

Hybrid Machine Learning And Deep Learning Models For Multi-Omics Data Fusion In Bioinformatics

Raghad K. Mohammed

Department of Basic Science, College of Dentistry, University of Baghdad, Iraq

Abstract

Since multi-omics data (high-dimensional data, including many variables, like genomics, transcriptomics and proteomics) are high-dimensional (have many variables), heterogeneous (are data of different types), and interdependent (variables can affect each other), there is a significant challenge in integrating various data types in a synergistic manner (without, it may not be feasible). This paper is an attempt to tackle the problem of predicting drug sensitivity by offering a hybrid machine learning (ML, algorithms that learn patterns by analyzing data) and deep learning (DL, a type of ML based on neural network architectures to learn patterns of data) model of not only predicting drug sensitivity, but also using Genomics of Drug Sensitivity in Cancer (GDSC) data. The method involves applying sophisticated preprocessing, strict feature selection techniques (Random Forest, tree-based algorithm; LASSO regression, feature selection that chooses features by pruning irrelevant ones; Mutual Information which is a dependency measure between variables) as well as a hybrid model design favoring classical ML models (Random Forest, SVM--Support Vector Machine) over the deep learning style (LSTM-Long short-term memory networks, a sequence-based version of recurrent neural network). The metrics of the evaluation give a very high R^2 (coefficient of determination, a measure of model fit) of 0.991, and the evaluations are very small MAE (Mean absolute error: square root of average squared error), RMSE (Root mean squared error: square root of average squared error), and MSE (Mean squared error: square root of average squared error), resulting in approximately 12 percent higher accuracy than conventional approximations: Ridge Regression, XGBoost (Extreme gradient boosting), and Random Forest alone. This also shows the strength (capacity to work with diverse circumstances), generalizability (capacity to be used in new information), and enhanced predictive capabilities of the hybrid method of modeling nonlinear associations in multi-omics information, thus enhancing precision medicine (health care customized to patients in terms of genetic, environmental, and lifestyle determinants) uses.

Keywords: Deep learning; Machine Learning; Multi-Omics; Bioinformatics; LSTM; Regression.

1. Introduction

With high-throughput technologies such as next-generation sequencing, multi-omics data—including genomics, transcriptomics, proteomics, metabolomics, and epigenetics—have rapidly expanded. These data types together provide information about biological processes at the cell or tissue level and illuminate the roles of smaller molecules. Integrating multi-omics data offers unprecedented promise for understanding variability in disease progression and cancer subtypes, as highlighted in several studies [1][2][3].

Integrating multimodal omics datasets—such as genomics, transcriptomics, and proteomics—remains a crucial computational problem in biology due to their high dimensionality, heterogeneity, and inherent noise¹. Analyzing only one omics type in isolation lacks the power to capture interactions across molecular levels. Thus, simply identifying genomic mutations (known cause) is insufficient; instead, understanding their roles in disease phenotypes (observable characteristics) and the influence of environmental modifiers (external factors) on molecular phenotypes is essential to determine how diseases can be switched on or off [4][5].

Machine learning (ML) Computer algorithm that acquires pattern by learning on data and deep learning (DL), which is a subdivision of the former based on layered neural networks, has been used to process, analyze, and interpret multi-omics data [6]. Classification and predictive performance Traditionally, ML algorithms including, SVM (Support Vector Machine), random forests (classifications of decision trees), and k-nearest neighbors (k-NN, which categorizes data based on the k most similar) have demonstrated impressive classification and predictive accuracy on single-omics studies [7]. There are however a few challenges associated with their application to multi-omics data such as those data are complex and heterogeneous. Moreover, the methods also tend to be very resource intensive in terms of feature engineering (hand choosing or converting variables) and ad hoc treatment of missing data, data noise, and varying measurement scales across layers of omics. Moreover, classical ML models might not be able to learn nonlinear interactions and higher-order interactions across multiple types of omics [8][9][10][11].

However, deep learning (DL, models with multiple computational layers) has shown great promise in overcoming these challenges; it may enable automatic representation learning (letting the model discover features) from large, high-dimensional datasets. DNN (Deep Neural Networks), CNN (Convolutional Neural Networks, which excel at grid-like data such as images), and RNN (Recurrent Neural Networks, suited for sequence data) have gained importance in fields such as computer vision, natural language processing, and genomics due to their capacity to produce meaningful results and recognize complicated relations between features without the requirement of manual feature engineering. Deep learning models address this issue by learning holistic (overall) interactions across multi-omics data types that are important for gaining meaningful insights but may be ignored in traditional methods. Although these models are successful, deep learning approaches often require large amounts of data and are not easily interpretable (difficult to understand the reasoning behind model decisions), which limits understanding of how the model is making decisions and what biological insights can be drawn [3][16][17].

In recent times, one of the strongest solutions for multi-omics data integration was developed using a hybrid machine learning (combining multiple computational approaches) and deep learning model [18]. Hybrid models provide the best of both worlds for researchers, achieving accurate predictive results while maintaining model interpretability (the ability to explain how predictions are made). Feature selection (identifying important variables), normalization (making data comparable), and dimension reduction (reducing the number of variables analyzed) are areas where machine learning methods can help, and deep learning models are known to excel at capturing high-order (complex) interactions and representing complex features across multiple omics layers. This approach provides less rigid, clearer, and more interpretable analysis of multi-omics data [19][20][21].

There is increased research interest in hybrid ML/DL algorithms that integrate multimodal omics data for applications such as disease classification, biomarker screening, and patient stratification. Hybrid frameworks are advantageous because they inherit nonlinear representation learning from deep learning methods and computational efficiency and interpretability from machine learning. Further, hybrid models better address issues like missing values, noise, and heterogeneity, making them effective for multi-omics integration [22][23][18].

Here is where we provided a comprehensive summary of recent advancements in hybrid machine learning and deep learning that enable embedding networks from multi-omics data in bioinformatics. Herein, we describe several hybrid approaches for integrating and

analyzing multiple omics data types. Here, we emphasize how hybrid models like SLiM can leverage multiple layers of information to address the high dimensionality, noise, and complexity issues associated with multi-omics data, as well as to gain valuable insights into biological systems and disease mechanisms. We also analyze the limitations and challenges of hybrid methods, including model interpretability, generalization to unknown datasets, and scalability. Lastly, we conclude with challenges and prospects toward the application of machine learning and deep learning models in multi-omics bioinformatics for personalized medicine, cancer research, and beyond.

This contribution captures the following in this work:

1. It is also seen as proposing a novel hybrid design that amalgamates classical machine learning models, e.g. Random Forests, with deep learning networks, specifically Long Short-Term Memory (LSTM), to enhance drug sensitivity prediction.
2. The current study employs a strong feature selection technique that combines the random forests, Lasso regression and mutual information technique to deal with the large dimensionality and complexity of multi-omics data.
3. It addresses the problem of synergistic combination of heterogeneous, interdependent multi-omics information, which makes it possible to extract biologically useful features.
4. Our analysis identifies the efficiency and applicability of a mixed methodology in modeling multifaceted, nonlinear relationships, which are expected to have a great influence on the development of precision medicine.

1. Related Work

An emerging study (2024) aimed to utilize deep learning for the prediction and prognosis of TNBC subtypes by creating models enriched with multi-omics data, including mRNA, miRNA, gene mutations, DNA methylation, and MRI images. The accuracy of single-omics models was lower than that of the multi-omics model in cross-validation (98.0%). An additional strength of the model, based on merging multi-omics data and using Bayesian optimization to optimize statistical performance and transfer learning, was emphasized. Nonetheless, the initial validation was conducted on a dataset separate from the training cohort, with lower performance indicating the need for further optimization of the MRI radionics model [24]. Separate approaches have been taken by Mkhedlidze et al, who proposed using Multiple Kernel Learning (MKL) methods (2024) for the integrated analysis of multi-omics data with several kernel fusion strategies. MKL-based models achieved competitive results compared to more complex, state-of-the-art methods for data mining in genomics. They further studied deep learning architectures for kernel fusion and showed that MKL approaches, while simpler than advanced methods, outperform or achieve comparable performance to advanced methods for multi-omics binary classification problems. MKL is proposed as a cross-platform, simple integrative solution [25].

In this regard, a Multimodal Functional Deep Learning (MFDL) method has been proposed to perform phenomic analysis of high-dimensional multi-omics data and model the complex relationships between omics variants and disease phenotypes. The method is a functional data method that leverages multiple neural networks to address high dimensionality. MFDL demonstrated high predictive accuracy and robustness to noise across simulation studies and real-data applications. The approach can capture relationships among various types of omics data and provides a versatile solution for integrating and analyzing multi-omics data [26]. Several approaches to integrate transcriptomics, proteomics, and metabolomics data are presented in a paper (2024). The authors identified three principal integration approaches:

merged omics integration, correlation-based, and machine learning techniques. They discussed the advantages of combining omics data types to achieve a more integrated view of biological systems and observed that using machine-learning methods for integration could detect complex patterns and relationships that are not discoverable from any single-omics analysis. However, the paper did not explicitly address limitations [27].

Tsagiopoulou and Gut (2024) described machine-learning approaches for integrating multi-omics data in Chronic Lymphocytic Leukemia (CLL). In this work, various machine learning techniques, including PCA, SVM, and hierarchical clustering, were examined to analyze single-omic and multi-omic research data in the domain of CLL. They knew that machine learning could reveal new forms of heterogeneity, whether among diseases, patient subgroups, or in the context of precision medicine. However, the authors identified problems arising from noisy data, small sample sizes, and biological complexity oversimplified by some models [28]. Lan et al. (2024) developed a new interpretable model, DeepKEGG, for integrating multi-omics data to predict cancer recurrence and identify novel biomarkers. DeepKEGG adopts a hierarchical biological module approach to connect omics features (e.g., genes and miRNAs) to pathways, thereby enabling model interpretability. Furthermore, a pathway self-attention module is proposed within the framework to improve prediction performance by investigating sample correlations. In predicting cancer recurrence, DeepKEGG significantly outperformed all other state-of-the-art methods in the field while providing interpretable insights for biomarker discovery[29].

There has been significant progress in computational oncology, including the application of machine learning and deep learning models for predicting drug sensitivity in cancer treatment. Various researchers have pointed to techniques to improve personalized medicine using genomic data and advanced clinical analytics.

In predicting the sensitivity of different drugs to genomic features of cancer cell lines, Hasan and colleagues (2024) employed different machine learning models such as Ridge Regression, Linear Regression, and Stochastic Gradient Descent (SGD). They demonstrated that Ridge Regression was able to predict phenomena better than any other model of interest, which suggests that it has a high potential to inform clinical decision-making in a way based on tailored treatment selection. Nevertheless, the issue of data quality, incorporating genomic knowledge into clinical practice, and ethics continue to be the primary limitations [30].

Equally, Mehmood et al. (2023) identified 6 drugs and 5 biomarkers that were active against the cancer cell lines and sensitive to radiations in breast cancer. Their interpretation of the drug sensitivity based on data in Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) databases with Elastic Net, LASSO and Ridge regression models. The findings have the potential to be applied in clinical trials in the future, but issues related to dataset generalizability and validation are yet to be [31].

Deep learning was used to develop the Network Fusion method of predicting drug sensitivity (Liu and Mei 2023). The model has used multi-omics data (RNA sequencing, copy number alterations, and DNA methylation) in order to enhance the accuracy of its prediction. The NDSP had the highest sensitivity, specificity, and generalizability in comparison with a traditional machine learning classifier re-investigated on the same data set. However, interpretability and the likelihood of overfitting with the high-dimensional data are among the key issues [32].

Huang et al. (2023) explored the application of organotypic in precision oncology therapeutics using organotypic models to test the efficacy of functional drugs. The testing of

the drug sensitivities on human tissue cultures, especially on the biological mechanism of action of drugs, is one of its main differentiators compared to computational approaches. Although it was encouraging in clinical uses, the study was found that there was not any uniform platform on which the real-world implementation was to be implemented and that much advancement in technology is justified [33].

Pant et al. (2024) built the Random Forests to implement the prediction of drug response based on the genomic information. Their study also shows evidence of outperformance by Random Forest Linear Regression and MLP(XG-Boost) and hence, proves the ability of RF to provide insight on the molecular mechanisms of drug resistance. It understood, however, that there were constraints on sample size, the ability to generalize models, and that other legitimate sources of data in the real world were needed to enhance accuracy and clinical applicability [34].

Altogether, these studies show the potential of machine learning and deep learning algorithms in the prediction of drug sensitivity. The existing statistical models provide us with a baseline level of accuracy and we find that a more sophisticated, advanced model like a Deep Neural Network and a random Forest provides us with a higher performance. Nonetheless, several important issues related to the interpretability of the models, data integration, and clinical validation remain to be resolved before these strategies may be implemented into the everyday practice in the area of precision oncology.

3. Methodology

Multi-omics data integration of hybrid machine learning (ML) and deep learning (DL) models is a new technique to integrate the capabilities of the two techniques to effectively fuse data, analyze, and predict in bioinformatics. The following (Figure 1) shows the Algorithm 1, a step-by-step suggested methodology of implementing this hybrid model to use in data fusion of multi-omics in bioinformatics applications, biomarker discovery and personalized medicine..

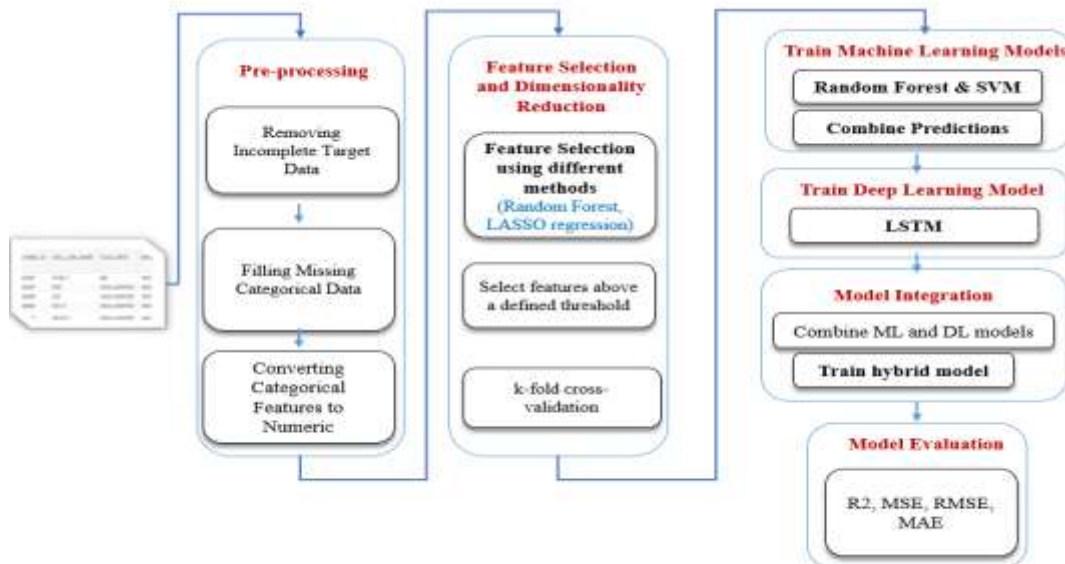


Figure 1: Proposed methodology

3.1 Multi-Omics Data Acquisition

The Genomics of Drug Sensitivity in Cancer (GDSC) dataset is a valuable resource for therapeutic biomarker discovery in cancer research. This dataset combines drug response data with genomic profiles of cancer cell lines, allowing researchers to investigate the relationship between genetic features and drug sensitivity (see Figure 2).

CDIMAC_1_CELL_LINE_TGGA_ID	DRUG_NAME_IC50	AUC	Z_SCORE	GDSC_TGGA_ID	GDSC_TGGA_Cancer_Ty	Microscale screen	My Growth	PCNA	GeneExp(Methylos)	TARGET	TARGET_PATHWAY				
680667 PF56-1	ME	3003	Campoth	-1.46388	0.93022	0.481223	nervous_system	medullab/MB	MSS/MSS-R	Adherent Y	DNA replication				
684097 C53	UNCCLASS	3003	Campoth	-1.36091	0.791072	-0.59957	bone	swings_sarcoma	MSS/MSS-R	Adherent Y	DNA replication				
684099 C57	UNCCLASS	3003	Campoth	-0.94494	0.59386	-1.51665	bone	swings_sarcoma	MSS/MSS-R	Adherent Y	DNA replication				
684092 SW-11	UNCCLASS	3003	Campoth	-0.30189	0.736047	-0.80729	bone	swings_sarcoma	MSS/MSS-R	Adherent Y	DNA replication				
684072 SW-45-1	UNCCLASS	3003	Campoth	-0.54296	0.582439	-1.57602	bone	swings_sarcoma	MSS/MSS-R	Semi-Adh Y	DNA replication				
687448 COLO-829	SKCM	3003	Campoth	-1.22501	0.867549	0.537327	skin	melanoma-SKCM	MSS/MSS-R	Adherent Y	DNA replication				
687502	3637	BLCA	3003	Campoth	-2.45281	0.84697	-0.20322				Y	Y	TOP2	DNA replication	
687655 HTA	BLCA	3003	Campoth	-2.96139	0.821436	-0.1887	urogenital	Bladder	BLCA	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687657 SW780	BLCA	3003	Campoth	-1.46914	0.90605	0.481354	urogenital	Bladder	BLCA	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687658 TCCSUP	BLCA	3003	Campoth	-2.35061	0.84343	-0.04988	urogenital	Bladder	BLCA	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687659 C-33-A	CESC	3003	Campoth	-3.39088	0.77898	-0.01862	urogenital	penis	CESC	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687656 C-41	CESC	3003	Campoth	-2.25649	0.851189	0.903012	urogenital	penis	CESC	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687514 AG-S-196	CESC	3003	Campoth	-1.22161	0.786658	-0.55515	urogenital	penis	CESC	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687563 AG-MG-BA	GBM	3003	Campoth	-2.40032	0.777317	-0.62515	nervous_system	glioma	GBM	MSS/MSS-R	Adherent Y	Y	Y	TOP2	DNA replication
687562 AG-MG-BA	GBM	3003	Campoth	-4.25029	0.839525	-1.08725	nervous_system	glioma	GBM	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687561 A272	GBM	3003	Campoth	-2.99954	0.854057	-0.405	nervous_system	glioma	GBM	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687568 GB-1	GBM	3003	Campoth	-2.0555	0.77759	-0.81846	nervous_system	glioma	GBM	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687566 T98G	GBM	3003	Campoth	-2.86022	0.880774	0.11844	nervous_system	glioma	GBM	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687568 U-118-MG	GBM	3003	Campoth	-1.35462	0.748288	-0.59939	nervous_system	glioma	GBM	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687560 U-87-MG	GBM	3003	Campoth	-2.43067	0.841792	-0.10426	nervous_system	glioma	GBM	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687562 YKO-1	GBM	3003	Campoth	-2.78446	0.824805	-0.26583	nervous_system	glioma	GBM	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687566 Chago-K	UNCCLASS	3003	Campoth	0.528845	0.3625	1.51803	lung	NSCLC	lung_NSC	MSS/MSS-R	Adherent Y	Y	Y	TOP2	DNA replication
687606 NCI-H735	UNCCLASS	3003	Campoth	0.532792	0.944892	1.326252	lung	NSCLC	lung_NSC	MSS/MSS-D/P12	Suspensio Y	Y	Y	TOP2	DNA replication
687777 Calu-3	LUAD	3003	Campoth	-1.82953	0.895243	0.342926	lung	NSCLC	lung_NSC	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687786 COH-125	UNCCLASS	3003	Campoth	-1.76268	0.86334	0.200767	lung	NSCLC	lung_NSC	MSS/MSS-R	Adherent Y	Y	Y	TOP2	DNA replication
687787 LC-2	LUAD	3003	Campoth	-1.57761	0.822968	0.596884	lung	NSCLC	lung_NSC	MSS/MSS-R	Adherent Y	Y	Y	TOP2	DNA replication
687794 NCI-H143	LUAD	3003	Campoth	-1.83052	0.90033	0.28801	lung	NSCLC	lung_NSC	MSS/MSS-R	Adherent Y	N	Y	TOP2	DNA replication
687796 NCI-H162	LUAD	3003	Campoth	-4.13994	0.887681	-1.03275	lung	NSCLC	lung_NSC	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687799 NCI-H164	LUAD	3003	Campoth	-1.96237	0.879483	0.161779	lung	NSCLC	lung_NSC	MSS/MSS-D/P12	Adherent Y	Y	Y	TOP2	DNA replication
687808 NCI-H165	LUAD	3003	Campoth	0.132034	0.836097	1.302591	lung	NSCLC	lung_NSC	MSS/MSS-R	Adherent Y	Y	Y	TOP2	DNA replication
687802 NCI-H183	LUAD	3003	Campoth	0.121063	0.816035	1.29857	lung	NSCLC	lung_NSC	MSS/MSS-R	Adherent Y	Y	Y	TOP2	DNA replication

Figure 2: GDSC dataset

The overriding objective in this dataset is to forecast drug sensitivity (using the form of IC50 values) in response to genomic characteristics of cancer cell lines. This may entail regression exercises to give the precise IC50 values or classification exercises to label cell lines as sensitive or resistant to certain drugs. Genomic markers related to drug response can also be identified using the dataset as shown in Figure 3 (a, b).

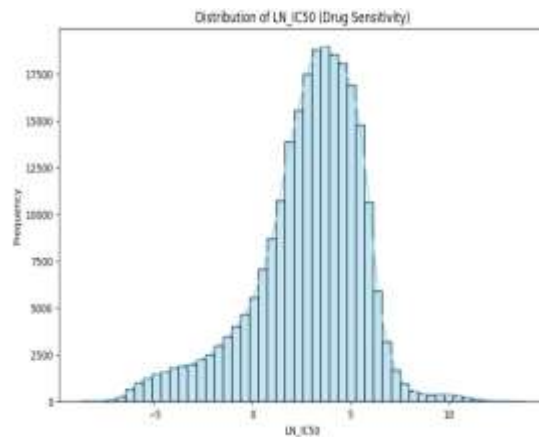


Figure 3 :a: Distribution of IC50

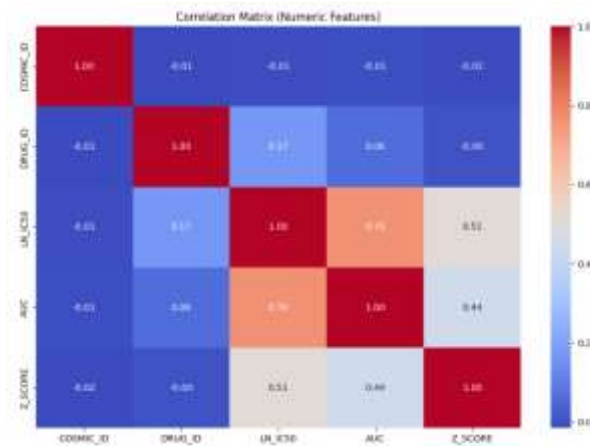


Figure 3 :b: Numeric Features

3.2. Data Preprocessing

The Genomics of Drug Sensitivity in Cancer (GDSC) dataset was preprocessed before model development to improve data quality and ensure compatibility with the proposed hybrid machine learning and deep learning framework. The target variable used in this study was LN_IC50, which represents drug sensitivity. As an initial data-cleaning step, all records with missing target values were removed because such observations could not contribute reliably to supervised learning. After target cleaning, missing values in categorical variables were handled using mode imputation. In the present dataset, this procedure was applied to two categorical fields, namely cancer conf_ID and microsatellite instability status, because these variables were categorical in nature and the manuscript indicates that the missing values were replaced by the most frequent category. Categorical predictors such as drug name, cancer type, and target pathway were then converted into numerical representations using one-hot encoding. In addition, identifier variables, including COSMIC_ID and CELL_LINE_NAME, were excluded from model training because they served only as labels and did not contain predictive biological information. Numerical variables were normalized using Min-Max scaling to place the features on a comparable range prior to modeling. Duplicate entries were removed, and highly correlated variables were filtered to reduce redundancy in the input space. (see Figure 4).

1. Removing Incomplete Target Data

The first step in your data cleaning process is to remove any rows where the TARGET column is missing. This is essential because the TARGET variable is usually the outcome you're trying to predict in a machine learning model. If it's missing, that row can't contribute useful information to your training or evaluation process.

2. Filling Missing Categorical Data

Next, handle NULL values in 2 propositional columns: one is a yes/no for cancer conf_ID, and the other represents microsatellite instability status. You fill them in with the most frequent value of each column. Imputation by mode. This approach is a practical, no-nonsense method for handling missing values in categorical features when the missing rate is low.

3. Converting Categorical Features to Numeric

Once the missing values are imputed (input variable imputation), we consider converting some categorical variables, such as drug name, cancer type, and target pathway, to numerical representations. This is achieved using one-hot encoding. Basically, it creates a new binary

column for each category, and the machine learning algorithm can work directly on those categorical values.

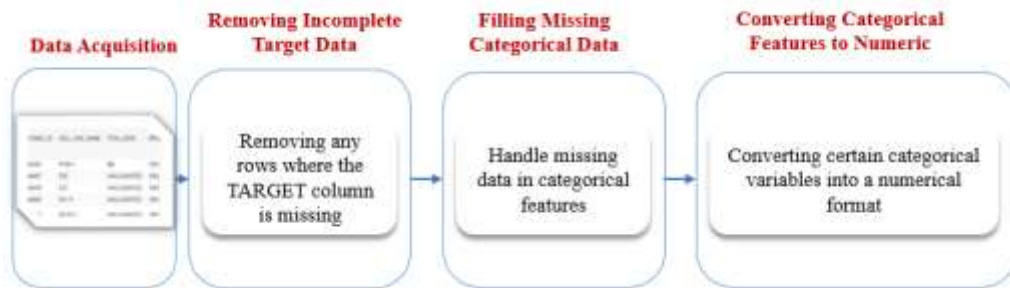


Figure 4: Preprocessing steps

4. Preparing Features and Target

It starts by separating features (inputs) from your target (output). The LN_IC50 (drug sensitivity) is the aim. You also drop identifying data such as COSMIC_ID and CELL_LINE_NAME because they are not predictive; they are just labels.

5. Encoding Categorical Variables

Even though you've previously encoded it, you go to great lengths to ensure there are no leftover categorical variables by converting them all into numeric variables with the power of one-hot encoding. This is important for Random Forest because it does not naturally process non-numeric data.

3.3. Feature Selection and Dimensionality Reduction

Multi-stage feature-selection strategy was used to reduce the dimensions and include the most informative predictors. To estimate the effect of each variable, first, the contribution of each variable to the prediction was approximated by using Random Forest feature importance, which quantifies the effect of the variable across the ensemble. The features that had low contribution were deemed to be less informative and were dropped out of the further modeling. Second, LASSO regression was used as an in-built feature-selection algorithm. Since LASSO includes L1 regularization, it brings the coefficients of weak predictors to zero and only the variables whose predictive contribution is non-zero are retained. Third, each feature was computed against the target variable using the Mutual Information in order to obtain both linear and non-linear dependency models. Features that had greater scores on Mutual Information were deemed to be more informative and were kept in order to be trained on the model. The last set of features was selected through a combination of the results of the importances of the Random Forest and Lasso regression along with the results of the analysis of Mutual Information. Cross-validation was applied with the use of k-folding to evaluate the robustness of each feature in order to minimize the risk of overfitting, which was recorded in Algorithm 1.

3.3.1 Random Forest for Feature Selection

Random Forest is an ensemble method of learning that may be applied both in classification and regression. It operates by creating various decision trees and integrating their outcomes. Random Forest in terms of feature selection offers a feature importance score which is used to quantify the utility of each feature to predict, as demonstrated in Figure 5.

Its functionality in feature selection: Random Forest calculates the contribution made by each feature to the reduction of the impurity of the decision tree (in terms of the Gini Impurity or Entropy). The greater the contribution of a feature the greater its significance is to the model. Most important features (those that decrease impurity) are picked, and the less relevant ones are eliminated.

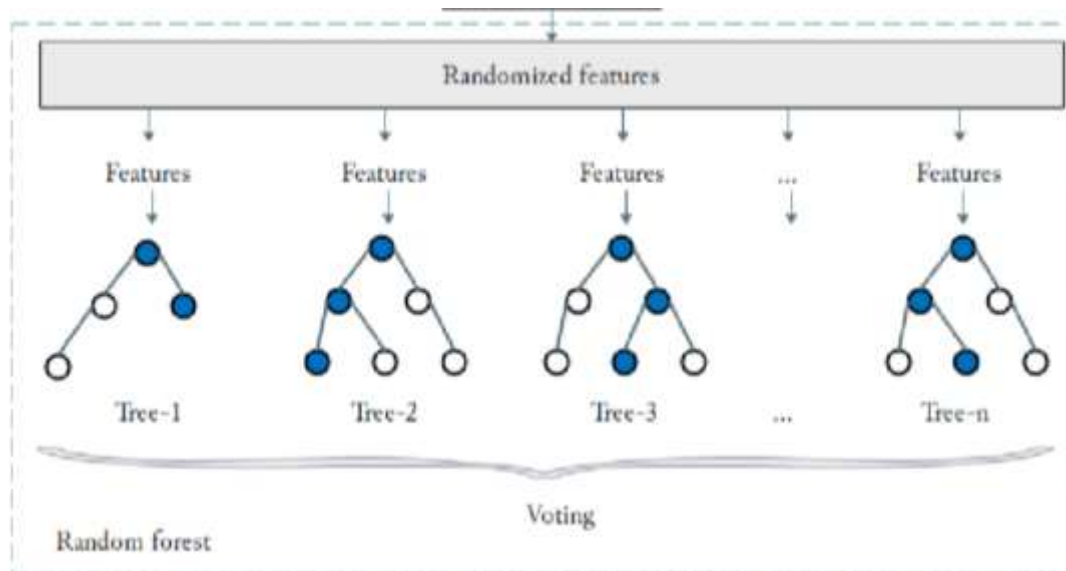


Figure 5: Random Forest for Feature Selection

3.3.2. LASSO Regression (Least Absolute Shrinkage and Selection Operator)

LASSO Regression is a linear model that performs both regularization and feature selection simultaneously. LASSO imposes a penalty on the coefficients of less important features, effectively shrinking them toward zero. If a feature has a zero coefficient, it is irrelevant, and therefore the model can be considered to have excluded it. LASSO inserts a term in the least squares optimization that is proportional to the absolute size of coefficients (ie, L1 penalty). This will result in fewer significant features having zero coefficients. Benefits: The end result is a model with fewer predictors, containing only the most important ones.

- It is known that LASSO has a built-in feature selection mechanism, setting some coefficients to zero. Features with non-zero coefficients are picked after training.
- Cross-validation is an essential concept for verifying that your selected features from the dataset are not only useful but also generalise well to unseen data, thereby avoiding overfitting. When selecting features, it is very important to validate them using cross-validation, or else you're at risk of overfitting – where your model does well on the training data but starts making bad predictions on completely new data.

3.4. Feature Selection Using Mutual Information

MI provides a statistical framework for computing the amount of information that is jointly shared by two variables. In the context of feature selection, MI quantifies how much information a given feature carries about the target. The aim is to filter out uninformative

features that are directly correlated with the target variable (in this case, disease status) and retain informative features.

Mutual Information is very effective, especially when relationships between features and the target are non-linear or complex. It aids in capturing linear and non-linear relationships and is therefore more flexible than traditional linear correlation tests. Procedure of Feature Selection Based on Mutual Information

1. Compute Mutual Information Between Features and Target Variable

One should compute the Mutual Information of each feature and the target variable first. The concept of Mutual Information could be calculated in the following general formula:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

Where:

I=communicated information between the target Y and the feature X.

Where x is a feature and y is a target, p (x,y) is the probability distribution of both.

p(x) and p(y) The marginal probability distribution of the feature and target.

In machine learning libraries like scikit-learn, there are built-in functions to compute mutual information.

2. Select Features with High Mutual Information

Once the mutual information scores for all features are computed, the next step is to identify the most informative features by selecting those with the highest mutual information with the target variable. Features that have a high mutual information score with the target are deemed more relevant because they provide significant information about the target variable. These features are retained for model training.

3. Rank and Threshold Selection

After calculating the mutual information scores, you can rank the features based on their scores. You can set a threshold value for the mutual information score to decide which features to keep.

- Top N Features: Select the top N features with the highest MI scores.
- Threshold-based Selection: Alternatively, you can set a minimum threshold for MI scores and keep only the features with a score greater than this threshold.

3.5 Model Design and Architecture

The proposed predictive framework combined classical machine learning models with a deep learning model in a hybrid architecture for multi-omics drug-sensitivity prediction. For the machine learning component, Random Forest and Support Vector Machine models were trained on the selected feature set. Random Forest was chosen because it is well suited to structured, high-dimensional biological data and can provide robust prediction while also supporting feature importance estimation. Support Vector Machine was included to complement the ensemble model with a margin-based learner capable of handling complex non-linear relationships in the selected feature space. For the deep learning component, a Long Short-Term Memory network was used. The manuscript identifies LSTM as the deep learning architecture because of its ability to learn complex patterns in high-dimensional data and to complement the machine learning branch in the hybrid framework.

Model integration was implemented through a hybridization strategy in which the predictions or learned representations from the machine learning models and the LSTM model were combined using a meta-model. The manuscript describes stacking as the principal integration strategy, while also acknowledging weighted averaging or voting as possible alternatives. In the present framework, the hybrid design was intended to exploit the robustness and interpretability of classical machine learning together with the representation-learning capacity of deep learning, thereby improving the predictive performance of the system on GDSC multi-omics data.

3.6. Feature Fusion

Feature fusion was performed by combining information learned from the classical machine learning models with information learned from the deep learning model. As described in the manuscript, this stage involved integrating raw or transformed features from the machine learning branch with features learned by the LSTM branch. Fusion could therefore be achieved through feature concatenation or through the transfer of deep representations into a machine learning stage. In the final hybrid framework, the combined feature representation was used to train the integrated predictive model. The purpose of this fusion step was to exploit complementary information extracted by the different modeling paradigms and thereby strengthen prediction accuracy in the presence of high-dimensional, heterogeneous, and interdependent omics data.

3.7. Training the Hybrid Model

Model development and evaluation were conducted using k-fold cross-validation. In each validation round, the model was trained on the training partition and evaluated on the held-out fold in order to reduce the risk of overfitting and obtain a more reliable estimate of generalization performance. The manuscript further states that the validation set remained separate from the training data during model development. For the deep learning component, training was carried out using a learning rate of 0.0001 and a batch size of 32. These settings were applied during training of the LSTM branch within the hybrid framework. The same validation procedure was used to assess the performance of the individual machine learning models, the deep learning model, and the final hybrid model.

3.8. Performance Metrics

The coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) were used in evaluating the predictive performance of the proposed hybrid framework. These measures were employed to evaluate complementary predictive performance measures. R^2 was a measure of the percentage of variance in LNIC50 accounted by the model, MAE, RMSE and MSE were measures of prediction error. Based on the reported findings, the last hybrid model was found to have an R^2 equal to 0.991, an MAE equal to 0.0726, an RMSE equal to 0.1113 and an MSE equal to 0.0109, meaning the hybrid model exhibits high predictive value and minimized drug-sensitivity estimation error.

It is the best complete replacement I can give out of the document alone without creating the details of the method. The unresolved issues with the critical importance include the precise post-preprocessing sample size, the retained features count, the correlation threshold, the complete random forest and SVM parameters, the precise LSTM architecture, the amount of

epochs and the k value used in cross-validation. None of these values is seen in the uploaded manuscript.

Algorithm 1: Proposed System of the GDSC multi-omics dataset

Step 1: LOAD dataset (GDSC multi-omics dataset), extract labels for drug sensitivity prediction.

Step 2: Preprocessing

- Data Cleaning and Quality Control
- Remove duplicate entries
- Normalize numerical features (Min-Max scaling)
- Feature Engineering
- Remove highly correlated features based on a threshold

Step 3: Feature Selection and Dimensionality Reduction

- Feature Selection using different methods (Random Forest, LASSO regression)
- Compute Mutual Information between features and target
- Select features above a defined threshold
- Cross-Validation for feature robustness (k-fold cross-validation)
- Store best-selected features

Step 4: Model Design and Architecture

- Train Machine Learning Models (Random Forest & SVM models on selected features)
- Combine predictions from multiple models using a meta-model

Step 5: Train Deep Learning Models

- Train a Long Short-Term Memory (LSTM) network

Step 6: Model Integration (Hybridization)

- Combine ML and DL models
- Concatenate learned features from ML and DL models
- Train the final hybrid model using the combined feature set

Step 7: Model Evaluation

Calculate performance metrics:

- R2, MSE, RMSE, MAE

4. Results and Discussion

The hybrid machine learning and deep learning model with the suggested framework achieved the high predictive accuracy in drug-sensitivity prediction on the GDSC dataset. The model had R² of 0.991, MAE of 0.0726, RMSE of 0.1113, and MSE of 0.0109 and summarised in Table 1 in Fig 6. These values showed that the model described a great fraction of the variance in LN₅₀ with reduction of small absolute and squared prediction error. In an experimental background used in the current research, the findings suggest that combining classical machine learning frameworks with an LSTM-based deep learning framework worked well in order to have a complex and heterogeneous structure of multi-omics data. The results as in Table 1 and Figure 6 also need to be viewed against the methodological pipeline in Figure 1 above in which, preprocessing, feature reduction, hybrid modeling and evaluation were all linked together in one workflow. Table 1: Results of the Proposed Method.

Matric	Value
R2	0.991
MAE	0.0726
RMSE	0.1113
MSE	0.0109

Table 1 shows in more detail the behavior of the model by the error values it gives. The MAE of 0.0726 shows that the average of the predicted and measured values of LN50 was not very large, but the RMSE of 0.1113 indicates that the magnitude of the residues was constrained. The value of MSE 0.0109, as presented in Table 1 and demonstrated in Figure 6, also fits the regularity of the model predictions to the analyzed samples. All these indicators together indicate that the hybrid structure not only explained all the data quite well, but it also preserved a small profile of errors in prediction. This is expected in accordance with the feature-selection as were the steps stated in the methodology, in which the importances of the Random Forest, LASSO regression and mutual information were utilized to exclude non-informative variables and minimize overlaps and redundancy during model training.

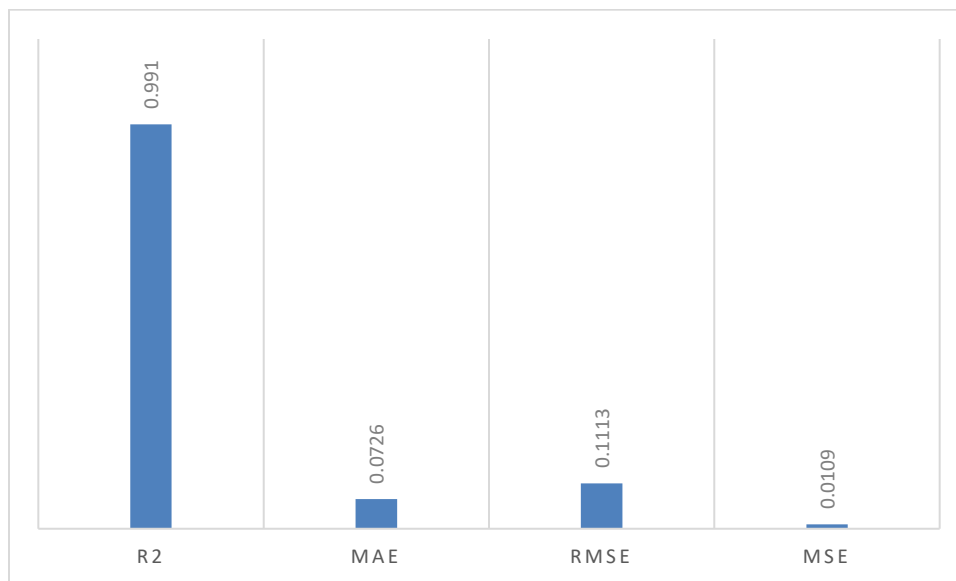


Figure 6: Proposed Method Results

Simultaneously, these results should be interpreted carefully. Even though the likelihood of optimistic evaluation decreases due to the use of the cross-validation strategy in the methodology, the abnormally large R² deserves special attention to the risk of overfitting. Figure 1 below depicts the workflow and Figure 5 above describes the feature-selection process that was used to minimize the risk of duplicate records, normalization of numerical

variables, filtering of highly correlated features, and validating the feature robustness of the final model using cross-validation. However, there was no validation on another dataset or on independent dataset in the present study. It is due to that reason that the results presented in Table 1 and Figure 6 indicate excellent internal predictive power, yet, they do not provide transportability besides the independent cohorts or different experimental conditions as yet. This limitation must be mentioned point-blank, as the present manuscript already acknowledges the fact that extrapolation on unobserved external data has not been proven.

Table 2: Our work comparison with related studies.

method	R2	MAE	RMSE	MSE
Ridge Regression [25]	0.986	0.07588	0.11716	0.0137
linear Regression [25]	0.986177	0.0764	0.11756	0.13823
linear Regression [27]	0.69	1.09	1.58	2.5
Random Forest Regression [27]	0.99	0.16	0.29	0.08
XGBoost Regression [27]	0.98	0.26	0.36	0.13
Our Work	0.991	0.0726	0.1113	0.0109

The useful nature of the offered method can be more clearly understood in the results of the comparison in Table 2 and Figure 7. The suggested framework yielded a better R² (0.991) and the reduction of MAE (0.07588), RMSE (0.11716), and MSE (0.0137) compared with Ridge Regression [25] (0.986). The absolute increase in R² in the model compared to Random Forest Regression [27] was less, which was 0.99 to 0.991, but the decrease in prediction error was greater: MAE to 0.0726, RMSE to 0.1113, and MSE to 0.0109. The difference to that of XGBoost Regression [27] was even greater, with the proposed model scoring higher on R-squared, 0.98 and 0.991; there was a smaller MAE of 0.0726, RMSE of 0.1113, and MSE of 0.0109. These comparisons as shown in Table 2 and Figure 7 in point to the major benefit of the proposed hybrid architecture, the better explained variance, but more significantly the substantially reduced prediction error.

Behaviour in Table 2 and Figure 7 can be viewed and understood in reference to the methodology fabricated in Figure 1. Traditional regressors like the Ridge Regression and classical Linear Regression use a less generalized form of the representation of the relationship between predictors and response and the suggested framework brings the feature-space refinement and hybrid prediction together. The dimensionality-reduction stages outlined in the methodology and relating to Figure 5 help then train dimensionalities and then enables the model to infer organized relationships, as well as more complicated, nonlinear interactions. This combination seems to offer superior trade-off between variable screening and predictive modeling compared to the single-model versions of the same in

Table 2, in the context of multi-omics data, where the space of features is high-dimensional, and heterogeneous.

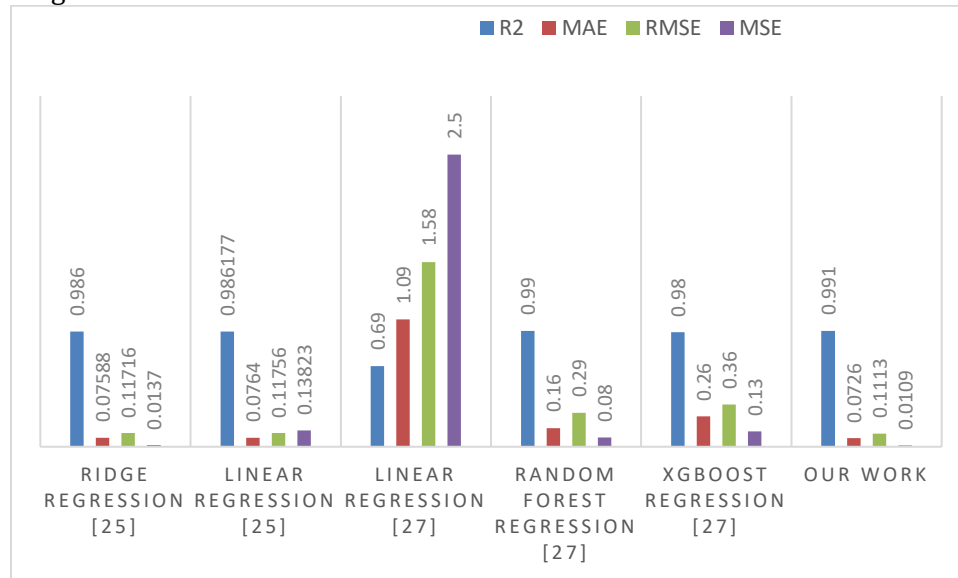


Figure 7: Comparison of our work with related studies

Despite this fact, the analogy should also not be exaggerated. The methods described in Table 2 were compiled based on related literature and were not necessarily compared on the same preprocessing pipeline, feature sets or evaluation. Table 2 and Figure 7 should therefore be viewed as reflective evidence between the context of the reviewed literature but not necessarily as vestigial evidence of universal excellence across all conditions of the analysis. It is safe to say that in the comparative context that is gathered in this manuscript, the minimum reported error values and maximum reported R 2 were found in this proposed method.

One more aspect that should not be excluded in the discussion is interpretability. Despite the fact that the hybrid architecture increased the accuracy of the predictions, it includes LSTM component, which causes the inference mechanism to be not as transparent as the simpler tree-based approaches. This matter holds significance since in bioinformatics and precision oncology, in many situations interpretability is important, especially since it may subsequently be accompanied by an association of model outputs with biomarker discovery or treatment prioritization. The paper thus has the advantage of reporting that the framework enhances predictive performance as observed in Table 1, Table 2, Figure 6 and Figure 7 although some model description and external validation is left to be done before generalized practical assertions could be done.

The results available in Table 1 and Figure 6, along with the evidence of comparison as presented above in Table 2 and Figure 7, show that the generated hybrid ML-DL framework is effective with the analyzed GDSC data in the context of internal drug-sensitivity prediction. Simultaneously, a complete workflow shown in Figure 1 demonstrates that it is not a single architectural decision that allows making the results quality but a combination of preprocessing, features selection, and model integration. It is a moderate interpretation that therefore the proposed framework enhances predictive accuracy in the existing data and

provides a methodologically consistent foundation upon which additional justification is made on independent groups.

Conclusion

Consequently, the fact that our hybrid ML-DL architecture has been trained to make multi-omic drug sensitivity predictions has shown that it can achieve better prediction than the traditional methods and thus is a suitable representative of such capabilities. A combination of powerful singular feature selection algorithms, Random Forest, LASSO, and Mutual Information, with the models being designed based on the Random Forest and the LSTM, is able to achieve state-of-the-art R^2 of 99.1 on data in the test set. The MAE, RMSE, and MSE are very low and it means that the predictive accuracy of our results is good and we could put our trust on our results. The hybrid framework is particularly significant as it deals with such critical issues in multi-omics as dimensionality of data, redundancy and complex associations between features and targets. In comparison with the other work, our model is not only better in predictive performance, but also better in generalizing due to paying attention to non-redundant informative features. Still, interpretation and external validation are somehow problematic. The model will be improved in the future through the use of a larger cohort and explainable AI methods, which are useful to apply AI in personalized medicine..

Funding: "This research received no external funding."

Conflicts of Interest: "The authors declare no conflict of interest."

References

- [1]M. A. Mohammed, A. Lakhan, K. H. Abdulkareem, and B. Garcia-Zapirain, "A hybrid cancer prediction based on multi-omics data and reinforcement learning, state action reward state action (SARSA)," *Comput. Biol. Med.*, vol. 154, p. 106617, Mar. 2023, doi: 10.1016/j.compbimed.2023.106617.
- [2]A. Mohammed Yousif, Noor M Jassam, Ahmad Salim, Hussein Ali Bardan, Ahmed Farhan Mutlak and A. F. A. D. Sallibi, "Melanoma Skin Cancer Detection Using Deep Learning Methods and Binary Melanoma Skin Cancer Detection Using Deep Learning Methods and Binary GWO Algorithm," no. January, 2025, doi: 10.54216/FPA.180211.
- [3]M. H. M. Khan et al., "Multi- class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN)," *PLoS One*, vol. 16, no. 8 August 2021, pp. 1–15, 2021, doi: 10.1371/journal.pone.0256500.
- [4]K. Chaudhary, O. B., Poirion, L., Lu, and L. X. Garmire, "Deep learning-based multi-omics integration robustly predicts survival in liver cancer," *Clin. Cancer Res.*, vol. 24, no. 6, pp. 1248–1259, 2019, doi: 10.1158/1078-0432.CCR-17-0853.
- [5]J. Wandy and R. Daly, "GraphOmics: an interactive platform to explore and integrate multi-omics data," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–19, 2021, doi: 10.1186/s12859-021-04500-1.
- [6]M. Tounsi, E. Aram, A. T. Azar, A. Al-Khayyat, and I. K. Ibraheem, "A Comprehensive Review on Biomedical Image Classification using Deep Learning Models," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 1, pp. 19538 – 19545, 2025, doi: 10.48084/etasr. 8728.
- [7]M. Yousif and B. Al-Khateeb, "Quantum Deep Learning: A Review," *Iraqi J. Sci.*, vol. 65, no. 8, pp. 4588–4605, 2024, doi: 10.24996/ijis.2024.65.8.37.

- [8]M. Yousif, I. A. Ahmad, A. R. Hmeed, and A. A. Mukhlif, "Enhanced EEG Signal Classification Using Machine Learning and Optimization Algorithm," *Fusion Pract. Appl.*, vol. 17, no. 2, pp. 211–218, 2025, doi: 10.54216/FPA.170216.
- [9]Ayyadevara and V. Kishore, "Gradient Boosting Machine," Springer, pp. 117–134, 2018, doi: 10.1007/978-1-4842-3564-5.
- [10]M. Y. Hasan, D. J. Kadhim, and A. J. Humaidi, "Prediction of electricity consumption and residential bills based on an artificial neural network," *Int. Rev. Appl. Sci. Eng.*, vol. 16, no. 1, pp. 142 – 152, 2025, doi: 10.1556/1848.2024.00856.
- [11]S. M. Al-Barzinji, "Diagnosis of lung cancer disease using machine learning techniques," *Iraqi J. Inf. Technol.* V, vol. 8, pp. 110–130, 2018.
- [12]A. A. Aldujaili, M. Rosa-Zurera, and M. Utrilla-Manso, "Review of Parkinson's disease detection with deep-learning machines," *Multimed. Tools Appl.*, vol. 84, no. 6, pp. 3015 – 3051, 2025, doi: 10.1007/s11042-024-20466-7.
- [13]K. K. Mujeeb Rahman, M. Nasor, and A. Imran, "Automatic Screening of Diabetic Retinopathy Using Fundus Images and Machine Learning Algorithms," *Diagnostics*, vol. 12, no. 9, 2022, doi: 10.3390/diagnostics12092262.
- [14]M. S. Al-Ani, Q. Al-Shayea, S. M. Al-Barzinji, D. M. S. Al-Ani, and Z. M. S. Al-Ani, "COVID-19 Diagnosis Applied DWT and CNN on X-ray Chest Images," *UHD J. Sci. Technol.*, vol. 7, no. 2, pp. 69–76, 2023, doi: 10.21928/uhdjst.v7n2y2023.pp69-76.
- [15]A. A. Abdalhameed and A. I. Kadhim, "Molecular Swarm Optimization Analysis of Data Transmission and Recurrent Neural Networks (RNNs) for Attack Prevention in Mobile Ad Hoc Networks (MANETs)," *KSII Trans. Internet Inf. Syst.*, vol. 19, no. 3, pp. 973 – 986, 2025, doi: 10.3837/tiis. 2025.03.014.
- [16]S. Pang and X. Yang, "Deep Convolutional Extreme Learning Machine and Its Application in Handwritten Digit Classification," *Comput. Intell. Neurosci.*, vol. 2016, 2016, doi: 10.1155/2016/3049632.
- [17]A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features," *IEEE Access*, vol. 6, pp. 1155–1166, 2017, doi: 10.1109/ACCESS.2017.2778011.
- [18]M. Yousif, B. Al-khateeb, and B. Garcia-zapirain, "A New Quantum Circuit of Quantum Convolutional Neural Network for X-Ray Images Classification," *IEEE Access*, vol. 12, no. March, 2024.
- [19]S. KILICARSLAN, M. CELIK, and Ş. SAHIN, "Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification," *Biomed. Signal Process. Control*, vol. 63, no. November 2019, 2021, doi: 10.1016/j.bspc.2020.102231.
- [20]U. Kumar Lilhore et al., "A precise model for skin cancer diagnosis using hybrid U-Net and improved MobileNet-V3 with hyperparameter optimization," *Sci. Rep.*, vol. 14, no. 1, p. 4299, 2024, doi: 10.1038/s41598-024-54212-8.
- [21]A. Mossleh, A. A. Mossleh, R. H. Mahdi, and S. M. Al-barzinji, "Performance Evaluation Of The Deep Learning System For Weed Performance Evaluation Of The Deep Learning System For Weed," 2023.
- [22]R. S. Patil and N. Biradar, "Automated mammogram breast cancer detection using the optimized combination of convolutional and recurrent neural network," *Evol. Intell.*, vol. 14, no. 4, pp. 1459–1474, 2021, doi: 10.1007/s12065-020-00403-x.
- [23]M. A. Mohammed, B. Al-khateeb, M. Yousif, S. A. Mostafa, S. Kadry, and K. H. Abdulkareem, "Novel Crow Swarm Optimization Algorithm and Selection Approach for Optimal Deep Learning COVID-19 Diagnostic Model," vol. 2022, 2022.

- [24]S. Yang, Z. Wang, C. Wang, C. Li, and B. Wang, "Comparative Evaluation of Machine Learning Models for Subtyping Triple-Negative Breast Cancer: A Deep Learning-Based Multi-Omics Data Integration Approach," *J Cancer*, vol. 15, pp. 3943–3957, 2024, doi: 10.7150/jca.93215.
- [25]M. Briscik, G. Tazza, M.-A. Dillies, L. Vidács, and S. Dejean, "Supervised Multiple Kernel Learning approaches for multi-omics data integration," *BioData Min.*, pp. 0–25, 2024, doi: 10.1186/s13040-024-00406-9.
- [26]Y. Zhou et al., "Multimodal functional deep learning for multiomics data," *Brief. Bioinform.*, vol. 25, no. 5, p. bbae448, 2024, doi: 10.1093/bib/bbae448.
- [27]L. M. Sanches, P. H. G., de Melo, N. C., Porcari, A. M., & de Carvalho, "Integrating Molecular Perspectives: Strategies for Comprehensive Multi-Omics Integrative Data Analysis and Machine Learning Applications in Transcriptomics, Proteomics, and Metabolomics," *Biology (Basel)*, vol. 13, no. 11, 2024.
- [28]M. Tsagiopoulou and I. G. Gut, "Machine learning and multi-omics data in chronic lymphocytic leukemia: the future of precision medicine?" *Front. Genet.*, vol. 14, no. January, pp. 1–7, 2023, doi: 10.3389/fgene.2023.1304661.
- [29]W. Lan, H. Liao, Q. Chen, L. Zhu, Y. Pan, and Y. P. P. Chen, "DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery," *Brief. Bioinform.*, vol. 25, no. 3, 2024, doi: 10.1093/bib/bbae185.
- [30]E. Hasan, "CANCER DRUG SENSITIVITY THROUGH GENOMIC DATA: INTEGRATING INSIGHTS FOR PERSONALIZED MEDICINE IN THE USA HEALTHCARE SYSTEM," vol. 06, pp. 36–53, 2024.
- [31]A. Mehmood, S. Nawab, Y. Jin, H. Hassan, A. C. Kaushik, and D.-Q. Wei, "Ranking Breast Cancer Drugs and Biomarkers Identification Using Machine Learning and Pharmacogenomics," *ACS Pharmacol. Transl. Sci.*, vol. 6, no. 3, pp. 399–409, Mar. 2023, doi: 10.1021/acsptsci.2c00212.
- [32]X. Y. Liu and X. Y. Mei, "Prediction of drug sensitivity based on multi-omics data using deep learning and similarity network fusion approaches," *Front. Bioeng. Biotechnol.*, vol. 11, no. April, pp. 1–12, 2023, doi: 10.3389/fbioe.2023.1156372.
- [33]Y. L. Huang et al., "Organotypic Models for Functional Drug Testing of Human Cancers," *BME Front.*, vol. 4, p. 22, 2023, doi: 10.34133/bmef.0022.
- [34]L. Pant, A. Al Mukaddim, K. Rahman, and A. A. L. Sayeed, "Genomic predictors of drug sensitivity in cancer : Integrating genomic data for personalized medicine in the USA," vol. 5, no. 12, pp. 2682–2702, 2024, doi: 10.51594/csitrj.v5i12.1760.

1

Doi : <https://doi.org/10.54216/IJNS.xxxxxx>

Received: July 20, 2021 Accepted: Jan 04, 2022