Thoracic Cancer*, vol. 12, no. 8, pp. 1244–1247, Apr. 2021.

[95] J. Lee, A. Franovic, Y. Shiotsu, S. T. Kim, K.-M. Kim, K. C. Banks, V. M. Raymond, and R. B. Lanman, "Detection of $ERBB_2$ (HER_2) gene amplification events in cell-free DNA and response to anti-HER_2 agents in a large Asian cancer patient cohort," *Frontiers Oncol.*, vol. 9, p. 212, Apr. 2019.
[96] N. Krishnamurthy and R. Kurzrock, "Targeting the Wnt/beta-catenin pathway in cancer: Update on effectors and inhibitors," *Cancer Treatment Rev.*, vol. 62, pp. 50–60, Jan. 2018.
[97] C. Gao, Y. Wang, R. Broaddus, L. Sun, F. Xue, and W. Zhang, "Exon 3 mutations of $CTNNB_1$ drive tumorigenesis: A review," *Oncotarget*, vol. 9, no. 4, p. 5492, 2018.
[98] K. Sewoon and J. Sunjoo, "Mutation hotspots in the β-catenin gene: Lessons from the human cancer genome databases," *Molecules Cells*, vol. 42, no. 1, pp. 8–16, 2019.
[99] C. Zhou, H. Jin, W. Li, R. Zhao, and C. Chen, "$CTNNB_1$ S37C mutation causing cells proliferation and migration coupled with molecular mechanisms in lung adenocarcinoma," *Ann. Transl. Med.*, vol. 9, no. 8, p. 681, Apr. 2021.
[100] J. Xu, Y. Chen, D. Huo, A. Khramtsov, G. Khramtsova, C. Zhang, K. H. Goss, and O. I. Olopade, "β-catenin regulatesc-MycandCDKN1Aexpression in breast cancer cells," *Mol. Carcinogenesis*, vol. 55, no. 5, pp. 431–439, May 2016.
[101] D. Braggio, A. Zewdu, P. Londhe, P. Yu, G. Lopez, K. Batte, D. Koller, F. C. C. de Faria, L. Casadei, A. M. Strohecker, D. Lev, and R. E. Pollock, "β-catenin S45F mutation results in apoptotic resistance," *Oncogene*, vol. 39, no. 34, pp. 5589–5600, Aug. 2020.
[102] H. Wu, Q. Ye, D. Razzano, O. Tugal, J. Rosenblum, T. Weigel, and M. Zhong, "Primary lung cribriform adenocarcinoma with squamoid morules harboring somatic CTNNB1 mutation in a never-smoked healthy adolescent," *Pediatric Develop. Pathol.*, vol. 23, no. 6, pp. 472–475, Dec. 2020.

IMENE ZENBOUT received the bachelor's and master's degrees in computer science and software engineering from Ferhat Abess University, Setif, Algeria, in 2014 and 2016, respectively. She is currently pursuing the Ph.D. degree in bionformatics with the Computer Science Department, Faculty of New Information and Communication Technologies, Universite Adelhamid Mehri Constantine 2. Her research interests include bioinformatics, next generation sequencing, artificial intelligence application in cancer research, biomarker discovery, and drug discovery.

ABDELKRIM BOURAMOUL received the joint Ph.D. degree in computer science from the University of Paris 11 and Constantine 2 University, in 2011, and the University Habilitation (H.D.R.) degree from the University of Constantine 2, in April 2017. He is currently a Research Professor at the Computer Science Department, Faculty of New Information and Communication Technologies, Constantine 2 University, where he is also a Founding Member of the MISC Research Laboratory. He has published many scientific papers in internationally renowned journals and conferences. His research interests include the areas of artificial intelligence, affective computing, virtual and augmented reality, information retrieval, and visualization in a big data context. He has directed many Ph.D. theses already defended, and he is also supervising nine Ph.D. thesis projects. He collaborates with renowned researchers in Algeria and abroad, particularly in France, Germany, Spain, South Korea, and Slovakia.

SOUHAM MESHOUL received the Ph.D. degree from Mentouri University Constantine, Algeria. She is currently a Full Professor at the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, where she also works as the Director of the Master of Science in Data Science Program with a previous work experience at Constantine 2 University and King Saud University. Her research interests include computational intelligence, optimization, machine learning, and data science.

MOUNIRA AMRANE received the Ph.D. degree in medicine from the Faculty of Medicine, Constantine, Algeria, in 1996, the D.E.M.S. degree in clinical biochemistry and molecular genetics in Constantine, in 1999, the D.E.S.M. degree in clinical biochemistry and molecular genetics, in 2008, and the University degree in next generation sequencing genetic diseases: experimental approaches and bioinformatics tools–Dijon–France University, in 2020. She is currently a Professor in biochemistry and molecular genetics at University Ferhat Abbes Setif 1, Algeria. She is also the Head of Department of the Central Laboratory (Cancer Center Mokhtari Abdelghani—Setif). In 2014, she became a Class A Lecturer, and a Full Professor, in 2017. She aims to develop molecular and protein marker diagnostics, therapeutic monitoring, and cancer prognosis.