Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

Predicting Drug Resistance in Gastric Cancer with Mutation in the Human Epidermal Growth Factor Receptor 2 (HER2) And Machine Learning Technique

Elham Soltanian[1],Amin Ramezani[2*]

[1]Ph.D. Student of Electrical Control Engineering, Electrical and Computer Faculty , Tarbiat Modares University (elhamsoltanian40@gmail.com)

[2]Associate Professor of Electrical and Computer Faculty of Tarbiat Modares University*
(ramezani@modares.ac.ir)

Abstract

Gastric cancer with mutations in the human epidermal growth factor receptor (HER2) can be regarded as one of the leading causes of cancer mortality in the world. Targeted tyrosine kinase inhibitors (TKIs) developed against HER2 yielded optimistic results in improving patients' survival rates and life quality. Nevertheless, drug resistance can influence the critical supportive documents of treatment plans and decrease the treatment effectiveness after about one year. Predicting the efficacy of HER2-TKI drugs or therapies for patients with HER2-mutated gastric cancer is a critical research field. In the present study, a personalized drug response prediction model based on molecular dynamics simulations and machine learning is presented to predict response to first-generation drugs approved by the Ministry of Health in patients with gastric cancer. In the molecular dynamics simulation, the patient's mutation status is considered. The patient's unique mutation status was modeled using molecular dynamics simulations to extract geometric features at the molecular level.Furthermore, additional clinical features are incorporated into the machine learning model to predict drug response. The complete features encompass demographic and clinical characteristics, geometrical properties of the drug-target binding site, and binding free energy (RBFE) of the drug-target complex from molecular dynamics simulations. Drug response prediction utilizes the XGBoost classifier, which achieved leading-edge performance for a 4-level drug response prediction task (PDRP) with 97.5% accuracy, 93% sensitivity, 96.5% specificity, and 94% F1 score.

The present research has demonstrated that modeling the binding cavity geometry, in tandem with the binding free energy, can effectively predict drug response. Interestingly, the clinical information, while significant, did not significantly influence the model's performance. This exciting finding opens up new avenues for testing the proposed model on various types of cancers, potentially revolutionizing drug development strategies.

Keywords: prediction of drug resistance, gastric cancer, mutation in the human epidermal growth factor receptor 2 (HER2), and machine learning technique

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

1. Introduction

Gastric cancer is considered as one of the principal causes of death all over the world [1], and has the lowest survival rate among the other cancers. It is the second most prevalent cancer and is often diagnosed when metastatic have spread to other parts of the body [2, 3]. In the last decade, rapid advancements have been made in the management of patients with gastric cancer [5]. *Molecular targeted* therapies have made many achievements, and the human epidermal growth factor receptor 2 (HER2) and ErbB family members are recognized as valuable therapeutic targets [4]. The US Food and Drug Administration has approved three generations of tyrosine kinase inhibitors (TKIs) as first-line treatment for gastric cancer patients with HER2 mutations [6]. These inhibitors have obtained hopeful results in the early stages of treatment and have increased patients' survival rates and quality of life [7].

Nevertheless, drug resistance is observed in many cases [8]. One of the principal reasons for drug resistance is secondary point mutation in the HER2 kinase domain [9]. Some research tried to decipher the drug resistance mechanism in gastric cancer with HER2 mutation [10, 11]. These studies revealed many causes, among which the secondary point mutation T790M [12], the in situ hydrogen bond breakage, and the AKT reactivation [11] can be regarded. Computational methods have been widely resorted to examine these drug resistance mechanisms [3, 13].

Molecular dynamics simulation is a computational material used to conceive the dynamics, stability and structural changes. A framework has been recently developed to visualize protein-drug interactions in the analysis of drug resistance in gastric cancer. Drug responses are still so diversified and unexplained in many patients. In this regard, the clinical-genetic features of patients might have a critical role in the mechanism of drug resistance and classification of patients.

The conclusion of the human genome project enabled a shift in treatment approach from the broad traditional medical model covering all individuals to targeted, customized treatments [21, 22]. Data from genomics and electronic health records (EHR) opens up new possibilities in patient care, prevention, and the development of effective treatment plans. Computational methods face challenges in predicting a patient's reaction to drug therapy and identifying the best treatment approach, including drug combinations and dosages, because of limited data, conflicting labels, and unknown biological evidence [24, 25]. Drug binding sites, binding free energy, geometrical features, and clinical information can be used to model multiclass drug responses. Several high-resolution structures of HER2 bound to various approved drugs are available in the Protein Data Bank (PDB), offering a chance to develop models using structured data. Since protein-drug interactions frequently determine how drugs affect individuals' drug response, the geometry of the drug-target binding site or binding cavity can be a valuable predictor of drug response. Using molecular dynamics simulations to analyze mutation-induced drug complexes and patient

Power System Technology
ISSN:1000-3673
Received: 16-01-2024 Revised: 12-02-2024 Accepted: 07-03-2024

characteristics, drug response levels can be accurately classified into two categories with a 95.3% success rate when fed into a machine learning algorithm [13].

By incorporating local geometric and energy-related features into a binding site approach, researchers could predict four classes of drug responses with an average accuracy of 69.35% [26]. Another study utilized demographic and lifestyle data from patients and geometric features of drugs to create a prediction model [30]. In contrast, a three-level drug response prediction model (PDRP) utilized tensors of drug-protein interaction footprints [31].

Despite demonstrating the possibility of integrating molecular dynamics features with patient information to forecast drug reactions, there still needs to be a need to enhance the accuracy of these predictions for practical clinical usage. This particular research project merges the geometrical properties of the drug-target binding site, binding free energy, and diverse clinical data of patients to create a personalized model for predicting four different types of drug responses. The aim is to develop a tailored drug response prediction system that excels in predicting drug responses with the utmost precision.

2- Method

The current research created a drug response prediction model for patients with gastric cancer. Information on demographic and clinical characteristics, such as age, gender, survival time, smoking history, tumor progression level, and mutation type, will be gathered from past studies. The HER2-Gefitinib complexes will be simulated in AMBER software for 2 ns, and the resulting simulated trajectories will be extracted. According to the simulated trajectories, we plan to introduce novel geometric features (matching rate, number of attached atoms, and number of hydrogen bonds) and (number of convex atoms and Euclidean distance) of HER2-Gefitinib mutant complexes to forecast the drug response level, including complete response, partial response, stable disease, or progressive disease. The binding free energy will be incorporated as a feature in the machine learning algorithm. Integrating features about clinical, geometric, and energy information will enhance the effectiveness of the machine learning algorithm.

2-1. Basic Information

The research involved 100 patients diagnosed with gastric cancer who were all given HER2-TKIs as the initial treatment. A response level of 0 or 1 signifies a complete or partial response to the medication. Response levels 2 and 3 indicate stable or progressive disease (no-response).

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

2-2. Summary of the proposed feature set

The demographic and clinical information and the features of energy and the HER2-Gefitinib mutation-induced complex were extracted for each patient to predict the drug response level using machine learning classifiers. The characteristics of these features are thoroughly explained in Table (1). Specifically, we extracted 4 features from patient demographics, 4 features from energy, and 5 from geometry.

Table 1. Clinical information, energy features, and geometric features: Description and values

Feature type	Classification of features	Description
Demographic information	Age	Patient's personal information
	gender	
	Smoking history	
	Response status	
Energy	VDW	Van der Waals energy
	EEL	Electrostatic interactions
	ESURF	No polar component of solvation energy
	EPB	Polar component of solvation energy
geometric	Matching rates	Identical atoms
	Convex atoms	The power of interaction
	Connectivity	bonded atoms
	Euclidean distance	The distance between the drug and the target
	Number of Hydrogen bonds	Number of hydrogen bonds

2-3. Data collection

The clinical information utilized in this study will be gathered from multiple sources [32-34]. A dataset comprising 100 individuals diagnosed with gastric cancer will be acquired.

2-4. Demographics and clinical information (DCI)

The clinical dataset will contain demographic information such as age, gender, and smoking history, as well as clinical information such as patient survival (0 or 1), drug response level, and functional status. Considering the presence of dominant categories (common values), age will be classified into 0 to 4 categories based on specific age ranges: 0 to 40 years, 41 to 50 years, 51 to 60 years, 61 to 70 years, and over 71 years.

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

2-5. 3D structural modeling

The crystal structure of HER2 (wildtype) obtained from the protein database with PDB 2ITY ID was used for the 3D mutation-induced structures. In Rosetta software, the high-resolution ddgmonomer protocol (HRDM)43 was employed to predict point mutations, while the comparative modeling protocol was used for predicting multi-point mutations. Verify3D45 and Q-mean46 were utilized to assess the predicted structures' quality.

3- Molecular dynamics simulation

For protein-drug complex molecular dynamics simulation, the QM/MM method was used in Amber47 software. The protein-drug complex was placed in a blue box with neutralizing sodium (Na^+) and chlorine (Cl^-) ions. ff9SB and GAFF force fields were used for simulation. The total energy of the system was the sum of bonding (stretching, bending, rotation) and non-bonding (electrostatic, van der Waals) energies.

(1)

$$E_{total} = E_{stretch} + E_{bend} + E_{torsion} + E_{electrosttic} + E_{vdw}$$

Energy minimization will be conducted before initiating the molecular dynamics simulation to enhance the model structure. The simulation will begin by applying heat to elevate the system's temperature from 0 degrees Kelvin to 300 degrees Kelvin. Subsequently, the system will be maintained at a constant pressure and density for 500 picoseconds to achieve equilibrium. In the next step, for another 50 picoseconds, equilibrium will be established with constant pressure but free compression. The SHAKE49 algorithm is used to control the temperature and limit the bond stretching. After reaching the stable state, molecular dynamics simulation will be executed for 2 nanoseconds at constant temperature (300 degrees Kelvin) and 1 atmosphere pressure. A 12-core central processor with a frequency of 3.47 GHz and a RAM of 8 GB will be used to execute the simulations.

3-1. Binding free energy

Estimating the binding tendency of a drug to a protein in a mdium containing solvent is done using the binding free energy. The molecular dynamics pathway is considered as input for MM-GBSA. The binding free energy is calculated based on the thermodynamic cycle theory in vacuum and the mdium containing solvent [35].

$$\Delta G = \Delta G_{Bind,Vacum} + \Delta G_{Solv,Complex} - (\Delta G_{Solv,ligand} + \Delta G_{Solv,\mathrm{Re}ceptor}) \quad (2)$$

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

The abovementioned equation describes the difference in binding free energy of the receptor-ligand system in a vacuum. The terms $\Delta G_{(Solv,Receptor)}$, $\Delta G_{(Solv,ligand)}$, and $\Delta G_{(Solv,Complex)}$ indicate the energy difference between the empty vacuum state and the solvent state for the protein, ligand, and the entire receptor-ligand system. Furthermore, the total energy ΔG (total energy) encompasses:

- Van der Waals forces (VDW)
- Electrostatic energy (EEL)
- Electrostatic contribution to solubility (ESURF)
- Nonpolar contribution to the free energy of solubility

3-2. Geometric features

In methods of prediction, interactions between the binding site residues[1] of a protein and molecule inhibitors are commonly utilized. The alpha shape will determine the geometric properties of the surface using the Computational Geometry Algorithms Library (CGAL)[2].

3-3. Convex atoms

In a mutation-induced drug system, each atom is characterized by a position and weight, represented as a = (p, w), with p being the position and w being the weight. Two atoms atoms $a_1=(p_1, w_1)$ and $a_2=(p_2, w_2)$ are identified as orthogonal or quasi-orthogonal based on the following equation .

$$\begin{cases} \left|p_1 p_2\right| = w_1 + w_2, & \alpha_1 \perp a_2 \\ \left|p_1 p_2\right| = w_1 \succ w_2, & \alpha_1 \perp_s a_2 \end{cases} \quad (3)$$

The solid angle of atoms was identified from the alpha shape to describe the geometric features of the nearby surface. The solid angle, Ω i, of a tetrahedron with vertices A, B, C, and D is determined as equal to:

[1] Binding site residues: amino acids that are in a protein and bind to a ligand (such as an inhibitor).

[2] Computational Geometric Algorithm Library (CGAL): a software library that contains algorithms for geometric calculations.

Power System Technology
ISSN:1000-3673
Received: 16-01-2024 Revised: 12-02-2024 Accepted: 07-03-2024

$$\Omega_i = \sum_i (\phi_i^{AB} + \phi_i^{BC} + \phi_i^{AC} - \pi) \quad (4)$$

where ϕ_i^{AB}, ϕ_i^{BC}, and ϕ_i^{AC} represent the vertices of tetrahedron i.

$$\Omega^{'} = \frac{\cos(\Omega_i)}{4} \quad (5)$$

If the calculated solid angle's value is positive, the resulting shape will be convex; conversely, if it is negative, it will be concave.

3-4. Matching rates

The atoms in the structures' bonding region create the connection between the drug and the target. An alpha algorithm was utilized to gather surface atoms, which was designated as point set A. Next, point sets B and C will be acquired to correspondingly illustrate the target and drug surface atoms. The interacting atoms were identified through set operations and subsequently classified based on the following equation:

- interacting atoms in the drug (I_d)
- interacting atoms in the target (I_t)

$$\begin{cases} I = (B \bigcup C) - A \\ I_t = (I \bigcap B) \\ I_d = (I \bigcap C) \end{cases} \quad (6)$$

The matching rate is established based on the selection of atoms in the drug and the target. When one atom is convex and the other is concave, it is considered a matched pair, signifying a solid interaction. Conversely, if both atoms are either convex or concave, it is deemed a mismatched pair, resulting in a weak interaction. The determination of matched and unmatched atoms is as follows:

$$f(B,C) = \begin{cases} 1 & \Omega_B \times \Omega_C \prec 0 \\ 0 & \text{Otherwise} \end{cases} \quad (7)$$

The matching rate for every frame in the molecular trajectory is determined in the following manner:

Power System Technology
ISSN:1000-3673
Received: 16-01-2024 Revised: 12-02-2024 Accepted: 07-03-2024

$$MR = \frac{\sum_{i,j} f(B_i, C_i)}{N} \quad (8)$$

• MR indicates the rate at which matches are made (matching rate).
• **$f(B_i,C_j)$** refers to a pair of matched atoms- one from atom **B_i** in point set B (target-related) and the other from atom **C_j** in point set C (drug-related).
• N denotes the overall count of molecular dynamics images.

Matching rate was a critical feature in the current study, as lower matching rates correlated with poor drug responses. Essentially, when there were fewer pairs of atoms with complementary convex and concave states during molecular dynamics simulations, the likelihood of the drug affecting the target decreased.

3-5. Binding measurement

Binding changes occur between the binding site residues and the drug molecule during molecular dynamics simulation. For this particular research, a local threshold value of 40 has been established using the Euclidean distance, and the number of atoms remaining within this threshold throughout the simulation is documented. The stability of these bindings could indicate significant atoms and serve as predictors of the drug response level. In summary, this section focuses on analyzing the changes in binding between the atoms at the binding site of the target and the drug molecule in molecular dynamics simulations. By establishing a specific minimum distance threshold, the aim is to pinpoint atoms that maintain their bond with the drug throughout the simulation. The durability of these bonds may be beneficial in predicting the drug's efficiency in terms of response.

$$C_{k,i} = \sum_j A_{k.i.j} \quad (9)$$

Suppose $A_{(k,\,i,j)}$ denotes the connection between the HER2 atom at index i and the drug atom at index j in snapshot k of the molecular dynamics simulation, where the value is 1 for a binding and 0 for no binding. Let us consider the following:

$$D_k = \sum_i C_{k,i} \succ 0 \quad (10)$$

In the current study, the value of C_k in molecular dynamics simulation represents the number of atoms connected in snapshot k. This feature, the total number of connected atoms in the entire path, is utilized in the research.

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

3-6. Binding site positioning

This position is evaluated using the Euclidean distance between the atoms of the HER2 binding site and the center of the drug molecule.

$$D(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2} \quad (11)$$

- Binding site atoms: Alpha carbon atoms (CA) depict the binding site residues.
- Example: If 14 CA atoms are located in the binding site and 2 atoms are in the center of the drug molecule, this binding is represented as a vector with dimensions 14 x 2 or 28 x 1.
- Distances: During the entire molecular dynamics simulation, a 200 x 28 matrix can be created to show the distances between the CA atoms in the binding site and the drug atoms at each of the 200 frames (snapshots).
- Binding site positioning: The position of the binding site is shown as the mean distance from the drug to the target.

$$D_{avg} = \frac{\sum_{i=1}^{N}(D_i)}{N} \quad (12)$$

D_{avg} signifies the location of the binding site.
D_i measures the distance in a specific snapshot demonstrating molecular dynamics simulation.
N represents the total number of snapshots in molecular dynamics simulation.

The feature values, specifically distances, are adjusted to fit within the range of [0,1]. Typically, Drug-sensitive mutations have a decreased distance between the drug and the target. Normalization facilitates data comparison and analysis. This process ensures that all values fall within a uniform range, allowing for more accurate comparisons.

3-7. Compound geometric features

In this section, , the computed features are merged to enhance simplicity and effectiveness, resulting in the creation of two new compound features:
The first compound feature (x_{g1}): This feature comprises the matching rate, the number of attached atoms and the number of hydrogen bonds. Visualize these three features merged together to form a solitary feature.

Power System Technology
ISSN:1000-3673
Received: 16-01-2024 Revised: 12-02-2024 Accepted: 07-03-2024

The second compound feature (x_{g2}) consists of the count of convex atoms and Euclidean distances. This feature is derived from combining the earlier two features. Likewise, personal features and energy features are merged to form two new compound features. The goal of merging features is to extract better the information they contain and help machine learning models predict drug response levels more accurately.

3-8. Normalization of features

To enhance data comparison and analysis, all features are standardized using the z-score normalization technique within a range of [-1,1]. The z-score normalization method involves determining each feature's mean and standard deviation, then adjusting each feature value by subtracting the mean and dividing by the standard deviation. Consequently, all feature values are transformed to fit into the new range of [-1,1].

$$z_i = \frac{x_i - \mu}{\sigma} \quad (13)$$

in which, z_i signifies the standardized value, μ signifies the average, and σ signifies the standard deviation for every feature.

3-9. Development of a cassification model

In the present research, geometric and energy features were extracted from HER2 mutation-induced drug complexes of gastric cancer patients with clinical data, and machine learning models were developed to predict the level of drug response. Five commonly used classifiers such as KNN, SVM, Artificial Neural Network, Random Forest, and XGBoost will be evaluated using Python libraries Scikit-learn and TensorFlow. Furthermore, feature visualization diagrams will be generated using the CARET package in RStudio software. A total of 141 samples will be utilized for model training, and the parameters will be fine-tuned using a network search approach. A neural network with four layers is trained using a sigmoid activation function for the hidden layers and Softmax for the output layer for 8000 periods. Various activation functions, such as ReLU and Dropout with varying thresholds, were tested during model selection. The loss function selected is categorical cross-entropy, and the optimizer chosen is RMSProp. Early stopping with a tolerance value of 100 is also implemented to monitor validation losses.

4- Results

In this study, we developed a model to predict drug response in patients with gastric cancer. We gathered demographic and clinical data, including age, gender, survival time, smoking history, tumor progression level, and mutation type from prior research. We simulated HER2-Gefitinib complexes using AMBER software for 2 nanoseconds and analyzed their trajectories. New geometric features x_{g1} such as matching rate, number of attached atoms, and number of hydrogen

Power System Technology
ISSN:1000-3673
Received: 16-01-2024 Revised: 12-02-2024 Accepted: 07-03-2024

bonds, as well as binding cavity features x_{g2} like number of convex atoms and Euclidean distance, were proposed based on the simulated trajectories of mutation-induced complexes HER2-Gefitinib to forecast the extent of drug response, such as complete response, partial response, stable disease, or progressive disease. In the machine learning model, binding free energy was also included as a feature. The amalgamation of clinical, geometric, and energy-related features enhanced the efficiency of the machine-learning model.

4-1. Basic information

The current research involved the analysis of data from 100 patients diagnosed with gastric cancer. The average age of the patients was 63 years. Among the patients, 56 were female (37%) and 94 were male (62%). Approximately 70% of the patients had no history of smoking. All patients included in the study were initially treated with HER2-TKIs. The response to the drug was categorized into four levels: 0 and 1 indicated complete and partial response, while 2 and 3 indicated stable and progressive disease (no response). The dataset analyzed in this study comprised 10 patients with complete response, 50 with partial response, 20 with stable disease, and 20 with progressive disease (no response).

The demographic and clinical information (DCI) and the energy and geometrical features of the mutation-induced complex HER2-Gefitinib were collected for each patient. These data were utilized to predict the drug response level using machine learning algorithms. A comprehensive overview of the features and their respective value ranges can be found in Table 2. Overall, we identified 4 DCI features, 4 energy features, and 5 geometry features. A box plot illustrating various features and their correlations is presented in Figure (1).

Table 2 - Clinical information as well as energy and geometric features: Description and values

Feature type	Classification of features	Description	Range
demographic information	Age	Patient's personal information	[0–4]
	gender		[0–2]
	Smoking history		[0–2]
	Response status		[0–3]
energy	VDW	Van der Waals energy	[− 60 to − 45]
	EEL	Electrostatic interactions	[− 23–11]
	ESURF	No polar component of solvation energy	[− 45 to − 1]
	EPB	Polar component of solvation energy	[27–40]
geometric	Matching rates	Identical atoms	[0, 17]
	Convex atoms	The power of interaction	[0, 43]
	Connectivity	bonded atoms	[0, 23]

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

	Euclidean distance	The distance between the drug and the target	[30–39]

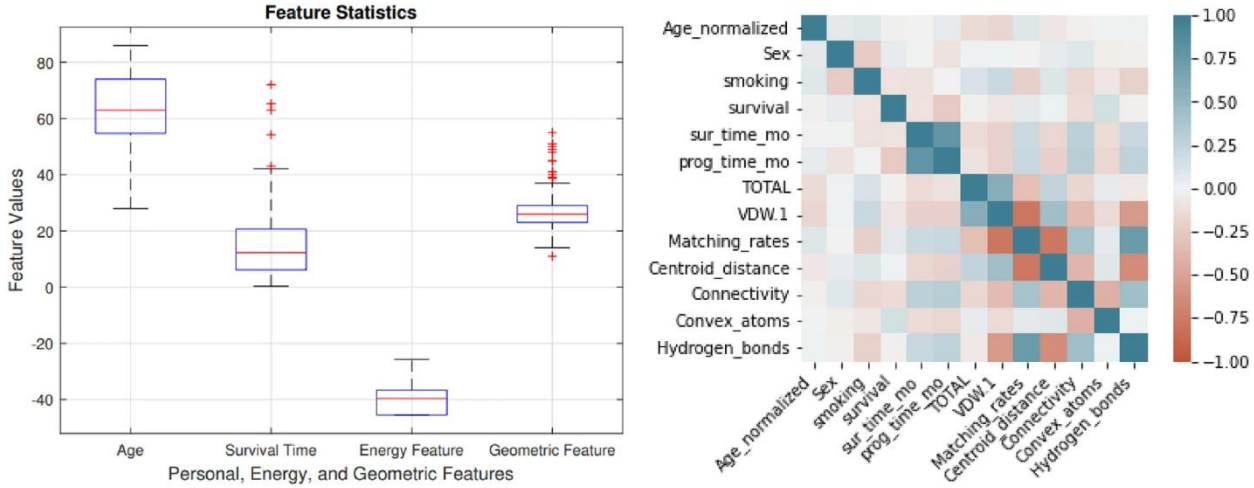

Figure 1. Box plot of normalized values for energy and geometric features (left panel) and correlation between features (right panel)

A total of 33 different HER2 mutations were identified in these patients. The standard deviation in the square root of the molecular dynamics simulation trajectories for the wild type (WT), the four mutation models, and the response classification of the disease are illustrated in Figure 2 for each patient according to the type of mutation. The most frequent mutations observed were L858R, delE746-750, and L858R-T790M. All mutations were simulated based on the 3D structure of HER2 utilizing Rosetta software. The effectiveness of an inhibitor can be evaluated by the duration of survival and the level of drug response. Figure 3 illustrates that there was no linear relationship between drug response and personal or energy features with drug response or survival time. Drug response was categorized into four levels based on the response evaluation criteria in solid tumors (RECIST).

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

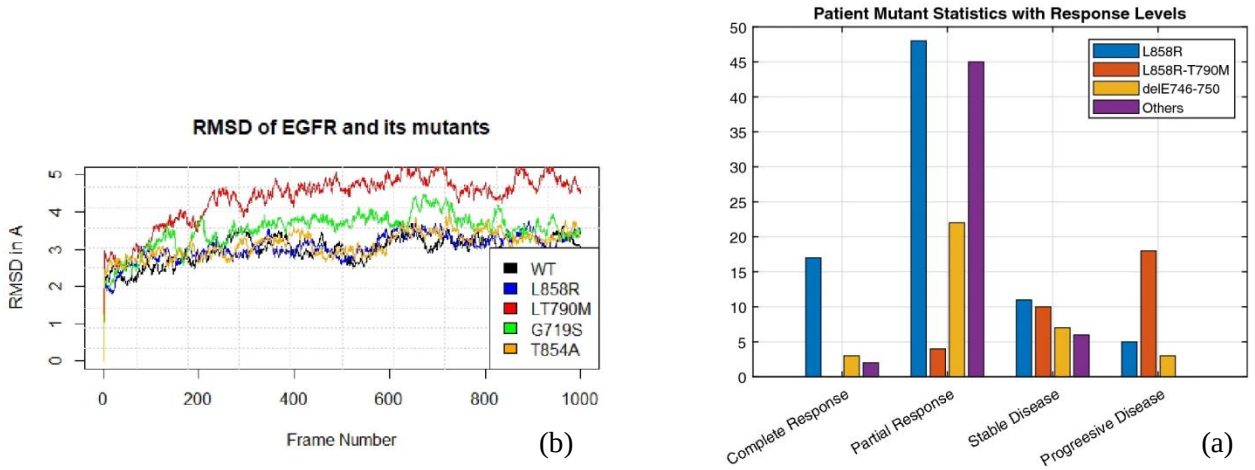

Figure 2. (a) Classification distribution of disease response for 150 patients based on three common mutations (L858R, L858R-T790M, del E746-750) and other mutations. (b) Molecular dynamics trajectories of HER2 and some mutants showing RMSD compared to the reference structure. Since the values are less than 5, the structures are reliable for additional analysis.

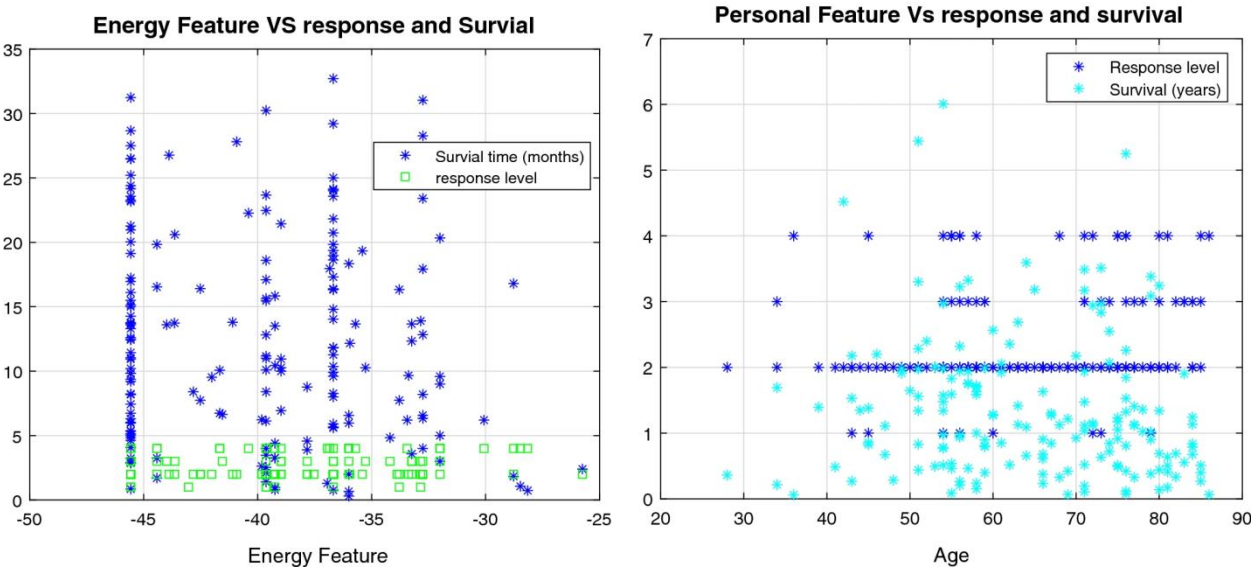

Figure 3- Classification of disease response and survival time (years) based on patient age (right panel); classification of disease response and survival time (months) using free binding energy as the criteria (left panel)

In order to develop the Drug Response Prediction (DRP) model, we integrated clinical data, results from molecular dynamics simulations, and geometric features of protein-drug interactions. It was

Power System Technology
ISSN:1000-3673
Received: 16-01-2024 Revised: 12-02-2024 Accepted: 07-03-2024

observed that geometric features were the most effective predictors, with DCI and energy features closely following based on model performance (Figure 4), based on each type of feature and performance of the relevant model. XGBoost was the top-performing model among those tested, and demographic and clinical features slightly boosted model performance. Combining all three features in XGBoost, random forest, and neural network further improved performance.

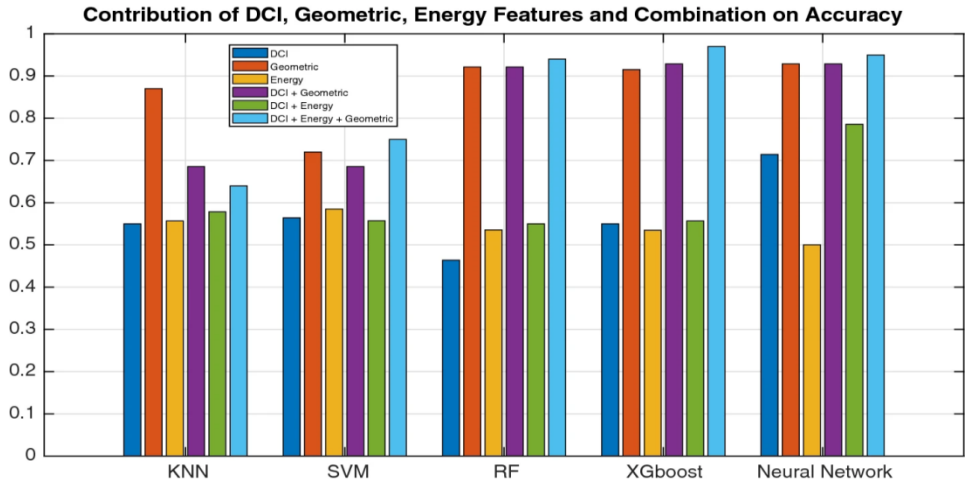

Figure 4- Geometric, DCI, and energy-related feature values in model accuracy

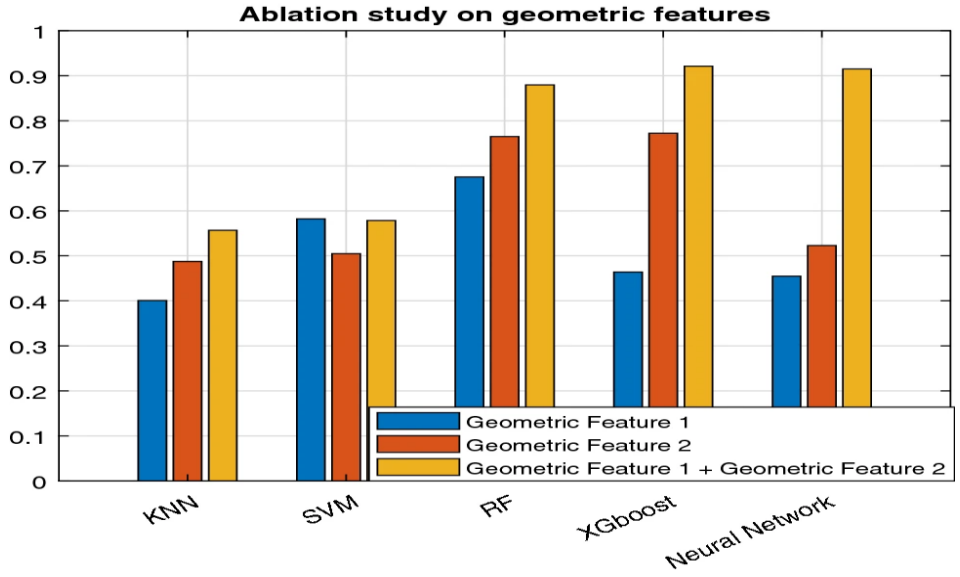

Figure 5- Study of ablation on geometric features

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

Figures (6 and 7) display the error rate matrices for the classification reports of the training set. The XGBoost classifier attained an accuracy of 97.5%, a recall sensitivity of 97%, a specificity of 93%, and an F1 score of 97%, with just two misclassifications in the independent test set of 61 samples. The random forest and neural network models performed similarly to the XGBoost classifier but did not surpass its performance. The classifiers underwent training utilizing a nested cross-validation approach (NCV), and their parameters were optimized through a network search method.

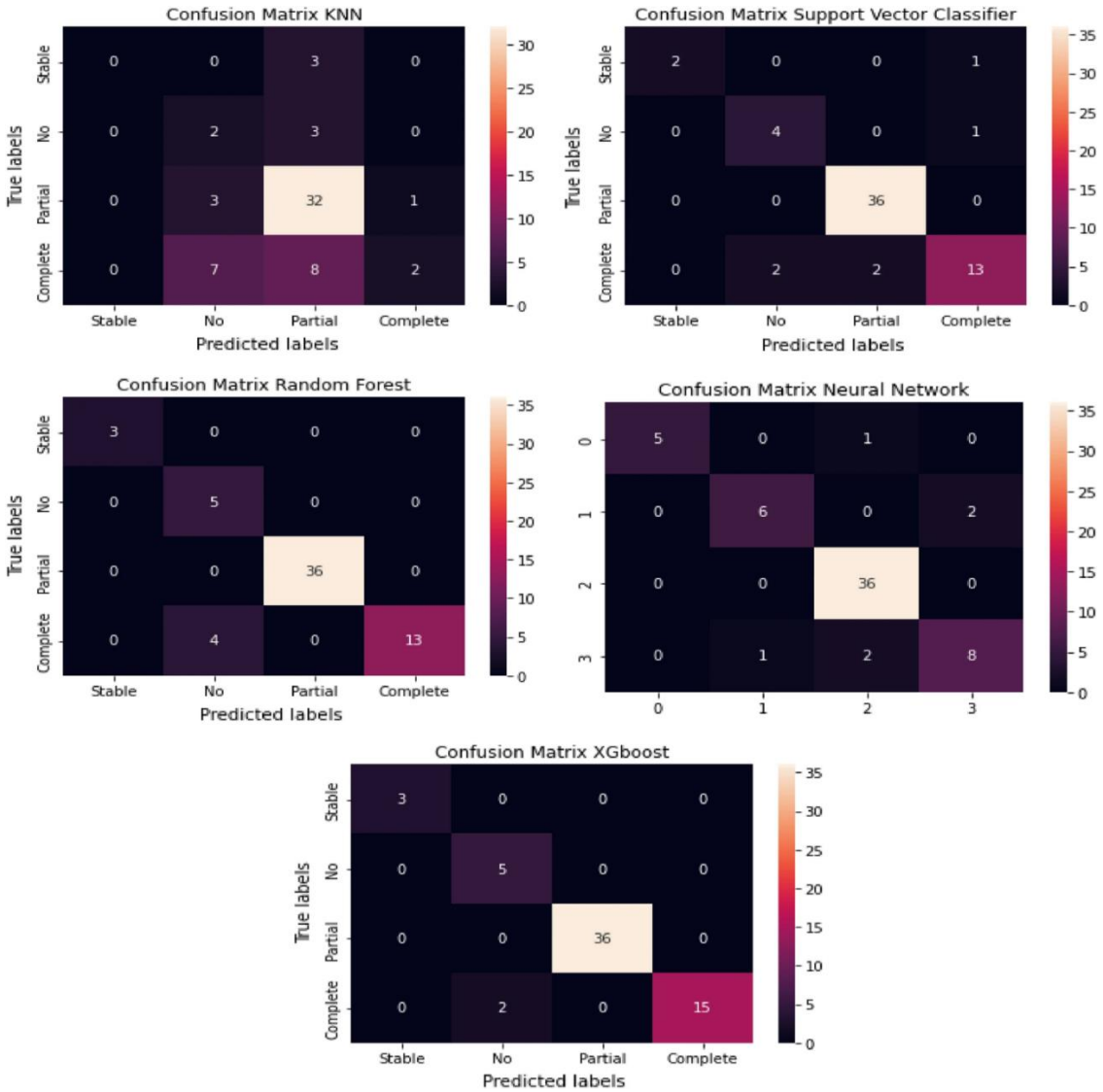

Figure 6- Error matrix for the test dataset

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

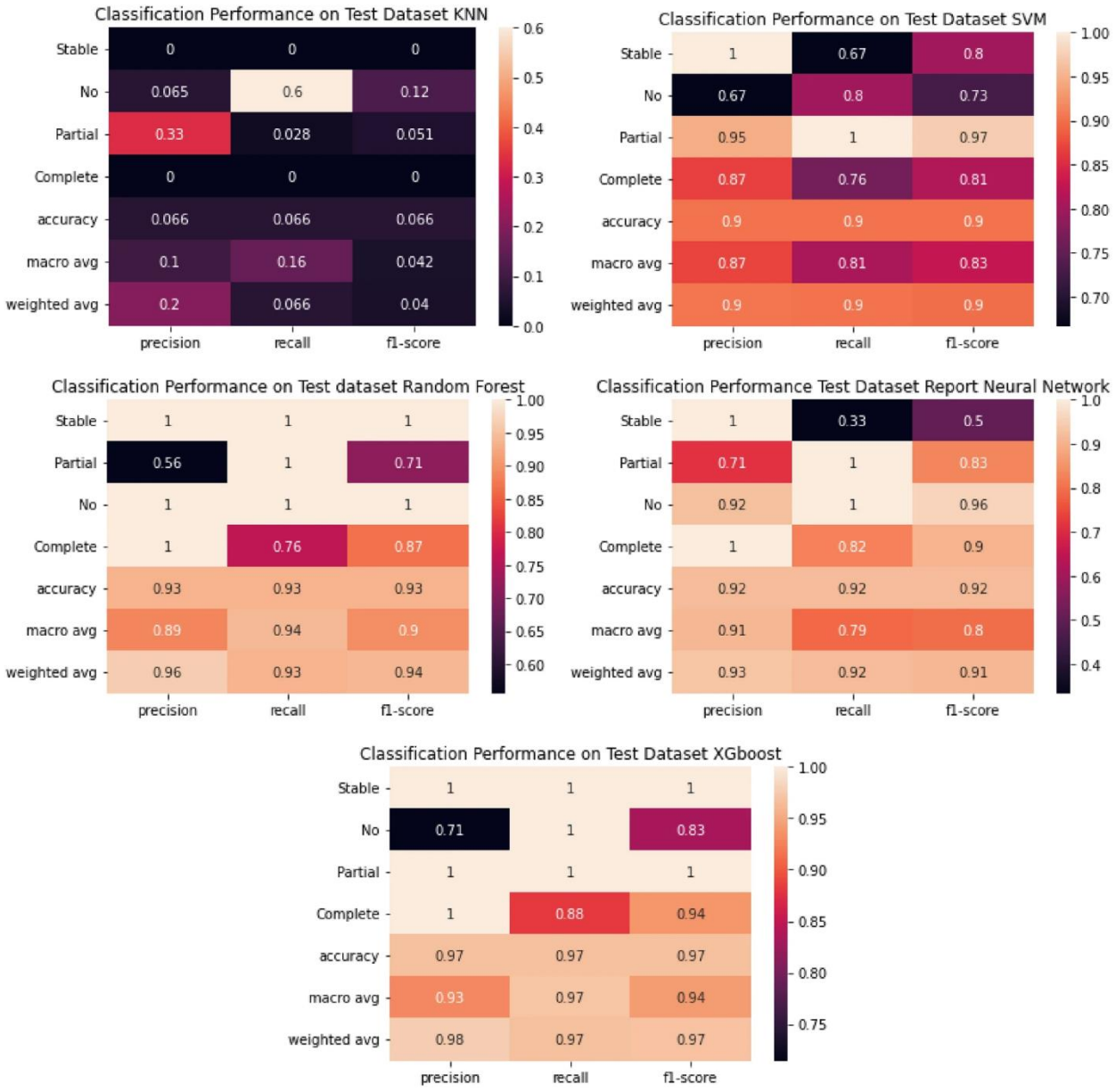

Figure 7- Classification performance of the test dataset

Conclusion

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

The primary aim of this study was to create a model of the drug binding cavity's shape and integrate it with patients' demographic and clinical data. The most distinguishing features in the analysis were the geometric features, such as the number of convex atoms on the interaction surface and the alignment of surface atoms, as well as the dynamic distances between the drug molecule's center and the binding site residues.

Using clinical and molecular predictors proved to be the most successful way of identifying patients sensitive to the drug. When a mutation occurs, it alters the structure of the complex where the drug binds, resulting in changes in geometric features and affecting how the patient responds to the drug. By delving deeper into this model, it may be possible to find geometric features that can more accurately categorize patients based on mutations and tailor treatments based on sex or age. Overall, this model could offer valuable tools to assist in selecting the most suitable drugs for specific patients.

At present, computational techniques, specifically those based on machine learning, are commonly employed for predicting how gastric cancer drugs will function. In this research, a machine learning-driven model for predicting drug responses was created, utilizing various data types such as demographics, clinical data, energy features, and geometric features to predict drug response levels within machine learning classifiers. Despite having access to data from a limited number of patients, the drug response prediction model attained an accuracy rate of 97.5% with the XGBoost classifier, demonstrating top-tier performance.

Our model offers personalized predictions of drug response levels with high accuracy, which can be validated across different types of cancer or diseases. Predicting drug response levels indicates that geometric modeling, including computer simulations, can be a valuable biomarker for anticipating how gastric cancer patients respond to treatment. In upcoming research, we will delve deeper into the dynamics and geometry of protein-drug complex binding sites. We plan to gather additional clinical data to enhance the prediction model and evaluate its effectiveness across various diseases.

References

[1] Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. CA Cancer J. Clin. (2022).
[2] Gupta, G. P. & Massagué, J. Cancer metastasis: Building a framework. Cell 127, 679–695 (2006).
[3] Qureshi, R. et al. Computational methods for the analysis and prediction of HER2-mutated lung cancer drug resistance: Recent advances in drug design, challenges and future prospects. IEEE/ACM Trans. Comput. Biol. Bioinform. (2022).
[4] Kawaguchi, T. et al. Randomized phase iii trial of erlotinib versus docetaxel as second-or third-line therapy in patients with advanced non-small-cell lung cancer: Docetaxel and erlotinib lung cancer trial (delta). J. Clin. Oncol. 32, 1902–1908 (2014).

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

[5] Pao, W. et al. Egf receptor gene mutations are common in lung cancers from "never smokers'' and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc. Natl. Acad. Sci. 101, 13306–13311 (2004).
[6] Zhang, H. Three generations of epidermal growth factor receptor tyrosine kinase inhibitors developed to revolutionize the therapy of lung cancer. Drug Des. Dev. Ther. 10, 3867 (2016).
[7] Singh, D., Attri, B. K., Gill, R. K. & Bariwal, J. Review on HER2 inhibitors: Critical updates. Mini Rev. Med. Chem. 16, 1134–1166 (2016).
[8] Tetsu, O., Hangauer, M. J., Phuchareon, J., Eisele, D. W. & McCormick, F. Drug resistance to HER2 inhibitors in lung cancer. Chemotherapy 61, 223–235 (2016).
[9] Rho, J. K. et al. Combined treatment with silibinin and epidermal growth factor receptor tyrosine kinase inhibitors overcomes drug resistance caused by t790m mutation. Mol. Cancer Ther. 9, 3233–3243 (2010).
[10] Balius, T. E. & Rizzo, R. C. Quantitative prediction of fold resistance for inhibitors of HER2. Biochemistry 48, 8435–8448 (2009).
[11] Tetsu, O., Phuchareon, J., Eisele, D. W., Hangauer, M. J. & McCormick, F. Akt inactivation causes persistent drug tolerance to HER2 inhibitors. Pharmacol. Res. 102, 132–137 (2015).
[12] Guardiola, S., Varese, M., Sánchez-Navarro, M. & Giralt, E. A third shot at HER2: New opportunities in cancer therapy. Trends Pharmacol. Sci. 40, 941–955 (2019).
[13] Wang, D. D., Zhou, W., Yan, H., Wong, M. & Lee, V. Personalized prediction of HER2 mutation-induced drug resistance in lung cancer. Sci. Rep. 3, 1–8 (2013).
[14] Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. Nat. Struct. Biol. 9, 646–652 (2002).
[15] Qureshi, R., Ghosh, A. & Yan, H. Correlated motions and dynamics in different domains of HER2 with l858r and t790m mutations. IEEE/ACM Trans. Comput. Biol. Bioinform. (2020).
[16] Wan, S. & Coveney, P. V. Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor. J. Comput. Chem. 32, 2843–2852 (2011).
[17] Qureshi, R., Zhu, M., Ghosh, A. & Yan, H. Computational analysis of structural dynamics of HER2 and its mutants. in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2784–2791 (IEEE, 2019).
[18] Rizwan, Q., Zhu, M. & Yan, H. Visualization of protein-drug interactions for the analysis of drug resistance in lung cancer. IEEE J. Biomed. Health Inform.(2020).
[19] Peng, Y. et al. Apatinib to combat HER2-tki resistance in an advanced non-small cell lung cancer patient with unknown HER2 status: A case report. Onco Targets Ther. 10, 2289 (2017).
[20] Mehner, C. et al. HER2 as a prognostic biomarker and therapeutic target in ovarian cancer: Evaluation of patient cohort and literature review. Genes Cancer 8, 589 (2017).
[21] Collins, F. S., Morgan, M. & Patrinos, A. The human genome project: Lessons from large-scale biology. Science 300, 286–290 (2003).
[22] Ashley, E. A. Towards precision medicine. Nat. Rev. Genet. 17, 507–522 (2016).
[23] Hoerbst, A. & Ammenwerth, E. Electronic health records. Methods Inf. Med. 49, 320–336 (2010).
[24] Mok, T. S. Personalized medicine in lung cancer: What we need to know. Nat. Rev. Clin. Oncol. 8, 661–668 (2011).
[25] French, B. et al. Statistical design of personalized medicine interventions: The clarification of optimal anticoagulation through genetics (coag) trial. Trials 11, 1–9 (2010).

Power System Technology
ISSN:1000-3673

Received: 16-01-2024 *Revised: 12-02-2024* *Accepted: 07-03-2024*

[26] Ma, L., Wang, D. D., Zou, B. & Yan, H. An eigen-binding site based method for the analysis of anti-HER2 drug resistance in lung cancer treatment. IEEE/ACM Trans. Comput. Biol. Bioinform. 14, 1187–1194 (2016).
[27] Basit, S. A., Qureshi, R., Shahid, A. R. & Khan, S. Survival prediction of lung cancer patients by integration of clinical and molecular features using machine learning. in 2021 15th International Conference on Open Source Systems and Technologies (ICOSST), 1–6 (IEEE, 2021).
[28] Berman, H. M. et al. The protein data bank. Nucleic Acids Res. 28, 235–242 (2000).
[29] Wang, R. et al. Taxirec: Recommending road clusters to taxi drivers using ranking-based extreme learning machines. IEEE Trans. Knowl. Data Eng. 30, 585–598 (2018).
[30] Duan, B., Zou, B., Wang, D. D., Yan, H. & Han, L. Computational evaluation of HER2 dynamic characteristics in mutation-induced drug resistance prediction. in 2015 IEEE International Conference on Systems, Man, and Cybernetics, 2299–2304 (IEEE, 2015).
[31] Zou, B., Lee, V. H. & Yan, H. Prediction of sensitivity to gefitinib/erlotinib for HER2 mutations in nsclc based on structural interaction fingerprints and multilinear principal component analysis. BMC Bioinform. 19, 1–13 (2018).
[32] Lee, V. H. et al. Association of exon 19 and 21 HER2 mutation patterns with treatment outcome after first-line tyrosine kinase inhibitor in metastatic non-small-cell lung cancer. J. Thoracic Oncol. 8, 1148–1155 (2013).
[33] Ma, L. et al. HER2 mutant structural database: Computationally predicted 3d structures and the corresponding binding free energies with gefitinib and erlotinib. BMC Bioinform. 16, 1–10 (2015).
[34] Zou, B. et al. Deciphering mechanisms of acquired t790m mutation after HER2 inhibitors for nsclc by computational simulations. Sci. Rep. 7, 1–13 (2017).
[35] Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the amber biomolecular simulation package. Wiley Interdiscip. Rev. Comput. Mol. Sci. 3, 198–210 (2013).
[36] Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. Protein structure prediction using rosetta. in Methods in Enzymology, vol. 383, 66–93 (Elsevier, 2004).