Personalized drug-response prediction model for lung cancer patients using machine learning

Rizwan Qureshi, Xinqi Fan, Mengxu Zhu *Student members, IEEE* and Hong Yan, *Fellow, IEEE*

***Abstract*—Lung cancer caused by mutations in the epidermal growth factor receptor (EGFR) is a major cause of cancer deaths worldwide. EGFR Tyrosine kinase inhibitors (TKIs) have been developed, and have shown increased survival rates and quality of life in clinical studies. However, drug resistance is a major issue, and treatment efficacy is lost after about an year. Therefore, predicting the response to targeted therapies for lung cancer patients is a significant research problem. In this work, we address this issue and propose a personalized model to predict the drug-response of lung cancer patients. This model uses clinical information, geometrical properties of the drug binding site, and the binding free energy of the drug-protein complex. The proposed model achieves state of the art performance with 97.5% accuracy, 100% recall, 95% precision, and 96.3% F1-score with a random forest classifier. This model can also be tested on other types of cancer and diseases, and we believe that it may help in taking optimal clinical decisions for treating patients with targeted therapies.**

***Index Terms*—Binding free energy, Cancer, Drug response, EGFR mutations, Machine learning, Personalized medicine, Protein-drug interactions**

I. Introduction

LUNG cancer is a leading cause of cancer deaths worldwide [1], and has the lowest survival rate among all cancer types. It is the second most common type of cancer, and often diagnosed at later stages when metastatic spread to other parts of the body may have occurred [2] [3].

In the last decade, great progress has been made in the management of non-small cell lung cancer NSCLC patients. Molecular targeting has made great advancements, and EGFR and ErbB family members have been identified as a useful therapeutic targets [4]. Over-expression of EGFR is found in about 60% of advanced NSCLC patients [5]. The US Food and Drug Administration (FDA) has approved small molecule inhibitors Gefitinib/Erlotinib as a first line treatment for lung cancer patients, harboring EGFR mutations [6].

Small molecule inhibitors produced encouraging results at the initial stage of therapy, and increased the survival rate and life quality of patients. However, drug resistance appeared due to secondary point mutation(s), which limited the effectiveness of the drug [7]. Using molecular dynamics, several computational studies have been conducted to decode the mechanism of drug resistance simulations [8], [9]. These studies provided useful insights about the conformational dynamics [10], stability [11], and structural changes [12] of EGFR and explaining several aspects of drug resistance. Recently, a framework is developed for the visualization of protein-drug interaction for lung cancer drug resistance analysis in [13]. There is still much unexplained variation in a patient's response to these drugs and the patient's personal characteristics may play a significant role in the mechanism of drug resistance.

The completion of human genome project [14] has allowed a move from the traditional medical model of targeting a large population, as a one-size fits all approach [15], towards personalized therapies. Information from genomic and genetic data provides new opportunities for patient care, prevention, and diagnosis [16].

Predicting a patient's response to a drug treatment [17] [18], or identifying their optimal treatment strategy, is challenging for computational methods due to limited data sources. Disease outcomes have been modeled for breast [19], and lung [20] cancers, and for large B-cell lymphoma [21] using clinical and molecular structural information and for NSCLC patients using supervised machine learning [22].

As response to a drug is often mediated by a protein-drug interaction, the geometry of the drug binding site, or pocket, and molecular dynamics (MD) modeling of the binding site, can be useful predictors. MD simulation of the binding energy of drug-mutant complexes and patient's personal characteristics when combined with an extreme learning machine (ELM) [23] could classify the drug response level into two classes [24]. They achived an accuracy of 95.3%. In another interesting work, Local geometrical properties combined with energy related features in an Eigen binding site method achieved 69.35% accuracy for four classes of drug responses [25] and personal and geometric features were used to make a similar prediction in [26], while Bin *et al.* used protein-drug interaction footprints in a three level drug response classificationl [27].

These studies demonstrate the potential of combining dynamic molecular features with patient's personal characteristics to predicting drug responses, but the quality of the predictions needs to be improved for potential clinical use. In this work, we combine geometry, energy and patient's personal information to predict four classes of drug response. The proposed model achieves state of the art performance with 97.5% accuracy, 100% recall, 95% precision, and 96.3% F1 score. The main contributions of this work can be stated as follows.

- We have developed a drug response prediction model using molecular dynamics simulation and machine learning, which achieves state of the art performance.
- Our work demonstrates the contribution of geometrical features to increase in prediction accuracy.

The authors are with City University of Hong Kong.

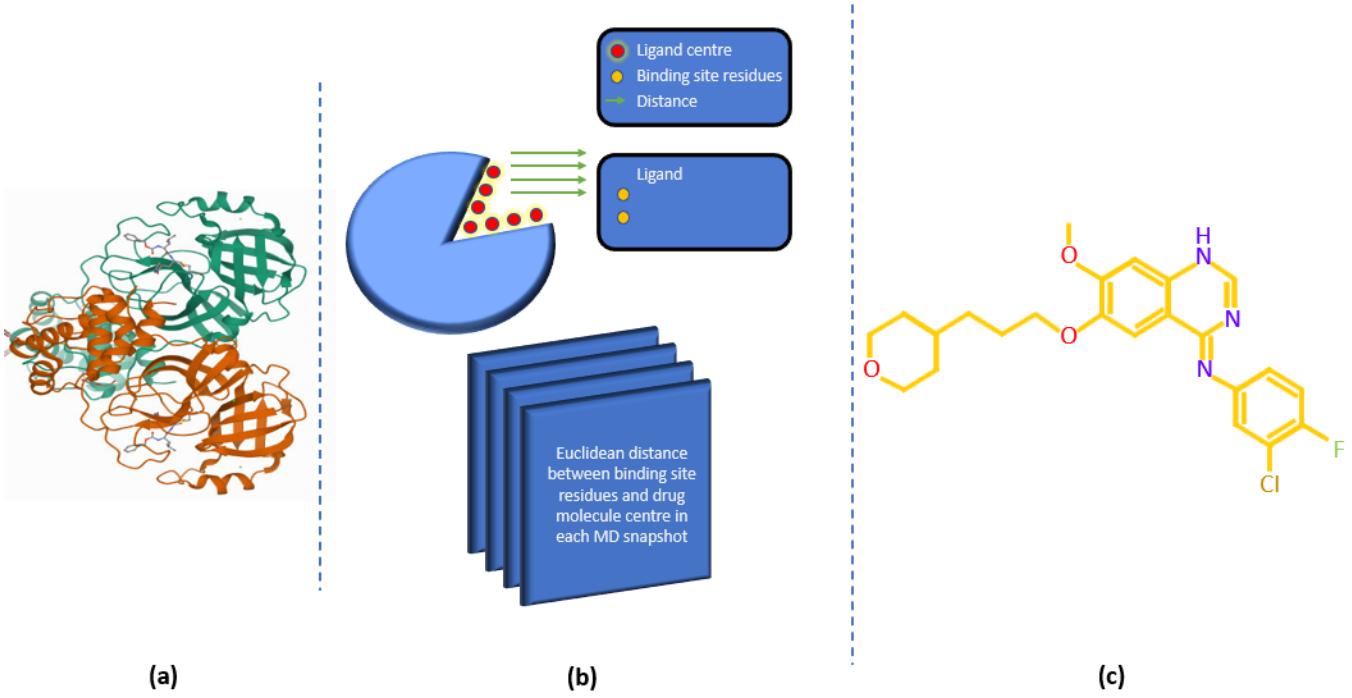

Fig. 1. Crystal structure of EGFR with Gefitinib (a). Euclidean distance between drug binding site residues (b), and (c) shows the molecular structure of a drug.

- As the proposed model is a general model, it can be tested on other types of cancer and and diseases and used in clinical decision support with minor modifications.

The paper is organized as follows. In section II, we formulate an improved method to classify a patient's response to drug treatment. Sections III, IV and V present the proposed methodology, geometrical feature extraction and classification, while results and discussion are given in section VI. Conclusions and future work are given in section VII.

II. Formulation of the Classification Model

The framework to classify individual patient outcomes is divided into three modules; computational modeling of mutant/drug structures, MD simulations, and classification. Figure 1 (a) shows the crystal structure of the EGFR dimer in complex with Gefitinib and this is the template from which mutant structures are modeled. Figure 1 (b) and (c) show the key interactions between the protein and drug and the detailed chemical structure of Gefitinib. Prediction of mutant structures from the template by computational modeling, using the Rosetta modeling tools [28], is shown in Figure 2 (a), MD simulations are shown in Figure 2 (b), and the classification module is shown in Figure 2 (c).

A. Datasets and Patients

The clinical information to conduct such study is always a major challenge. The clinical information used in this study was taken from several published sources [24], [25], [29]–[31]. A dataset of 201 NSCLC patients was obtained. These patients had a median age of 63 years, 35% (71) were female and 65% (130) male, and about 75% were non-smokers. All patients received EGFR-TKIs as their first line of treatment. A total of 31 different EGFR mutations occurred in these patients at frequencies shown in Figure 3. The most common mutations were L858R, delE746−750 and L858R−T790M. All mutations were modelled into the EGFR 3D structure using Rosetta [28].

The potency of an inhibitor can be measured by a patient's survival time or their drug response level. Drug response is classified into four levels, based on the response evaluation criteria in solid tumors RECIST [32]. Response levels 1 and 2 indicate complete and partial responses to the drug. Response levels 3 and 4 correspond to stable and progressive disease. The dataset used here consisted of 19, 118, 30, 34 patients at response levels 1, 2, 3 and 4, respectively. The dataset was divided into training (80%, 163 patients) and testing (20%, 38 patients) subsets.

B. Personal features

The personalized information used for each patient was age, sex, smoking history, performance status, and drug response level. Age was coded as [0,4] based on the ranges (0,40), (41, 50), (51, 60), (61, 70), and (70+). Figure 4 shows that the drug response level and survival time are not correlated with the age of a patient.

A detailed description of the features and their value ranges are presented in Table I. For a patient with specific personal and clinical information, the energy and geometric features of

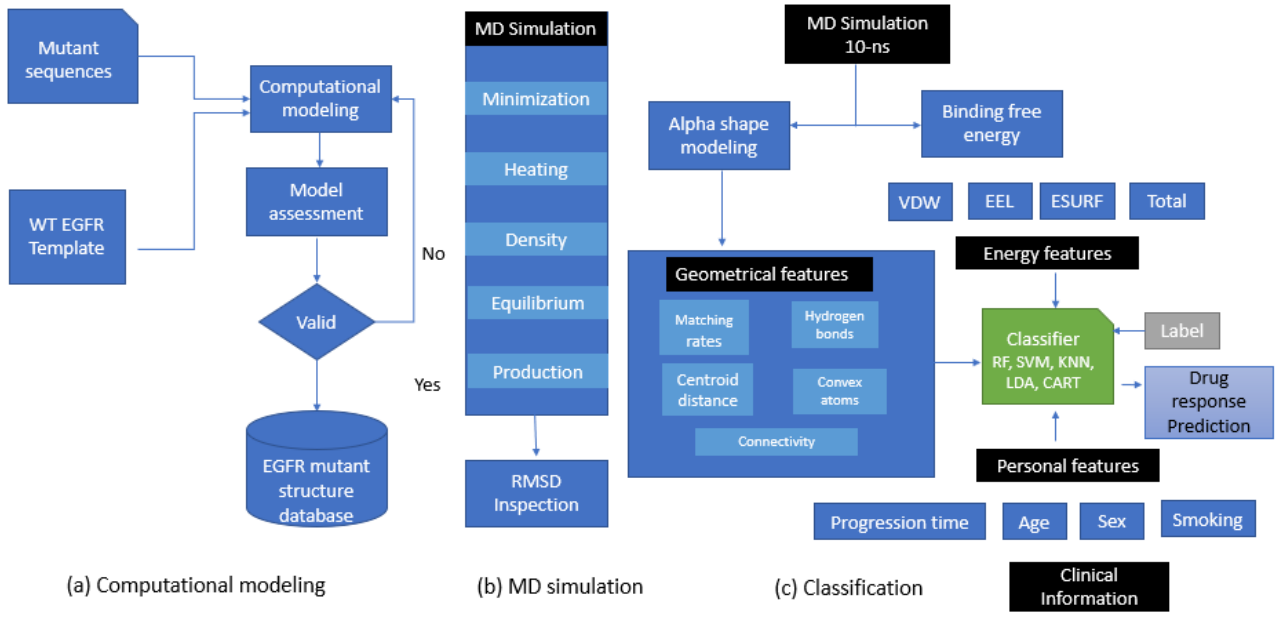

Fig. 2. The framework for predicting the drug response in lung cancer patients based on personal data, energy, and geometric features. Mutant structures are predicted by computational methods then molecular dynamics simulations extract energy and geometrical features. Machine learning classifiers then predict four classes of drug response from these features

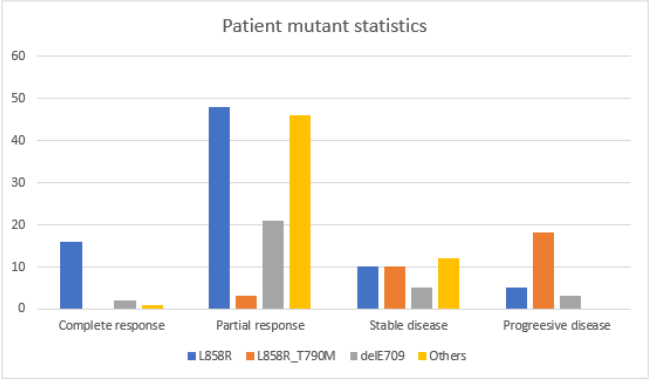

Fig. 3. Distribution of mutation statistics for 201 patients with drug response. L858R, L858R-T790M, and delE746-750 were the most common mutants.

their mutant EGFR-Gefitinib complex were obtained and used to predict their drug response level through machine learning classifiers.

III. Proposed Methodology

The framework for predicting drug response level is divided into three parts (Figure 2). Initially, the mutation is modeled into the EGFR 3D structures so that MD simulations [33] can be performed to extract the energy and geometric features of the protein-drug interaction which are passed to machine learning classifiers to predict the drug response level.

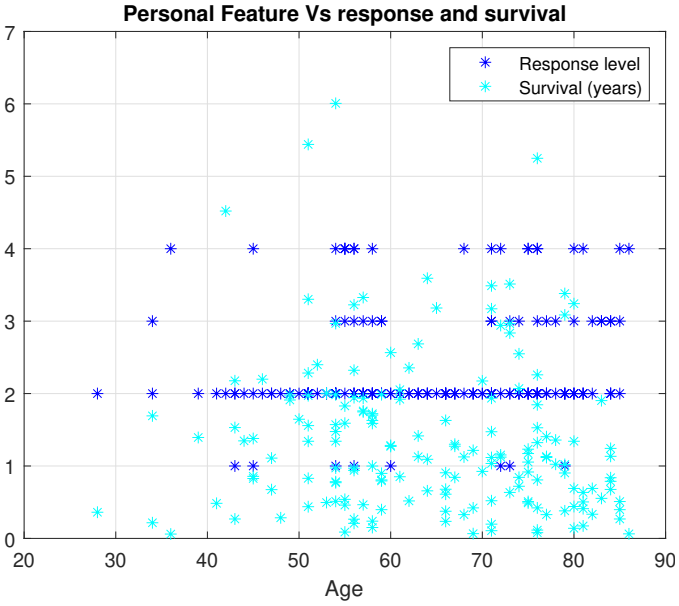

Fig. 4. Drug response and survival time by patient age

A. Computational modeling of the structures

The 3D mutant structures are predicted based on the crystal structure of wildtype EGFR, taken from Protein Data Bank (PDB) [34]. The high resolution ddgmonomer (HRDM) [35] protocol in Rosetta is used to predict point mutations, and the comparative modeling protocol is used to predict multi-point mutations [36]. Quality assessment of predicted structures is performed by Verify3D [37], and Q-mean [38].

Feature type	Attributes	Description	Discrete/Continuous	Range
Clinical information	Age	Patient's personal information	Discrete	[0 - 4]
	Sex		Discrete	[0 - 2]
	Smoking history		Discrete	[0 - 2]
	Performance status		Discrete	[0 - 2]
Energy feature	VDW	Binding free energy between protein-drug mutant	Continuous	[-60 - -45]
	EEL		Continuous	[-23 -11]
	ESURF		Continuous	[-45 - -1]
	EPB		Continuous	[27 - 40]
Geometrical features	Matching rates	Matched atoms	Discrete	[0, 17]
	Convex atoms	Strength of interaction	Discrete	[0, 43]
	Connectivity	Connected atoms	Discrete	[0, 23]
	Euclidean distance	Distance between drug and target	Continuous	[30 - 39 Å]
	Hydrogen bonds	Number of hydrogen bonds	Discrete	[775 - 1650]

TABLE I
CLINICAL INFORMATION AND ENERGY AND GEOMETRICAL FEATURES: DESCRIPTION AND VALUES

B. Molecular dynamics simulation

MD simulations of the protein-drug complex were performed using the QM/MM method in Amber [39] with a surrounding waterbox neutralized using Na+ and Cl- atoms and the ff9SB [40] and gaff force fields. The total energy of the system is the sum of the bonded (stretch, bend, torsion) and non-bonded (electrostatic, van derWaals) terms.

$$E_{total} = E_{stretch} + E_{bend} + E_{torsion} + E_{electrostatic} + E_{vdw} \quad (1)$$

Energy minimization is performed to refine the modeled structure bfeore the MD run, which starts with a heating of the system from 0 ° K to 300 ° K, followed by density equilibrium for 50-ps and constant pressure for 500-ps. The SHAKE [41] algorithm was used to constrain bond stretching and for efficient temperature control. After achieving a stable state, production MD runs were performed at constant temperature (300 ° K) and pressure (1 atm) for 2-ns. The MD simulations were performed using a 12 core 3.47 Ghz I-7 processor, with 8 GB RAM [42]. A Tesla C2075 GPU [43] was used for production files. Each simulation was completed in about 12 hours. The CPPTRAJ [44] package in Amber was used to extract the trajectory with frames collected every 10-ps, giving 200 frames for each run.

1) Root mean square deviation: The root mean square deviation is used to measure the spatial variance between a reference structure and superimposed structures.

$$RMSD(t) = \sqrt{\frac{1}{N}\sum_{i=0}^{N} ||X_i - Y_i||^2} \quad (2)$$

where N is the number of frames, X_i is the target and Y_i is the reference structure. RMSD [45] is used to measure the stability of the MD simulations. The trajectories of the RMSD values of EGFR and its mutants from the reference structure are shown in Figure 5.

2) Binding free energy: The free energy of binding [46] of a drug to a protein in a solvent environment estimates the binding affinity [47]. The parallel version of MM-GBSA [48] on a 12 core, 3.47 GHz processor is used for the simulation.

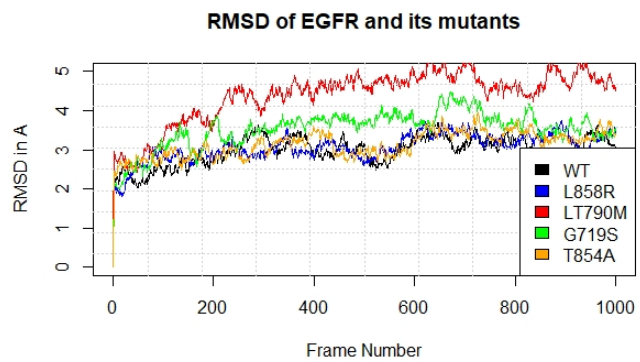

Fig. 5. RMSD trajectories of EGFR and its mutants from the reference structure. As the values are within 5 Å, the structures are reliable for further analysis.

The MD trajectory was input to the MM-GBSA, and each simulation took about 12 hours for computation. The binding free energy is calculated based on the theory of the thermodynamic cycle in vacuum and solvent environments [49], as:

$$\Delta G = \Delta G_{Bind,Vacuum} + \Delta G_{Solv,Complex} - (\Delta G_{Solv,ligand} + \Delta G_{Solv,Receptor}) \quad (3)$$

where ΔG is the binding free energy difference of the receptor-ligand system in a vacuum. $\Delta G_{Solv,Complex}$, $\Delta G_{Solv,ligand}$, and $\Delta G_{Solv,Receptor}$ represent their energy differences between vacuum and solvent states.

The energy component is composed of Van der Waals forces (VDW), electrostatic energy (EEL), the electrostatic contribution to solvation, and non-polar contributions to the solvation free energy (ESURF). The binding free energy and its components for EGFR and its mutants are shown in Figure 6.

In this work, we have used the binding free energy and its components as energy features in our prediction model. Drug-sensitive mutants generally have higher binding energy values than drug resistant mutants. The energy features vs response level and survival are shown in Figure 7. Energy feature are not

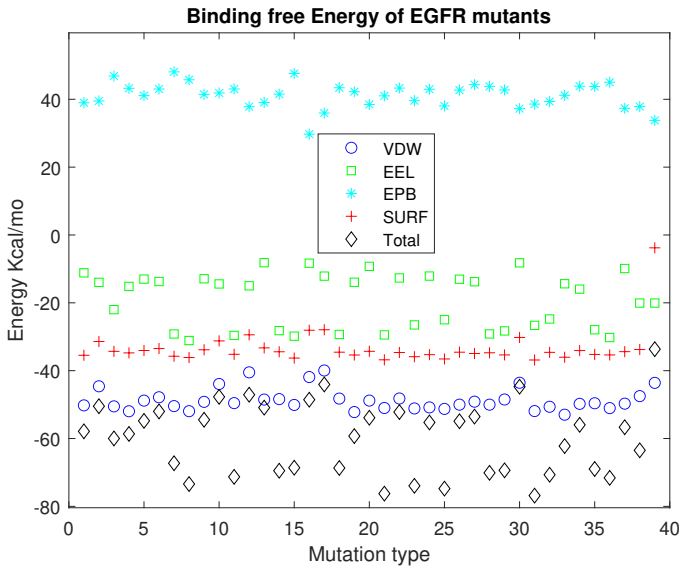

Fig. 6. Binding free energy between the EGFR-mutants and the EGFR-TKI

one-to-one or linearly related to the response level, indicating an influence of patient specific features.

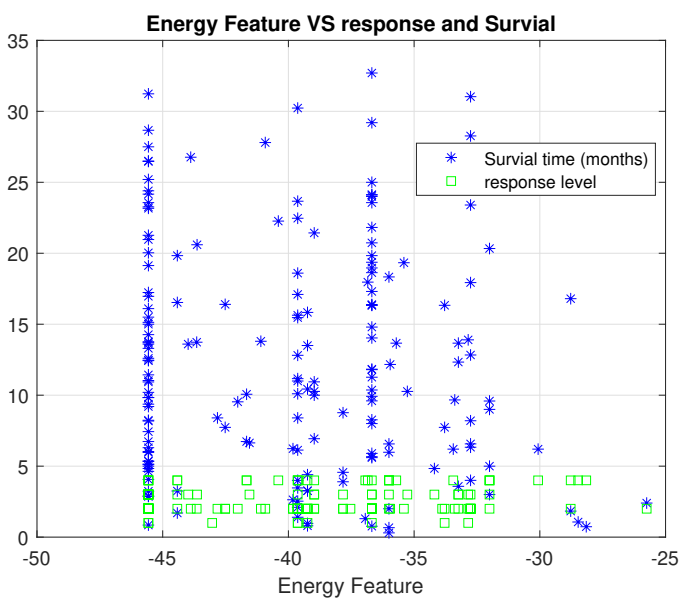

Fig. 7. Binding free energy vs drug response levels and survival times

IV. Geometrical Features

Interactions between the binding site residues of a protein and small molecule inhibitors are commonly used in prediction methods [50]. Local geometric surface properties were determined based on the alpha shape [51] using the computational geometry algorithm library (CGAL) [52].

A. Alpha shape

The theory of the alpha shape algorithm is based on 3D Delaunay triangulation, which aims to maximize the minimum angle of all the angles of the triangle in triangulation [53]. Given four atoms A, B, C, D, in 3D space, after successful triangulation, no atom will be located in the circumcircle of any triangle. The Delaunay algorithm maximizes the angle using the following rule.

$$\begin{bmatrix} x_A & y_A & x_A^2+y_A^2 & 1 \\ x_B & y_B & x_B^2+y_B^2 & 1 \\ x_C & y_C & x_C^2+y_C^2 & 1 \\ x_D & y_D & x_D^2+y_D^2 & 1 \end{bmatrix} > 0$$

Here x_A, y_A shows the location of atom in a 2D-plane. If the determinant is positive, the atom D lies in the circumcircle of A, B, C, as shown in Figure 8.

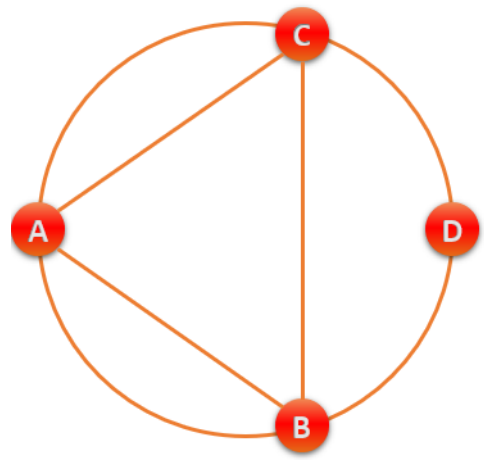

Fig. 8. Unsuccessful triangulation using the Delaunay algorithm. Atom D lies in the circumcirle of A, B, C

1) Convex atoms: Each atom in a drug-mutant system has a position and mass, represented as a = (p, w), where p is the position and w is the mass of the atom. Two atoms a_1 = (p_1, w_1) and a_2 = (p_2, w_2) are defined as orthogonal or sub-orthogonal using the following equation.

$$\begin{cases} |p_1p_2| = w_1 + w_2, & a_1 \perp a_2 \\ p_1p_2| = w_1 > w_2, & a_1 \perp_s a_2 \end{cases}$$

From the alpha shape, the solid angle [54] of atoms was determined to characterize the geometric properties of the local surface . If A, B, C, and D are the vertices of a tetrahedron, the solid angle, Ω_i, is:

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

where ϕ_i^{AB}, ϕ_i^{BC}, and ϕ_i^{BC} represent the dihedral angles of tetrahedron i.

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

Ω' results in a convex shape if positive and a concave shape if negative (Figure 9). The number of atoms in a convex shape at the local surface of the drug-dimer complex are used.

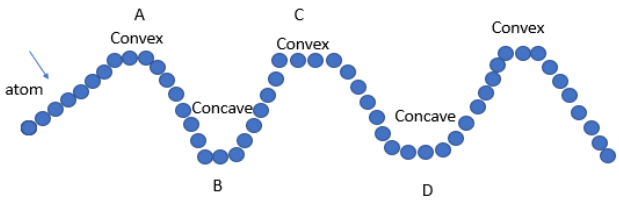

Fig. 9. Atoms in Convex and Concave shapes at the surface curvature. At point A and B, the atoms are matched and there is a strong interaction between them, whereas at point A and C, the atoms are unmatched, and there is a weak interaction.

2) Matching rates: The atoms at the interface of the structures create the interaction between the drug and the target. The surface atoms are collected using the alpha shape algorithm, and named as point set A. After that, point sets B and C are obtained, to represents the surface atoms of the target and the drug, respectively. The interacting atoms (I) are obtained using set operations and further classified as, interacting atoms on the drug I_d or target I_t.

$$\begin{cases} I = (B \cup C) - A \\ I_t = (I \cap B) \\ I_d = (I \cap C) \end{cases}$$

(6)

I represents the interacting atoms, I_t the interacting atoms in the target, and I_d the interacting atoms in the drug. The matching rate is determined by selecting atoms at the drug and the target. If one of the atoms is convex and other is concave, the pair is recorded as matched and there is a strong interaction between them. If both atoms are convex or concave, the pair is unmatched, and the interaction is weak. The matched and unmatched atoms are determined as:

$$f(B,C) = \begin{cases} 1 & \Omega_B \times \Omega_C < 0 \\ 0 & otherwise \end{cases}$$

(7)

Matching rates are calculated for each frame of the MD trajectory as:

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

MR represents the matching rate, $f(B_i, C_i)$ is a matched atom pair, and N is the total number of MD snapshots. The matching rate is used as a feature in this work, and low matching rates were linked to low drug responses.

3) Connectivity measure: Connectivity changes between binding site residues and the drug molecule throughout the MD simulation. The consistency of these connections could be used as a predictor of the drug response level.

$$C_{k,i} = \sum_j A_{k,i,j} \tag{9}$$

Where $A_{k,i,j}$ represents the connection between the ith EGFR atom and the jth drug atom in the kth MD snapshot, and is 1 if there is a connection and zero otherwise.

$$D_{k,i} = \sum_i C_{k,i} > 0 \tag{10}$$

D_k represents number of connected atoms in the MD snapshot. The number of connected atoms over the entire trajectory is used as a feature.

$$E_k = \frac{\sum_{i=1}^{N} (D_{k,i})}{N} \tag{11}$$

4) Binding site positioning: The positioning is evaluated using the Euclidean distance between drug-binding site atoms and the center of the drug-molecule.

$$D(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2} \tag{12}$$

The binding site residues are represented by their CA atoms, and if there are 14 CA atoms at the binding site, and two atoms at the drug molecule center, then a 14 $\times$ 2 or 28 $\times$ 1 vector will represent this. The distance over the entire MD simulation of 200 frames can be represented as 200 $\times$ 28 matrix. The binding site position is represented as the average distance between the drug and the target.

$$D_{avg} = \frac{\sum_{i=1}^{N} (D_i)}{N} \tag{13}$$

where D_{avg} shows the binding site position, D_i shows the ith MD snapshot distance, and N shows the number of MD snapshots. All the feature values were normalized to [0,1]. Figure 10 shows the geometrical features relative to the response level. Generally, drug sensitive mutants have less distance between the drug and the target.

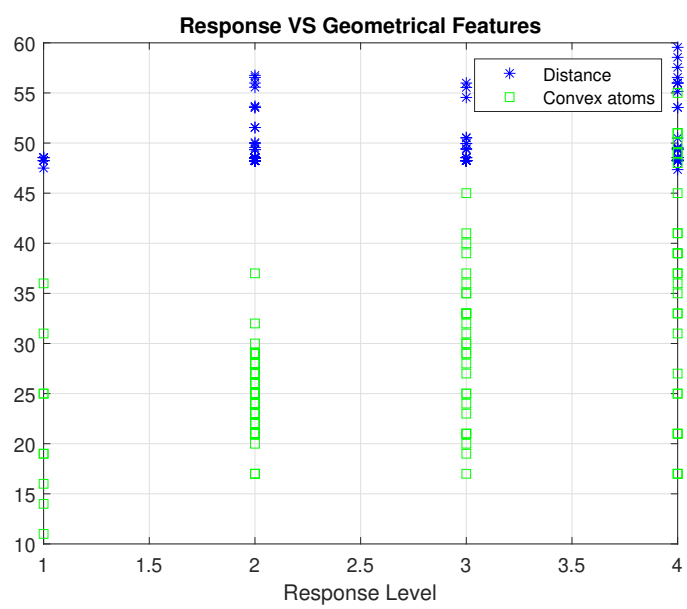

Fig. 10. Geometric features by response level. As the distance and number of convex atom increases, the response level also increases.

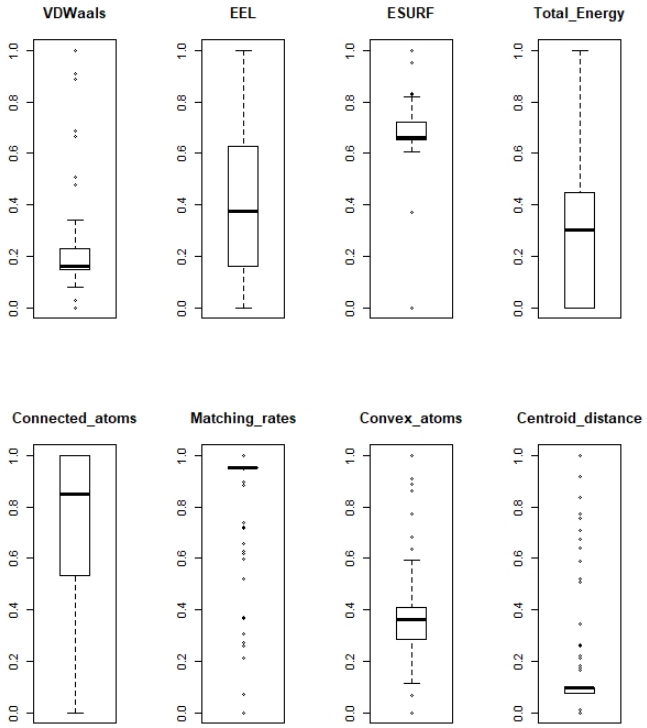

Fig. 11. Normalized values for energy, and geometrical features

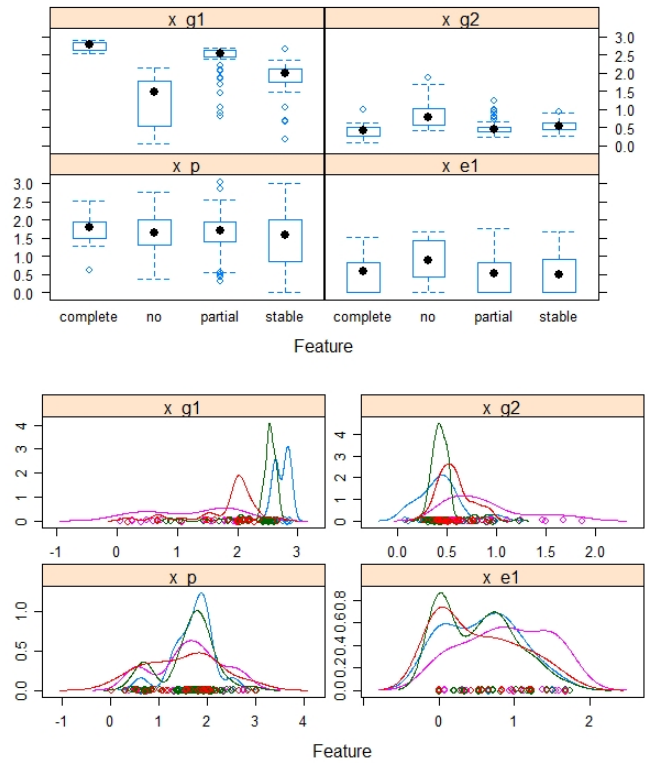

Fig. 12. Feature and density distributions. The geometrical features are most discriminative.

5) Hydrogen Bonds: Hydrogen bonds contribute to the stability of a structure and can provides insights about interactions within the structure. Stable systems tend to have more hydrogen bonds. The number of hydrogen bonds in the EGFR-drug complex were calculated using the hbond command in Amber.

B. Composite Geometric Features

The geometric features were combined to make two composite features, with the matching rate, number of connected atoms, and number of hydrogen bonds as one feature, x_{g1} and the number of convex atoms and Euclidean distance, x_{g2}, as the other. Similarly, the personal features and energy features were combined as x_p and x_{e1}, respectively.

C. Feature normalization

Each feature was normalized to the [0, 1] range using min - max normalization.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{14}$$

where z_i represents the normalized value, $min(x)$ represents the minimum value, and $max(x)$ represents the maximum value for each feature. The normalized values are shown in Figure 11. Composite features and density distributions are presented in Figure 12.

V. Classification

For patients with clinical information, geometrical and energy features were obtained from their EGFR mutant drug complex, and classifiers were trained to predict one of the four-classes of drug response level. Five popular classifiers were tested using Rstudio with the CARET package [55] and 10-fold cross-validation.

A. Classification performance metric

The classification performance metrics used are precision, recall, F1-measure and balanced accuracy.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

$$F1 - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{17}$$

$$BalAcc = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

The terms TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative respectively. The performance metrics are shown in Figure 13.

We also used Kappa (κ) statistic to deal with multi-class classification and and imbalanced problems.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{19}$$

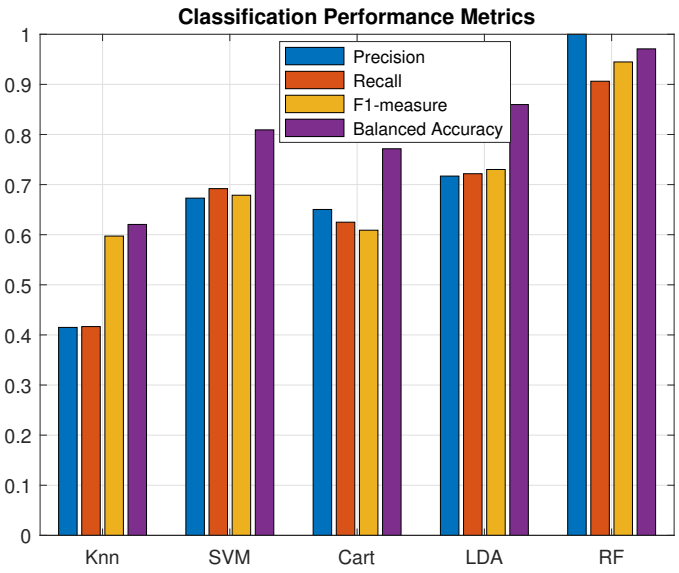

Fig. 13. Classification performance metric. Random forest achieved the best performance with 100% precision and 97.5% balanced accuracy.

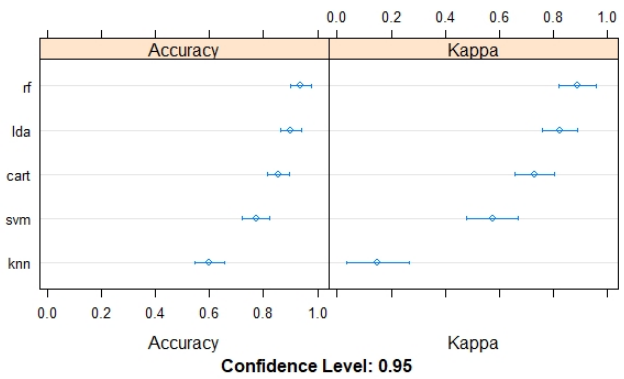

Fig. 14. Dot plots for classification accuracy and Kappa values using a 10-fold cross-validation

where p_o is the observed agreement and p_e is the expected agreement.

VI. Results and Discussion

In this work, we have designed a drug-response prediction model for lung cancer patients. Clinical data, age, survival, sex, smoking history, and type of mutation, were collected from various previous studies [24], [25], [29]–[31] and were used with energy and geometric features from the EGFR mutant-drug complex to predict the drug-response level as one of complete response, partial response, stable disease and progressive disease. The accuracy of the five classifications methods on these data and their Kappa scores [56] showing the agreement between the reference and predicted values are shown in Figure 14. The average accuracy of Knn at 71% with Kappa at 39% is lowest, whereas the random forest achieves an accuracy of 97.5% with Kappa at 95%.

The confusion matrices for training and testing are shown in Figure 15 and 16 respectively. The random forest classifier achieves 100% accuracy with no mis-classification for training data and only one mis-classification in the testing data.

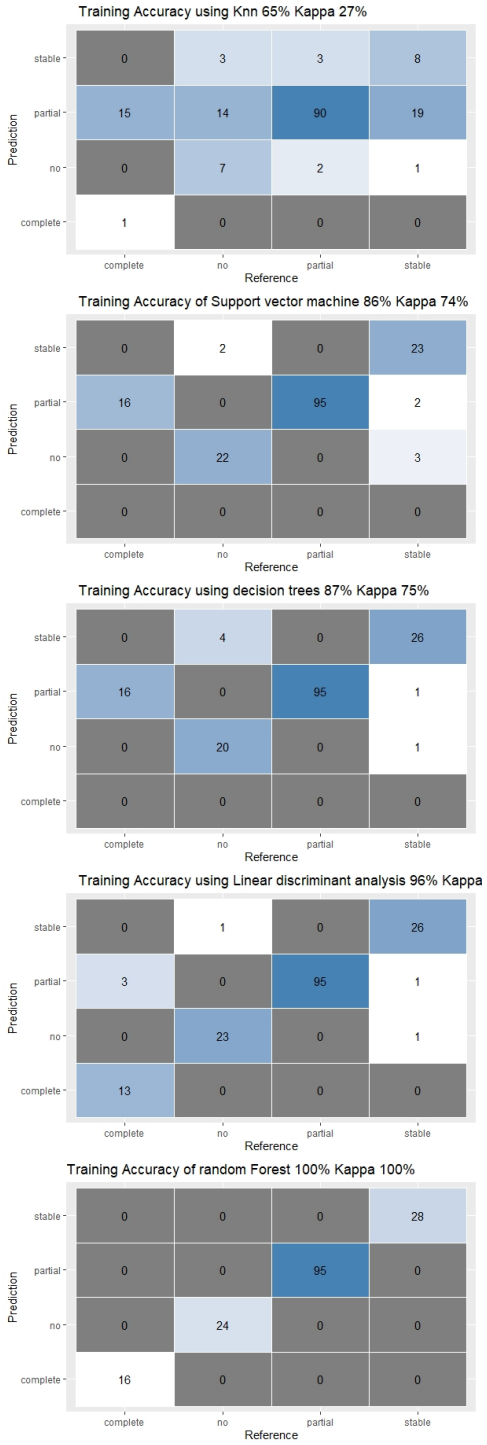

Fig. 15. Confusion matrix for the training dataset. The labels correspond to complete response, no response, partial response and stable disease.

A. Comparison

Table II shows that the classification accuracy of the proposed method achieves state of the art performance relative to related works and for prediction using four classes of

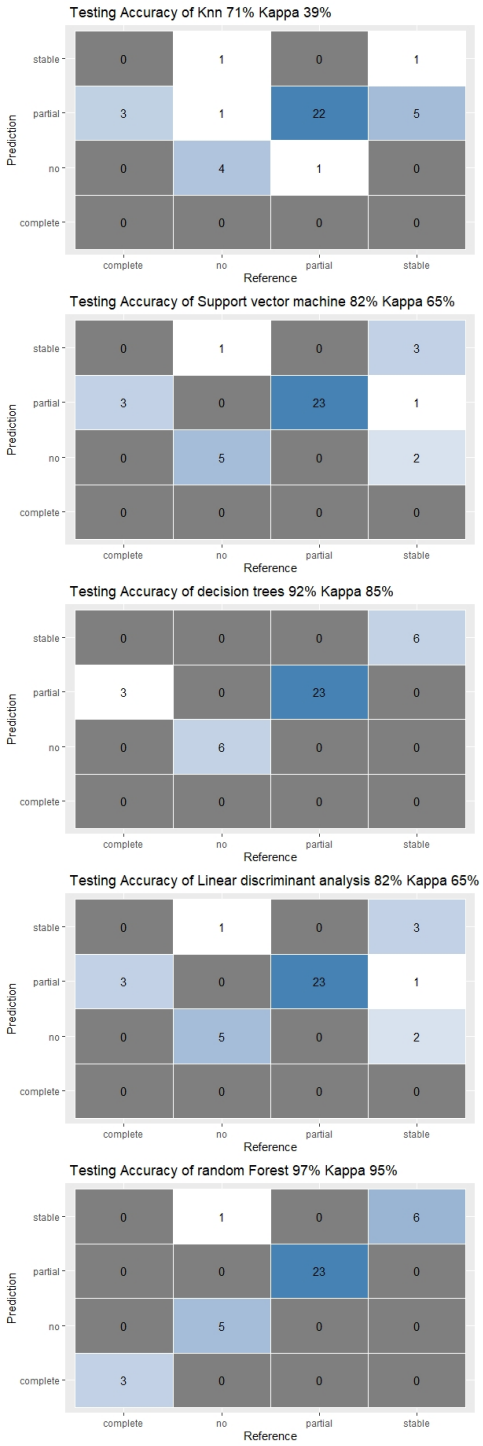

Fig. 16. Confusion matrix for testing dataset, labels as in Figure 17.

drug response. The pioneer work [24] used binding free energy, and personal features of 168 patients to predict a two-class drug response. Methods predicting fewer response levels performed better than earlier four-level predictors, however, our method outperforms all previous work. The combination of geometrical, energy, and personal features seems to be the optimal strategy for predicting the drug response. Figure 17, shows the contribution of each feature.

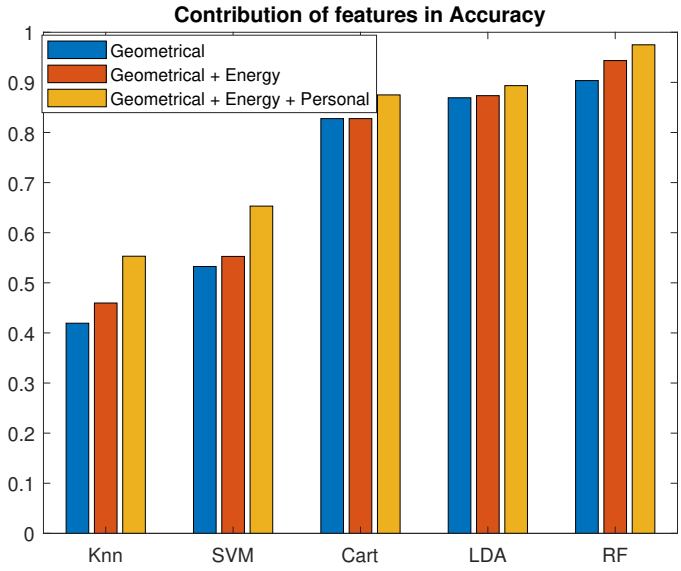

Fig. 17. Contribution of geometrical, personal, and energy features in the accuracy

B. Discussion

Rapid developments in the field of bioinformatics and the large amount of genomic data available today, allows development of personalized drug response models [58]. Previous studies showed that drug response is associated with clinical information, the type of EGFR mutation, binding free energy, and geometrical properties of the drug binding pocket. Different studies combined different properties to achieve higher accuracy (Table II).

In this work, three features, clinical information, protein-drug interactions, and geometrical properties of the drug-binding pocket, were combined to achieve state of the art performance with 97.5% accuracy, 100% recall, 95% precision, and 96.3% F1 - score with a random forest classifier. The classifier performed well on all the classes, with only one mis-classification between stable disease and progressive disease.

Our main focus in this work was on modeling the geometry of the drug-binding pocket and combining this with clinical information. The geometrical features, convex atoms at the interaction surface of the complex, the matching rates of surface atoms, the distances between the center of the drug molecule and binding-site residues, and hydrogen bonds, were the most discriminative features. Combining clinical and molecular predictors to identify drug-sensitive patients was most effective. Further investigation of this model may result in mutation, age or gender specific therapies and the model can also help in selecting the optimal drug for specific patients.

A limitation of this study was that it contained only most common 33 EGFR mutations from a possible 594 EGFR mutations available in COSMIC database [59]. However, the mutations used account for about 90% of all mutations. It is difficult to determine drug sensitivity to rare mutations due to limited patient data. Another limitation is the small dataset of 201 patients. Since, obtaining clinical data is difficult due

Reference	Number of patients	Features	Method	Response Level	Accuracy
[24]	167	Personal and energy	Extreme learning machines	2	95.13
[57]	355	Personal and genetic	Sequential minimization optimization	2	76.56
[26]	137	Geometrical and personal	Softmax regression	4	70.78
[25]	311	Energy and geometrical	Support vector machine	4	69.35
[27]	NA	Protein-drug interactions	Naive Bayes	3	95.50
Proposed method	201	Personal, energy, and geometrical	Random forest	4	97.50

TABLE II
COMPARISON WITH OTHER METHODS

to privacy and ethical considerations, most clinical studies consist of fewer than 400 patients, e.g. Table II, and may have imbalanced numbers of patients at each response level. Despite this, our model achieved a highly accurate prediction rate.

Personalized or precision medicine separates patients into different groups based on their individual medical decisions, interventions, risk of disease and drug-responses. Our model provides a personalized drug response model with a highly accurate prediction rate that could be tested on other types of cancer and other diseases.

VII. CONCLUSION

Computational methods, especially machine learning [60]–[62] are widely used to analyze, visualize and predict responses to lung cancer drugs. In this work, we developed a systematic model that uses personal, energy, and geometrical features in machine learning classifiers to predict the four levels of drug response. This method achieved state of the art performance at 97.5% accuracy with a random forest classifier, even though only a small patient dataset was available. This demonstrates the potential of the random forest method to deal with difficult learning situations and to be implemented in daily clinical practice to optimize treatment strategies for individual patients. In the future, more clinical data will be collected to further refine the prediction model, and test it on other diseases.

ACKNOWLEDGMENT

This work is supported by the Hong Kong Research Grants Council (Project 11200818) and City University of Hong Kong (Project 9610034)

The authors would like to thank David smith for his significant help in improving the paper.

REFERENCES

[1] L. A. Torre, R. L. Siegel, and A. Jemal, "Lung cancer statistics," in *Lung cancer and personalized medicine*, pp. 1–19, Springer, 2016.

[2] G. P. Gupta and J. Massagué, "Cancer metastasis: building a framework," *Cell*, vol. 127, no. 4, pp. 679–695, 2006.

[3] T. Oskarsson, E. Batlle, and J. Massagué, "Metastatic stem cells: sources, niches, and vital pathways," *Cell stem cell*, vol. 14, no. 3, pp. 306–321, 2014.

[4] T. Kawaguchi, M. Ando, K. Asami, Y. Okano, M. Fukuda, H. Nakagawa, H. Ibata, T. Kozuki, T. Endo, A. Tamura, *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)," *Journal of clinical oncology*, vol. 32, no. 18, pp. 1902–1908, 2014.

[5] W. Pao, V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria, B. Singh, R. Heelan, V. Rusch, L. Fulton, *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," *Proceedings of the National Academy of Sciences*, vol. 101, no. 36, pp. 13306–13311, 2004.

[6] S. Khozin, G. M. Blumenthal, X. Jiang, K. He, K. Boyd, A. Murgo, R. Justice, P. Keegan, and R. Pazdur, "Us food and drug administration approval summary: erlotinib for the first-line treatment of metastatic non-small cell lung cancer with epidermal growth factor receptor exon 19 deletions or exon 21 (l858r) substitution mutations," *The oncologist*, vol. 19, no. 7, p. 774, 2014.

[7] M. Günther, M. Juchum, G. Kelter, H. Fiebig, and S. Laufer, "Lung cancer: Egfr inhibitors with low nanomolar activity against a therapy-resistant l858r/t790m/c797s mutant," *Angewandte Chemie International Edition*, vol. 55, no. 36, pp. 10890–10894, 2016.

[8] R. Qureshi, M. Nawaz, A. Ghosh, and H. Yan, "Parametric models for understanding atomic trajectories in different domains of lung cancer causing protein," *IEEE Access*, vol. 7, pp. 67551–67563, 2019.

[9] S. Ikemura, H. Yasuda, S. Matsumoto, M. Kamada, J. Hamamoto, K. Masuzawa, K. Kobayashi, T. Manabe, D. Arai, I. Nakachi, *et al.*, "Molecular dynamics simulation-guided drug sensitivity prediction for lung cancer with rare egfr mutations," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 10025–10030, 2019.

[10] R. Qureshi, A. Ghosh, and H. Yan, "Correlated motions and dynamics in different domains of egfr with l858r and t790m mutations," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[11] S. Wan, R. Yan, Y. Jiang, Z. Li, J. Zhang, and X. Wu, "Insight into binding mechanisms of egfr allosteric inhibitors using molecular dynamics simulations and free energy calculations," *Journal of Biomolecular Structure and Dynamics*, vol. 37, no. 16, pp. 4384–4394, 2019.

[12] R. Qureshi, M. Zhu, A. Ghosh, and H. Yan, "Computational analysis of structural dynamics of egfr and its mutants," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2784–2791, IEEE, 2019.

[13] Q. Rizwan, M. Zhu, and H. Yan, "Visualization of protein-drug interactions for the analysis of drug resistance in lung cancer," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[14] F. S. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from large-scale biology," *Science*, vol. 300, no. 5617, pp. 286–290, 2003.

[15] C. C. Bennett, T. W. Doub, and R. Selove, "Ehrs connect research and practice: Where predictive modeling, artificial intelligence, and clinical decision support intersect," *Health Policy and Technology*, vol. 1, no. 2, pp. 105–114, 2012.

[16] P. De Meo, G. Quattrone, and D. Ursino, "Integration of the hl7 standard in a multiagent system to support personalized access to e-health services," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 8, pp. 1244–1260, 2010.

[17] M. Thafar, A. B. Raies, S. Albaradei, M. Essack, and V. B. Bajic, "Comparison study of computational prediction tools for drug-target binding affinities," *Frontiers in Chemistry*, vol. 7, 2019.

[18] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg, "Machine learning approaches to drug response prediction: challenges and recent progress," *NPJ precision oncology*, vol. 4, no. 1, pp. 1–10, 2020.

[19] J. Pittman, E. Huang, H. Dressman, C.-F. Horng, S. H. Cheng, M.-H. Tsou, C.-M. Chen, A. Bild, E. S. Iversen, A. T. Huang, *et al.*, "Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes," *Proceedings of the National Academy of Sciences*, vol. 101, no. 22, pp. 8431–8436, 2004.

[20] A. Potti, S. Mukherjee, R. Petersen, H. K. Dressman, A. Bild, J. Koontz, R. Kratzke, M. A. Watson, M. Kelley, G. S. Ginsburg, *et al.*, "A genomic strategy to refine prognosis in early-stage non–small-cell lung cancer," *New England Journal of Medicine*, vol. 355, no. 6, pp. 570–580, 2006.

[21] L. Li, "Survival prediction of diffuse large-b-cell lymphoma based on both clinical and gene expression information," *Bioinformatics*, vol. 22, no. 4, pp. 466–471, 2006.

[22] A. López-Encuentra, F. Lopez-Rios, E. Conde, R. Garcia-Lujan, A. Suarez-Gauthier, N. Manes, G. Renedo, J. Duque-Medina, E. Garcia-Lagarto, R. Rami-Porta, *et al.*, "Composite anatomical–clinical–molecular prognostic model in nonsmall cell lung cancer," *European Respiratory Journal*, vol. 37, no. 1, pp. 136–142, 2011.

[23] R. Wang, C. Chow, Y. Lyu, V. C. S. Lee, S. Kwong, Y. Li, and J. Zeng, "Taxirec: Recommending road clusters to taxi drivers using ranking-based extreme learning machines," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 3, pp. 585–598, 2018.

[24] D. D. Wang, W. Zhou, H. Yan, M. Wong, and V. Lee, "Personalized prediction of egfr mutation-induced drug resistance in lung cancer," *Scientific reports*, vol. 3, no. 1, pp. 1–8, 2013.

[25] L. Ma, D. D. Wang, B. Zou, and H. Yan, "An eigen-binding site based method for the analysis of anti-egfr drug resistance in lung cancer treatment," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 5, pp. 1187–1194, 2016.

[26] B. Duan, B. Zou, D. D. Wang, H. Yan, and L. Han, "Computational evaluation of egfr dynamic characteristics in mutation-induced drug resistance prediction," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2299–2304, IEEE, 2015.

[27] B. Zou, V. H. Lee, and H. Yan, "Prediction of sensitivity to gefitinib/erlotinib for egfr mutations in nsclc based on structural interaction fingerprints and multilinear principal component analysis," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–13, 2018.

[28] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker, "Protein structure prediction using rosetta," in *Methods in enzymology*, vol. 383, pp. 66–93, Elsevier, 2004.

[29] V. H. Lee, V. P. Tin, T.-s. Choy, K.-o. Lam, C.-w. Choi, L.-p. Chung, J. W. Tsang, P. P. Ho, D. K. Leung, E. S. Ma, *et al.*, "Association of exon 19 and 21 egfr mutation patterns with treatment outcome after first-line tyrosine kinase inhibitor in metastatic non–small-cell lung cancer," *Journal of Thoracic Oncology*, vol. 8, no. 9, pp. 1148–1155, 2013.

[30] L. Ma, D. D. Wang, Y. Huang, H. Yan, M. P. Wong, and V. H. Lee, "Egfr mutant structural database: computationally predicted 3d structures and the corresponding binding free energies with gefitinib and erlotinib," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–10, 2015.

[31] B. Zou, V. H. Lee, L. Chen, L. Ma, D. D. Wang, and H. Yan, "Deciphering mechanisms of acquired t790m mutation after egfr inhibitors for nsclc by computational simulations," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[32] R. Lencioni and J. M. Llovet, "Modified recist (mrecist) assessment for hepatocellular carcinoma," in *Seminars in liver disease*, vol. 30, pp. 052–060, © Thieme Medical Publishers, 2010.

[33] D. C. Rapaport and D. C. R. Rapaport, *The art of molecular dynamics simulation*. Cambridge university press, 2004.

[34] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[35] E. H. Kellogg, A. Leaver-Fay, and D. Baker, "Role of conformational sampling in computing mutation-induced changes in protein structure and stability," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 3, pp. 830–838, 2011.

[36] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Šali, "Comparative protein structure modeling of genes and genomes," *Annual review of biophysics and biomolecular structure*, vol. 29, no. 1, pp. 291–325, 2000.

[37] D. Eisenberg, R. Lüthy, and J. U. Bowie, "[20] verify3d: assessment of protein models with three-dimensional profiles," in *Methods in enzymology*, vol. 277, pp. 396–404, Elsevier, 1997.

[38] Z. Thakur, R. Dharra, V. Saini, A. Kumar, and P. K. Mehta, "Insights from the protein-protein interaction network analysis of mycobacterium tuberculosis toxin-antitoxin systems," *Bioinformation*, vol. 13, no. 11, p. 380, 2017.

[39] D. A. Case, T. Darden, T. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, *et al.*, "Amber 10," tech. rep., University of California, 2008.

[40] C.-Y. Zhou, F. Jiang, and Y.-D. Wu, "Residue-specific force field based on protein coil library. rsff2: modification of amber ff99sb," *The journal of physical chemistry B*, vol. 119, no. 3, pp. 1035–1047, 2015.

[41] V. Kräutler, W. F. Van Gunsteren, and P. H. Hünenberger, "A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations," *Journal of computational chemistry*, vol. 22, no. 5, pp. 501–508, 2001.

[42] H. M. Aktulga, J. C. Fogarty, S. A. Pandit, and A. Y. Grama, "Parallel reactive molecular dynamics: Numerical methods and algorithmic techniques," *Parallel Computing*, vol. 38, no. 4-5, pp. 245–259, 2012.

[43] A. W. Gotz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker, "Routine microsecond molecular dynamics simulations with amber on gpus. 1. generalized born," *Journal of chemical theory and computation*, vol. 8, no. 5, pp. 1542–1555, 2012.

[44] D. R. Roe and T. E. Cheatham III, "Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data," *Journal of chemical theory and computation*, vol. 9, no. 7, pp. 3084–3095, 2013.

[45] K. Sargsyan, C. Grauffel, and C. Lim, "How molecular size impacts rmsd applications in molecular dynamics simulations," *Journal of chemical theory and computation*, vol. 13, no. 4, pp. 1518–1524, 2017.

[46] H. Gohlke, C. Kiel, and D. A. Case, "Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the ras–raf and ras–ralgds complexes," *Journal of molecular biology*, vol. 330, no. 4, pp. 891–913, 2003.

[47] H. Öztürk, A. Özgür, and E. Ozkirimli, "Deepdta: deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.

[48] S. Genheden and U. Ryde, "The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities," *Expert opinion on drug discovery*, vol. 10, no. 5, pp. 449–461, 2015.

[49] M. R. Reddy and M. D. Erion, "Calculation of relative binding free energy differences for fructose 1, 6-bisphosphatase inhibitors using the thermodynamic cycle perturbation approach," *Journal of the American Chemical Society*, vol. 123, no. 26, pp. 6246–6252, 2001.

[50] M. Naderi, J. M. Lemoine, R. G. Govindaraj, O. Z. Kana, W. P. Feinstein, and M. Brylinski, "Binding site matching in rational drug design: Algorithms and applications," *Briefings in bioinformatics*, vol. 20, no. 6, pp. 2167–2184, 2019.

[51] J. A. Wilson, A. Bender, T. Kaya, and P. A. Clemons, "Alpha shapes applied to molecular shape characterization exhibit novel properties compared to established shape descriptors," *Journal of chemical information and modeling*, vol. 49, no. 10, pp. 2231–2241, 2009.

[52] A. Fabri, G.-J. Giezeman, L. Kettner, S. Schirra, and S. Schönherr, "On the design of cgal a computational geometry algorithms library," *Software: Practice and Experience*, vol. 30, no. 11, pp. 1167–1202, 2000.

[53] D.-T. Lee and B. J. Schachter, "Two algorithms for constructing a delaunay triangulation," *International Journal of Computer & Information Sciences*, vol. 9, no. 3, pp. 219–242, 1980.

[54] L. Ma, B. Zou, and H. Yan, "Identifying egfr mutation-induced drug resistance based on alpha shape model analysis of the dynamics," *Proteome science*, vol. 14, no. 1, p. 12, 2016.

[55] M. Kuhn *et al.*, "Building predictive models in r using the caret package," *Journal of statistical software*, vol. 28, no. 5, pp. 1–26, 2008.

[56] A. J. Viera, J. M. Garrett, *et al.*, "Understanding interobserver agreement: the kappa statistic," *Fam med*, vol. 37, no. 5, pp. 360–363, 2005.

[57] N. Kureshi, S. S. R. Abidi, and C. Blouin, "A predictive model for personalized therapeutic interventions in non-small cell lung cancer," *IEEE journal of biomedical and health informatics*, vol. 20, no. 1, pp. 424–431, 2014.

[58] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2013.

[59] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P. A. Futreal, M. R. Stratton, *et al.*, "The cosmic (catalogue of somatic mutations in cancer) database and website," *British journal of cancer*, vol. 91, no. 2, pp. 355–358, 2004.

[60] O. Frunza, D. Inkpen, and T. Tran, "A machine learning approach for identifying disease-treatment relations in short texts," *IEEE transactions on knowledge and data engineering*, vol. 23, no. 6, pp. 801–814, 2010.

[61] G.-F. Hao, G.-F. Yang, and C.-G. Zhan, "Structure-based methods for predicting target mutation-induced drug resistance and rational drug design to overcome the problem," *Drug discovery today*, vol. 17, no. 19-20, pp. 1121–1126, 2012.

[62] L.-H. Loo, L. F. Wu, and S. J. Altschuler, "Image-based multivariate profiling of drug responses from single cells," *Nature methods*, vol. 4, no. 5, pp. 445–453, 2007.