Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 5, No. 3, July 2023, pp: 168-176; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received May 27, 2023; revised June 20, 2023; accepted June 21, 2023; date of publication July 30, 2023
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeemi.v5i3.307

How to cite: Syarifah Aini, Wisnu Ananta Kusuma, Medria Kusuma Dewi Hardhienata, Mushthofa, "Network-Based Molecular Features Selection to Predict the Drug Synergy in Cancer Cells", vol. 5, no. 3, pp. 168–176, July 2023.

Network-Based Molecular Features Selection to Predict the Drug Synergy in Cancer Cells

Syarifah Aini, Wisnu Ananta Kusuma, Medria Kusuma Dewi Hardhienata, Mushthofa
Department of Computer Science, Faculty of Mathematic and Natural Science, IPB University, Bogor, West Java, Indonesia

Corresponding author: Mushthofa (mush@apps.ipb.ac.id)

ABSTRACT Identifying synergistic drug combinations in cancer treatment is challenging due to the complex molecular circuitry of cancer and the exponentially increasing number of drugs. Therefore, computational approaches for predicting drug synergy are crucial in guiding experimental efforts toward finding rational combination therapies. This research selects the molecular features of cancer cells with a diffusion network-based approach. Additionally, a model is developed using non-linear regression algorithms, namely Random Forest, Extremely Randomized Tree, and XGBoost, to predict the synergy score of drug combinations against the selected cancer cell features. The data used are 118 drug combination screening data and 85 cancer cell molecules provided by AstraZeneca-Sanger DREAM Challenge. The prediction results indicate that as the data size increases, the correlation value of the model improves, leading to better prediction accuracy. The influential feature analysis revealed that the three most influential mutation features in the AKT_1 and PIK3C drug combination model were ATP8B3, ERBB2, and RNF8. In the drug combination model BCL2_BCL2L1 and FGFR, the three most influential mutation features were BACH1, ODF2, and BFAR. In the MAP2K_1 and PIK3C drug combination model, TP53, IL12p40*, and SOX4 were the most influential features. All of these features have a connection between the mutation features and cell lines, aligning with the therapeutic targets of the three-drug combinations, which were the focus of this study.

INDEX TERMS cancer, drug combination, drug synergy, network diffusion kernel, non-linear regression

I. INTRODUCTION

Cancer is a disease caused by uncontrolled cell growth [1]. According to [1], several cancer treatments are available for patients, including surgery, chemotherapy, radiation therapy, drug therapy, and molecular biological therapy. However, most cancer patients experience resistance, rendering the drugs ineffective in killing cancer cells [2]. Giving a single drug to cancer patients is ineffective and often leads to resistance [3]. In contrast, drug combinations offer higher effectiveness by overcoming resistance. The effectiveness of drug combinations is divided into three categories: positive (synergistic), neutral, and negative (antagonist). Synergistic drug combinations occur when drug A can kill 40% of cancer cells, and drug B can kill 10% of cancer cells, resulting in a combined effectiveness of 80% against cancer cells. The drug combination is considered neutral if drug A combined with drug B does not increase in killing cancer cells. Drug combinations are categorized as antagonists if drugs A and B are combined, decreasing the killing of cancer cells.

Identifying synergistic drug combinations in cancer treatment is challenging due to the complex molecular circuits of cancer [4]. Moreover, the preparation of drug combinations requires pre-clinical stages to assess potential synergies between drug pairs. Considering the exponentially growing number of drug combinations, it is impractical to screen all these combinations through direct experimentation. Hence, computational approaches play a crucial role in predicting drug synergy and guiding experimental efforts toward finding effective combination therapies [5].

In this context, the computational approach uses mathematics, statistics, and computer science to study the mechanisms and behavior of complex systems through computer simulations. By using a computational approach, the search space for large datasets of drug combinations can be reduced. This approach enables the selection of optimal drug combinations for experimental testing based on predefined priority criteria.

Several approaches have been developed to model synergistic drug combinations using chemical, biological, and molecular data from cancer cells [6], although with limited translation modeling capabilities. The main challenge in developing such models lies in the availability of sufficiently large and diverse public data to train computational approaches [7]. As a step towards addressing this challenge, in 2015, the AstraZeneca Dream Challenge

released experimental data comprising 11.5 thousand drug combinations, measuring cell viability in 118 drugs and 85 cancer cell lines. This dataset also includes comprehensive monotherapy drug responses for each drug and cell line. In addition to limited public data, another challenge in developing computational models for drug combinations in cancer lies in the dependence on the feature selection of cancer cell molecular data [8]. Feature selection involves identifying and selecting data to address specific goals or problems. Several approaches can be used to select or extract molecular features of cancer cells, including network or graph-based methods [9].

Previous research focused on selecting molecular features of cancer cells using a network-based method, as conducted by [10]. They investigated relevant features or paths in breast and ovarian cancer data from the Cancer Genome Atlas. They used a network diffusion approach that calculates the similarity scores of neighboring nodes in the constructed network. The study demonstrated that the network diffusion approach performs well in identifying relevant features in cancer data.

The researchers employed machine learning techniques to examine the relevance of feature selection using the proposed approach [11]. They developed models to predict the effects of drug synergy based on selected biological information features from O'Neil's cancer data [4], utilizing linear and non-linear regression algorithms. The result showed that the non-linear regression algorithm better predicted drug synergy in cancer patients using genomic information or features.

This study proposes a method by which we can perform feature selection to identify relevant molecular features used to predict drug synergy in combinations of cancer drugs. The approach involves network diffusion of gene mutations in cancer cells and the analysis of gene interactions in cancer cells. Model prediction was conducted to assess the model's accuracy in predicting the synergy of drug combinations based on the selected gene mutations' features. The prediction model is developed using three non-linear algorithms: Extremely Randomized Trees (ERT), Random Forest (RF), and XGBoost. The data used includes drug and molecular cancer cell screening data provided by the AstraZeneca-Sanger DREAM Challenge. This research provides insights into influential molecular features of cancer cells in predicting drug synergy effects. Additionally, the study yields a predictive model with reasonable accuracy for specific drug combinations from the AstraZeneca DREAM Challenge's data.

This research contributes to identifying relevant molecular features in cancer cells that influence the synergy score of cancer drug combinations. In practical terms, the findings of this study could aid in the early identification of drugs for newly diagnosed cancer patients whose cancer cell structures resemble those analyzed in this research. Additionally, this research only focuses on the selection of molecular features in cancer cells. The features of drug compounds were not considered to produce a predictive model for the synergy score of cancer drug combinations.

II. MATERIAL AND METHODS

A. RESEARCH FLOW

The research was conducted through several stages: data collection, data pre-processing, selection of molecular features of cancer cells, modeling, analysis, and model evaluation using the Pearson correlation. During the data collection stage, mutation, leaderboard, training, and testing data were downloaded from the AstraZeneca website. Subsequently, interaction data of cancer cell genes were downloaded from the Atlas of Cancer Signaling Network website. In the next stage, the data is pre-processed by reducing, transforming, and integrating. The third stage involved the selection of the molecular features in cancer cells, starting with the construction of the network. Following the network construction, diffusion was performed on the network. Additionally, feature selection was conducted using univariate feature selection and LASSO techniques. The next step is to create a model using the Random Forest, ERT, and XGBoost algorithms. Finally, the model was evaluated and analyzed using Pearson's correlation. The stages can be seen in FIGURE 1.

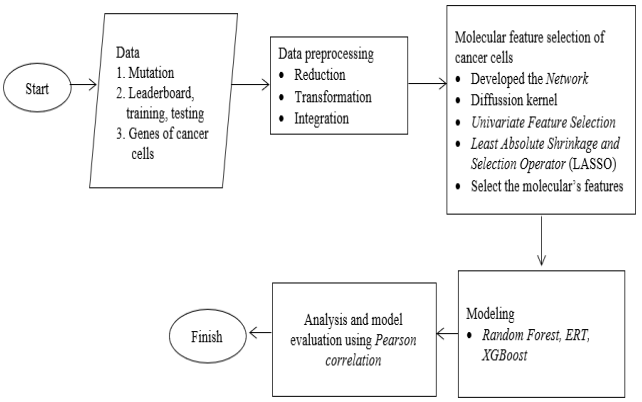


FIGURE 1. Research flow

B. DATA

The data was obtained from the AstraZeneca Dream Challenge and Atlas of Cancer Signaling Network (ACSN) websites. In this research, we integrated monotherapy data from the leaderboard, training, testing datasets, mutation data, and a gene-gene interaction network to predict drug synergy. A total of 11,500 synergy scores involving 118 drug combinations and 85 cell lines were experimentally assessed and provided by the AstraZeneca Dream Challenge [8]. However, this study focused only on three-drug combinations to examine the relevant features of each specific combination. The cancer cell lines used in this study originated from various sources, including breast (n=34), lung (n=22), urinary tract (n=14), gastrointestinal tract (n=12), male genital system (n=2), and lymphoma (n=1). AstraZeneca generated the mutation dataset using whole exome sequencing for all 85 cell lines, resulting in 20,521 gene mutations. The gene-gene interaction network was interconnected with cancer-related signaling and metabolic maps provided by ACSN [12]. TABLE 1 presents various data collected from these two websites.

TABLE 1
Variety of data

Data Name	Website	Attribute	Explanation
Leaderboard	AstraZeneca	14	Data containing cell name, *name of drug, combination of drugs, the value of the effectiveness of a single drug against cells, and the value of drug effectiveness when combined.
Training	AstraZeneca	14	Data that contains the same attributes as leaderboard data.
Testing	AstraZeneca	14	Data that contains the same attributes as leaderboard data.
Mutation	AstraZeneca	32	Data containing mutated genes in cancer cells.
Data between cancer cell genes.	ACSN	3	Data containing genes relevant to cancer.

*The drug name attribute does not contain the name of the drug but the name of the protein or inhibitor.

C. PRE-PROCESS DATA

Pre-process data consists of three processes: reduction, integration, and transformation. Data reduction involves removing unimportant data and selecting essential data. The data reduction was applied to the leaderboard, mutation, training, and testing data. Data integration involves combining multiple datasets into one. In this study, data integration was performed on the mutation data. Data transformation involves changing the format of the data. The data transformation process was applied to the cancer cell gene data.

The first pre-processing step was carried out on the leaderboard and training data. The attributes in the leaderboard and training data were reduced to only five: CELL_LINE, COMPOUND_A, COMPOUND_B, Einf_A, and Einf_B. Next, the data was transformed to generate drug-specific data, including CELL_LINE and Einf attributes. The Einf values in the new data were given a threshold to classify the drug response to cells, using the median Einf value as a threshold. The threshold value used was 40.51634. If the Einf value exceeds this threshold, the drug response is classified as synergistic (1); otherwise, it is classified as an antagonist (0). The pre-processing results on the leaderboard and training data yielded monotherapy data per drug, including the cell name, Einf value, and response value. Only data with more than ten cells were selected from the 69 monotherapy drug data. Ten drugs met this criterion and had a response value of 0 or 1: AKT, BCL2_BCL2L1, FGFR, MAP2K_1, MTOR_1, PIK3C, PIK3CB_PIK3CD, AKT_1, ATR_4, and IAP.

The second pre-processing step was applied to the mutation data. The 32 attributes in the mutation data were reduced to three attributes: cell_line_name, Gene.name, and FATHMM.prediction. Data cleaning was performed on the FATHMM.prediction attribute by removing data with the value "PASSENGER/OTHER". A value of "PASSENGER/OTHER" indicates that the somatic gene in the cancer cell does not contribute to cancer development. Next, a further reduction was made by selecting the cell_line name and Gene.name attributes. The transformation process was then applied to produce an adjacency matrix measuring 85 cells x 20,521 genes. The third pre-processing step involved transforming the data between cancer cell genes into an adjacency matrix measuring 2748 rows x 2748 columns.

The fourth preprocessing step was performed on the leaderboard, training, and testing data, following the same process as the first pre-processing step to obtain monotherapy data. However, the monotherapy data was based on drug combinations instead of single drugs for this fourth step. Only the drug combination data with the highest number of cancer cells in the leaderboard, training, and testing data were processed in the subsequent stages. From this pre-processing, three-drug combinations were selected: AKT_1.PIK3C, BCL2_BCL2L1.FGFR, and MAP2K_1.PIK3C.

Furthermore, integration was conducted between the pre-processed mutation data and the pre-processed result dataset per drug in the training, leaderboard, and testing data. This integration was performed for the selected monotherapy drugs based on the pre-processing results from the leaderboard, training, and testing data. The five drugs were AKT_1, BCL2_BCL2L1, FGFR, MAP2K_1, and PIK3C.

D. SELECTION OF MOLECULAR FEATURES OF CANCER CELLS

At this stage, feature selection is done on the cancer cell genes. First, a 0 and 1 adjacency matrix was built on cancer cell gene data and gene mutation data. The two adjacency matrices are then combined. A value of 0 means no mutation or the gene is normal, and a value of 1 indicates a mutation or abnormal gene. Then do the diffusion using the Laplacian Exponential Diffusion (LED) method with the formula:

$$K = \exp(-\alpha L), \quad (1)$$

$$L = D - A, \quad (2)$$

$$D(i,i) = \sum_{j=1}^{n} A(i,j), \quad (3)$$

A is the network adjacency matrix, D is A's diagonal degree association matrix, L is the Laplacian network matrix, and α is the independent parameter. This diffusion results in matrices with a value of 0, around which a value of 1 also has a value (no longer has a value of zero). Then the normalization process is carried out on the diffusion result dataset to change the numeric value of the diffusion result into the same scale. But not distort the differences in the range of values. Furthermore, feature selection is made with univariate feature selection to reduce the molecular features with numbers in the tens of thousands. Next, feature selection is performed using the LASSO algorithm to select the best features from the reduced features using univariate feature selection. FIGURE 2 is a visualization of the feature selection that was done.

After obtaining the gene features of the cells for each drug with LASSO, the datasets were combined according to

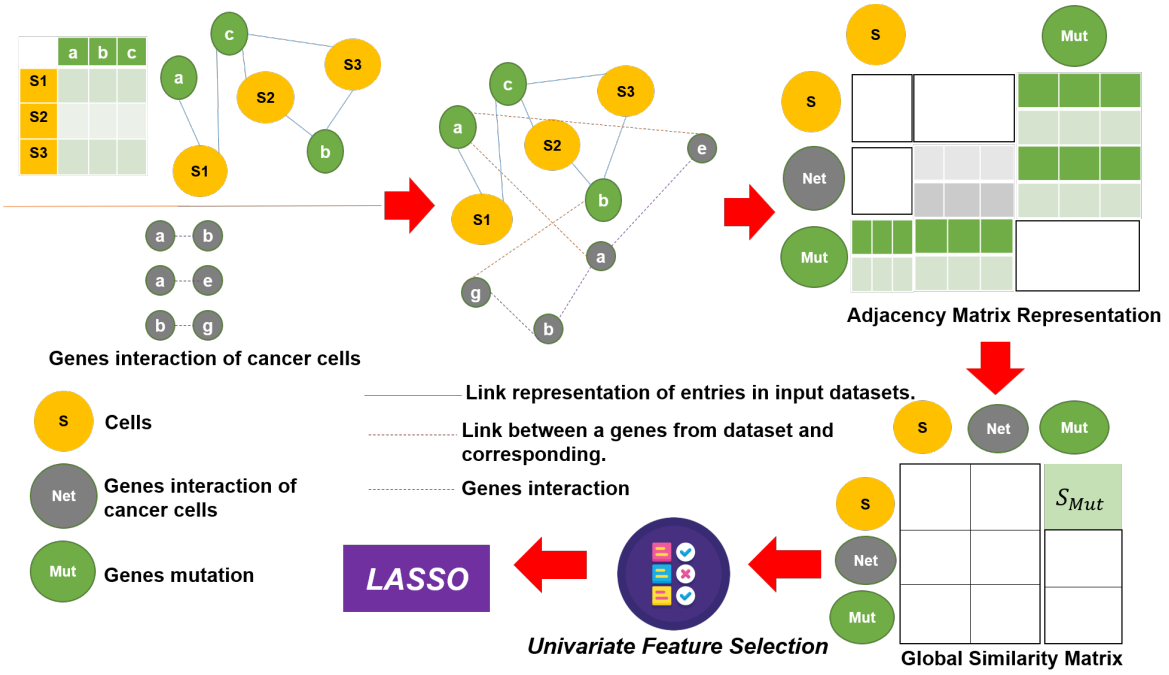


FIGURE 2. Selection of molecular features of cancer cells

the selected drug combinations. These combined datasets will become training data for the modeling process.

E. MODELING

Modeling is done using ERT, RF, and XGBoost algorithms. These algorithms are non-linear regression algorithms. The comparison of the three algorithms is that the RF algorithm used the bootstrap method to build a decision tree. The decision trees are built into all collected data in the ERT algorithm. While in the XGBoost algorithm, the decision trees are taken by combining several weak trees into a robust model that produces a strong prediction.

F. MODEL ANALYSIS AND EVALUATION

In this stage, an evaluation of the model is performed to see the ability of the model to predict synergy scores from the given drug combinations. Pearson correlation is used as a based accuracy prediction measure in every drug combination that is calculated with the formula below:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}, \quad (4)$$

where x is a predicted sample, y is an observed sample, $\bar{x}$ is the average of x, and $\bar{y}$ is the average of y. Four results can be obtained in the evaluation model using the Pearson correlation, as shown in TABLE 2.

TABLE 2
Pearson correlation evaluation matrix

Prediction results	Explanation
Positive	If x increases, y also increases. If <0.50, weak correlation; if >0.50, strong correlation.
Negative	If x increases, y decreases. If <-0.50 weak correlation, if >-0.50 strong correlation.
Zero	There is no connection between variables x and y.
NaN	One of the variables has a constant value, so the correlation coefficient is not defined.

III. RESULTS

A. SELECTION OF MOLECULAR FEATURES OF CANCER CELLS

Feature selection was conducted on the selected drug monotherapy from the pre-processed data. TABLE 3 presents each drug's selected gene mutation cell features obtained through lasso selection.

TABLE 3
Features selected from the selection using LASSO for each drug

Drugs	Selected Features of Gene Mutation Cells
AKT_1	RNF8; PDGFRB; MET; OCLN; ATP8B3; ERBB21P_ENST00000284037; FAM1358; TNXB.
BCL2_BCL2L1	IDAAM1; IGSF11; GNB2L1; PLEKHM2; NMNAT2_ENST00000294868.
FGFR	STAB1; BFAR; BACH1; ODF2; ALG13_ENST00000394780; FBXW7_ENST00000281708; c10orf68_ENST00000375025; Q8N0W1_HUMAN.
MAP2K_1	CAMKK2; THBS2; POLQ; ITGB2; NEURL; BAIAP2; OGG1.
PIK3C	TP53; SOX4; BCL3; ITGB1; IL1RAP; MIZ1*; NADE*; IL12p40*.

Based on the literature study, the average selected features from the five drugs exhibit a close relationship with cancer. However, there is one feature, Q8N0W1_HUMAN, on the FGFR drug, which has no direct association with cancer.

According to [13], it is one of the uncharacterized fragments in human GPCRs (G protein-coupled receptors). For instance, RNF8 is one of the selected features. As stated in [14], AKT regulation mediated by the RNF8 gene can activate AKT signaling in lung cancer.

B. MODELING

Prior to modeling, the pre-processed dataset and feature selection results are merged. The combined data serves as the training dataset for the three-drug combinations, incorporating the selected features and the monotherapy dataset. Following the data combination process, the features obtained consist of 22 features in the AKT_1 and PIK3C drug combination dataset, 19 features in the BCL2_BCL2L1 and FGFR drug combination dataset, and 21 features in the MAP2K_1 and PIK3C drug combination dataset. Subsequently, each dataset is split into training data (70%) and test data (30%). The training data is used to train the constructed model, while the test data is used to validate the model that has been constructed. Modeling was carried out on the three-drug combination datasets using three algorithms: Extremely Randomized Tree (ERT), Random Forest (RF), and XGBoost. The modeling is implemented using the sci-kit-learn library [15] in Python, utilizing the RandomForestRegressor module for RF, the XGBRegressor module for XGBoost, and the ExtraTreesRegressor module for ERT.

TABLE 4
Parameter configuration of the RF algorithm for each drug combination model

Parameter	Selected value in drug combination 1	Selected value in drug combination 2	Selected value in drug combination 3	Search space
n_estimators	400	600	1500	[200, 400, 600, 800, 1000, 1500, 2000]
max_depth	10	40	15	[5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]
max_features	log2	log2	sqrt	[auto, sqrt, log2]
criterion	mae	mae	mae	[mse, mae]
min_samples_split	3	6	9	[2, 3, 4, 5, 6, 7, 8, 9, 10]
min_impurity_decrease	0.1	0.5	0.1	[0.0, 0.05, 0.1, 0.5]
bootstrap	false	true	true	[true, false]

The three algorithms are then compared to determine which performs best. The first modeling round was performed on the AKT_1 and PIK3C drug combination datasets. The second round was conducted on the BCL2_BCL2L1 and FGFR drug combination datasets. The third modeling was done on the MAP2K_1 and PIK3C drug combination datasets. Parameter tuning is accomplished using the GridSearchCV library in Python by providing a range of values for each parameter to determine the optimal parameter for each model. Subsequently, each model is trained using the hyperparameters obtained from the parameter search with GridSearchCV. The tuning parameter values for each algorithm, derived from the GridSearchCV results for the three-drug combination dataset models, can be seen in TABLE 4, TABLE 5, and TABLE 6.

TABLE 5
Parameter configuration of the ERT algorithm for each drug combination model

Parameter	Selected value in drug combination 1	Selected value in drug combination 2	Selected value in drug combination 3	Search space
n_estimators	800	200	200	[200, 400, 600, 800, 1000, 1500, 2000]
max_depth	45	20	10	[5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55]
max_features	auto	sqrt	log2	[auto, sqrt, log2]
criterion	mse	mae	mae	[mse, mae]
min_samples_split	4	10	6	[2, 3, 4, 5, 6, 7, 8, 9, 10]
min_impurity_decrease	0.05	0.5	0.5	[0.0, 0.05, 0.1, 0.5]
bootstrap	true	true	false	[true, false]

TABLE 6
Parameter configuration of the XGBoost algorithm for each drug combination model

Parameter	Selected value in drug combination 1	Selected value in drug combination 2	Selected value in drug combination 3	Search space
n_estimators	1800	1800	800	[200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]
max_depth	18	18	14	[2, 4, 6, 8, 10, 12, 14, 16, 18, 20]
min_child_weight	10	10	4	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
xgb_tree_method	gpu_hist	gpu_hist	auto	[auto, exact, approx, hist, gpu_hist]
xgb_eta	0.1	0.1	0.30000000000000004	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6]
xgb_gamma	0	0	0	[0, 0.1, 0.2, 0.3, 0.4, 0.5]
xgb_objective	reg:squarederror	reg:squarederror	reg:squarederror	[reg:squarederror, reg:squaredlogerror]

C. MODEL ANALYSIS AND EVALUATION

After modeling, the model is analyzed and evaluated using the Pearson correlation method. According to [16], Pearson correlation is suitable for analyzing normally distributed data. In the AstraZeneca dream challenge dataset, the synergy score attributes for each data provided by the

challenge are normally distributed. Because of that, the Pearson correlation is used as a suitable measure for predicting accuracy in each drug combination [8]. This correlation method compares the correlation between the predicted synergy score and the observed synergy score. The Pearson module in the SciPy Python library is utilized to calculate the Pearson correlation value in this study. FIGURE 3 illustrates that the synergy score attribute in this study follows a typical distribution. TABLE 7 presents the correlation values obtained from the three models built using the three algorithms.

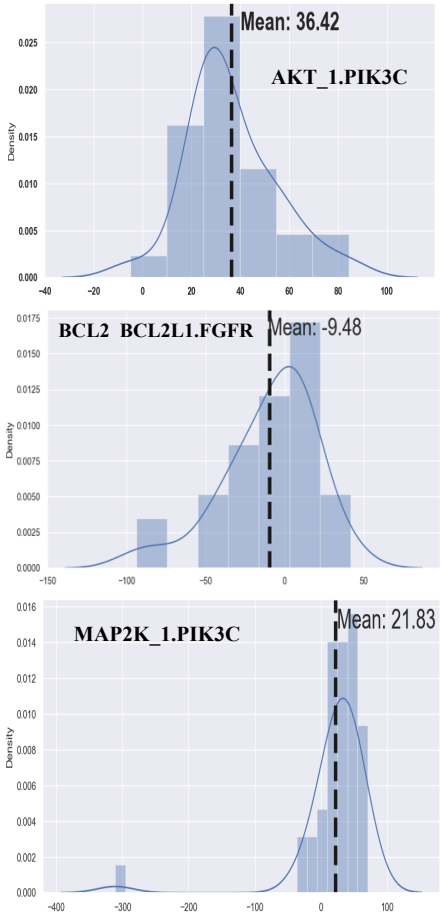


FIGURE 3. **The normal distribution of attribute synergy scores on the three datasets**

TABLE 7
Pearson correlation values with three models with three algorithms

Drug Combination	Number of cells	RF	ERT	XGBoost
AKT_1 dan PIK3C	29	**0,045**	-0,411	NaN
BCL2_BCL2L1 dan FGFR	30	**0,570**	0,130	NaN
MAP2K_1 dan PIK3C	42	0,830	0,662	**0,932**

TABLE 7 demonstrates that the RF algorithm has the highest correlation value compared to the ERT and XGBoost algorithms for the model on the AKT_1 and PIK3C drug combination dataset with 29 cells, as well as the BCL2_BCL2L1 and FGFR drug combination dataset with 30 cells. Conversely, for the model on the combined dataset of MAP2K_1 and PIK3C drugs with a cell count of 42, the XGBoost algorithm achieves the best accuracy, with a Pearson correlation value of 0.932, which is 0.1 higher than the Pearson correlation value obtained with the RF algorithm.

Furthermore, in the XGBoost algorithm, the Pearson correlation value for the AKT_1 and PIK3C drug combination dataset and the BCL2_BCL2L1 and FGFR drug combination dataset is NaN (undefined) due to the prediction of synergy scores on the AKT_1 and PIK3C drug combination test data and the BCL2_BCL2L1 and FGFR drug combination test data using the XGBoost algorithm resulting in a constant value. In the AKT_1 and PIK3C drug combination test data, each data row is predicted to have a value of 36.974. In the test data for the BCL2_BCL2L1 and FGFR drug combination, each data row is predicted to have a value of -13.192. In the Pearson correlation formula (Equation 4), there is a division by zero when subtracting the average of the predicted value (y) from each value. This division by zero results in NaN values for the Pearson correlation in the first and second drug combinations models using the XGBoost algorithm.

Additionally, the ERT algorithm produces a negative Pearson correlation value in the AKT_1 and PIK3C drug combination model due to an inverse relationship between the actual values (x) and the predicted results (y), resulting in a negative correlation coefficient.

IV. DISCUSSION

A. PREDICTION MODEL ANALYSIS

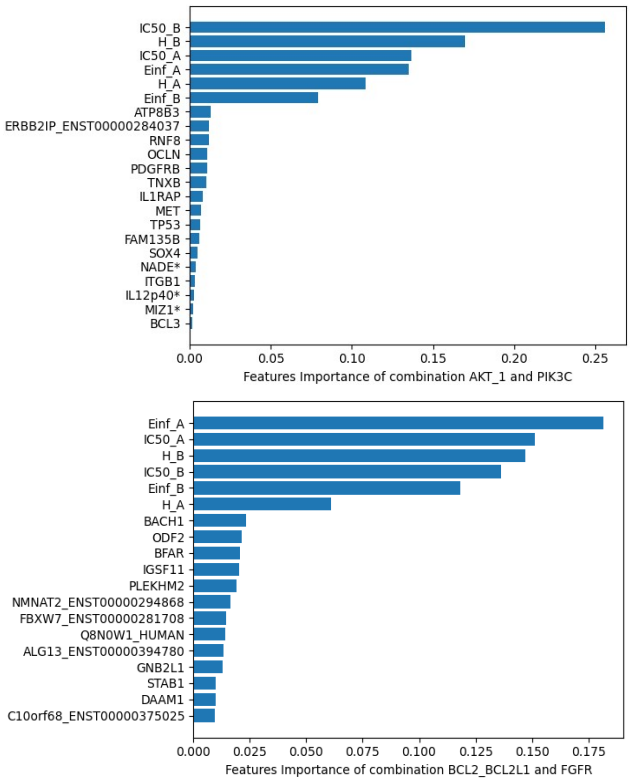
The results of the evaluation model on AKT_1 and PIK3C aslo BCL2_BCL2L1 and FGFR align with the research conducted by [11], which predicted the synergy score of drug combinations by comparing the accuracy of models built with RF and ERT. When the entire training dataset of 16,575 samples was used, ERT showed a higher correlation value than RF, with values of 0.738 and 0.731, respectively. However, when only 780 data samples were used, RF had a better correlation value than ERT, with values of 0.827 and 0.821, respectively.

A study by [17] on 28 datasets from the UCI repository [18] compared the accuracy of the random forest, XGboost, and gradient boosting algorithms. The results showed that XGboost has good accuracy on data with fewer features or columns compared to the number of instances or rows. This aligns with the results obtained from the correlation value of the MAP2K_1 and PIK3C drug combination model, where the number of cells is exceeded the number of features, resulting in a better correlation value than RF and ERT.
Additionally, TABLE 7 indicates that as the data size increases, the correlation value in the model also increases, indicating improved prediction accuracy. This result is consistent with the finding of [19], which states that larger data sizes lead to enhanced model performance. According

to [19], this improvement suggests that the built model can be generalized to include more cells, thereby improving predictive performance. Furthermore, in the AstraZeneca dream challenge, one of the teams who achieved the best performance in predicting the synergy of drug combinations was the DMIS team (https://www.synapse.org/#!Synapse:syn5816530). For the drug combination BCL2_BCL2L1 and FGFR, the DMIS team obtained a Pearson correlation value of 0.162. This value is smaller when compared to the value obtained in this study, measuring 0.570.

B. INFLUENTIAL FEATURES ANALYSIS WITH RANDOM FOREST ALGORITHM

The important features of each drug combination can be obtained by using the feature_importance function in the sklearn module in Python. FIGURE 4 illustrates the sequence of important features for each drug combination model. As shown in FIGURE 4, the three most influential features in the AKT_1 and PIK3C drug combination model are ATP8B3, ERBB2, and RNF8. According to the proteinatlas.org website, ATP8B3, ERBB2, and RNF8 have gene expression in all cell lines of the training and test data, except for the M-14 and EVSA-T cell lines, where all cell lines exhibit expression of these three influential features that are relevant to breast cancer. Research by [20] found that direct inhibition of AKT_1 may represent a therapeutic strategy for breast cancer. In addition, [21] found that PIK3C inhibition could be a therapeutic strategy for breast cancer after adjunctive therapy.



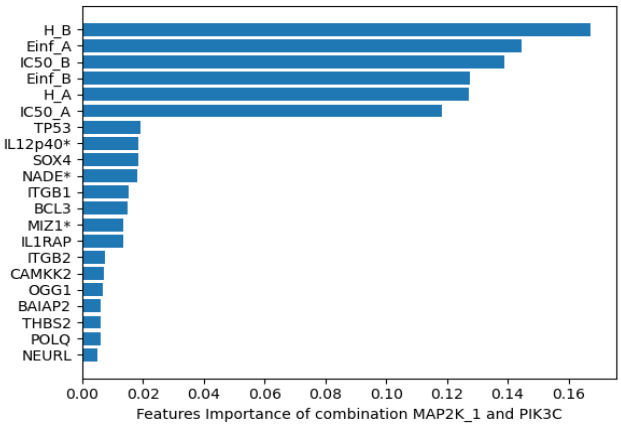


FIGURE 4. **Important features of the three drug combination models**

In the BCL2_BCL2L1 and FGFR drug combination model, the three most influential features of cell mutations are BACH1, ODF2, and BFAR. According to the proteinatlas.org website, these three genes are expressed in all training and test cell lines, except for the MFM-223 and M14 cell lines in the test data. All cell lines expressing these three influential features are relevant to breast, bladder, and lung cancer. Research by [22] indicates that the BCL2 family is an important clinical prognostic marker for breast cancer. Research by [23] demonstrated that targeting BCL2L1 could be a plausible therapeutic strategy in bladder cancer patients. [24] targeted the BCL2 family as a therapeutic strategy in lung cancer patients. Research by [25] showed that FGFR can be a therapeutic target in breast cancer. Research by [26] stated that FGFR could pose challenges in the clinical practice of bladder cancer treatment. Research by [27] highlighted that FGFR could inhibit personalized treatment in lung cancer patients.

The three most influential cell mutation features in the MAP2K_1 and PIK3C drug combination model are TP53, IL12p40*, and SOX4. According to the proteinatlas.org website, these three genes are expressed in the training and test data cell lines, except for the C32 and COLO-205 cell lines in the training data, and RKO and MFM-223 in the test data. The cell lines expressing these three influential features are relevant to breast, bladder, and colon cancer. Research by [28] demonstrated that using MAP2K_1 inhibitors combined with radiation therapy significantly reduced cell migration capacity in breast cancer cells. Research by [29] hypothesized that the knockdown of MAP2K_1 via miRNA-1826 could be a new therapeutic approach for bladder cancer. Research by [30] targeted MAP2K_1 as an inhibitor for the treatment of colon cancer. Research by [31] found that PIK3C could be a therapeutic target for bladder cancer cells. Research by [32] provided clinical trial data of PIK3C as a target for drug therapy in various cancers, including breast, colon, and bladder. Based on the above analysis, it can be concluded that there is a relationship between the influential cell mutation features and the cell lines with the therapeutic targets of the three-drug combinations, which are the focus of this study.

Furthermore, the results of this research imply that computational approaches for predicting drug synergy are crucial in guiding experimental efforts toward finding rational combination therapies in cancer treatment. This study demonstrates the potential of a network-based molecular features selection approach to predict drug synergy in cancer cells. Furthermore, selecting influential mutation features in specific drug combination models provides insights into the therapeutic targets and potential mechanisms underlying the observed synergistic effects. This information can inform further investigations and drug development strategies. The comparison of three regression algorithms (Random Forest, Extremely Randomized Tree, and XGBoost) highlights the importance of algorithm selection in predicting drug synergy. The Random Forest algorithm generally performs well in the drug combination datasets analyzed in this study, but the XGBoost algorithm shows superior accuracy in some instances. The results of this study align with previous research findings, validating the effectiveness of the selected algorithms and highlighting the potential of the network-based molecular features selection approach in predicting drug synergy. These findings contribute to the growing knowledge of computational drug synergy prediction. Finally, identifying influential features in each drug combination model offers potential biomarkers or targets for further investigation. These features, which show a connection between mutation features and cell lines, align with the therapeutic targets of the drug combinations studied. Further studies can explore the functional roles of these features and their potential in personalized medicine approaches. Additionally, the weakness of this study lies in its limited focus on cancer cell interaction genes, which resulted in a restricted search space limited to gene interactions within cancer cells.

IV. CONCLUSION

This study successfully developed a network diffusion-based approach for selecting molecular features of cancer cells. The data used includes drug and molecular cancer cell screening data provided by the AstraZeneca-Sanger DREAM Challenge. A prediction model for the synergy score of drug combinations against cancer cells was also successfully developed. The development of the model is based on the selected mutational features of cancer cells using diffusion network. In the drug combination models of AKT_1 and PIK3C with the RF, ERT, and XGBoost algorithms, the Pearson correlation values were found to be 0.045, -0.411, and NaN, respectively. In the BCL2_BCL2L1 and FGFR drug combination model, the Pearson correlation results were RF (0.570), ERT (0.130), and XGBoost (NaN). In the MAP2K_1 and PIK3C drug combination model, Pearson correlation values were RF (0.830), ERT (0.662), and XGBoost (0.932). These results indicate that as the data size increases, the correlation value of the model improves, leading to better prediction accuracy. This improvement suggests that the developed model can be generalized to include more cells and enhance prediction performance. The feature selection using network diffusion, univariate feature selection, and LASSO yielded relevant gene mutation features related to cancer. The RF algorithm's influential feature analysis revealed that the three most influential mutation features in the AKT_1 and PIK3C drug combination model were ATP8B3, ERBB2, and RNF8. In the drug combination model BCL2_BCL2L1 and FGFR, the three most influential mutation features were BACH1, ODF2, and BFAR. In the MAP2K_1 and PIK3C drug combination model, TP53, IL12p40*, and SOX4 were the most influential features. All of these features have a connection between the mutation features and cell lines, aligning with the therapeutic targets of the three-drug combinations, which were the focus of this study. Future research can expand the dataset to improve the predictive value of drug combination synergy scores. Additionally, other types of data, such as copy number variant (CNV) data, methylation, and gene expression, can be incorporated as features for selection. Various feature selection techniques and model training approaches can be explored to enhance algorithm performance.

ACKNOWLEDGMENT

We thank Biofarmaka IPB University for providing the necessary resources, including Google Collab Pro and server access, for data processing in this research.

REFERENCES

[1] D. T. Debela *et al.*, "New approaches and procedures for cancer treatment: Current perspectives," *SAGE Open Medicine*, vol. 9. SAGE Publications Ltd, 2021. doi: 10.1177/20503121211034366.
[2] G. Housman *et al.*, "Drug resistance in cancer: An overview," *Cancers (Basel)*, vol. 6, no. 3, pp. 1769–1792, 2014, doi: 10.3390/cancers6031769.
[3] R. S. Narayan *et al.*, "A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities," *Nat Commun*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-16735-2.
[4] J. O'Neil *et al.*, "An unbiased oncology compound screen to identify novel combination strategies," *Mol Cancer Ther*, vol. 15, no. 6, pp. 1155–1162, 2016, doi: 10.1158/1535-7163.MCT-15-0843.
[5] B. Al-Lazikani, U. Banerji, and P. Workman, "Combinatorial drug therapy for cancer in the post-genomic era," *Nat Biotechnol*, vol. 30, no. 7, pp. 679–692, 2012, doi: 10.1038/nbt.2284.
[6] K. C. Bulusu *et al.*, "Modelling of compound combination effects and applications to efficacy and toxicity: State-of-the-art, challenges and perspectives," *Drug Discov Today*, vol. 21, no. 2, pp. 225–238, 2016, doi: 10.1016/j.drudis.2015.09.003.
[7] H. Gao *et al.*, "High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response," *Nat Med*, vol. 21, no. 11, pp. 1318–1325, 2015, doi: 10.1038/nm.3954.
[8] M. P. Menden *et al.*, "Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen," *Nat Commun*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-09799-2.
[9] M. Netzer *et al.*, "A network-based feature selection approach to identify metabolic signatures in disease," *J Theor Biol*, vol. 310, pp. 216–222, 2012, doi: 10.1016/j.jtbi.2012.06.003.
[10] L. P. C. Verbeke, J. Van Den Eynden, A. C. Fierro, P. Demeester, J. Fostier, and K. Marchal, "Pathway relevance ranking for tumor samples through network-based data integration," *PLoS One*, vol. 10, no. 7, pp. 1–22, 2015, doi: 10.1371/journal.pone.0133503.
[11] M. Jeon, S. Kim, S. Park, H. Lee, and J. Kang, "In silico drug combination discovery for personalized cancer therapy," *BMC Syst Biol*, vol. 12, no. Suppl 2, 2018, doi: 10.1186/s12918-018-0546-1.
[12] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011, doi: 10.1016/j.cell.2011.02.013.
[13] J. C. Kim *et al.*, "Enhancement of colorectal tumor targeting using a novel biparatopic monoclonal antibody against carcinoembryonic antigen in experimental radioimmunoguided surgery," *Int J Cancer*, vol. 97, no. 4, pp. 542–547, Feb. 2002, doi: 10.1002/ijc.1630.

[14] Y. Xu *et al.*, "RNF8-mediated regulation of Akt promotes lung cancer cell survival and resistance to DNA damage," *Cell Rep*, vol. 37, no. 3, Oct. 2021, doi: 10.1016/j.celrep.2021.109854.
[15] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," 2011. [Online]. Available: http://scikit-learn.sourceforge.net.
[16] P. Schober and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesth Analg*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.
[17] C. Bentéjac, C. Anna, and O B Gonzalo Martínez-Muñoz, "A Comparative Analysis of XGBoost," 2019. [Online]. Available: https://www.researchgate.net/publication/337048557
[18] M. Lichman, *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*. 2013.
[19] H. Li, T. Li, D. Quang, and Y. Guan, "Network propagation predicts drug synergy in cancers," *Cancer Res*, vol. 78, no. 18, pp. 5446–5457, Sep. 2018, doi: 10.1158/0008-5472.CAN-18-0740.
[20] Q. B. She *et al.*, "Breast tumor cells with P13K mutation or HER2 amplification are selectively addicted to Akt signaling," *PLoS One*, vol. 3, no. 8, Aug. 2008, doi: 10.1371/journal.pone.0003065.
[21] Y. A. Cho *et al.*, "PIK3CA Mutation as Potential Poor Prognostic Marker in Asian Female Breast Cancer Patients Who Received Adjuvant Chemotherapy," *Current Oncology*, vol. 29, no. 5, pp. 2895–2908, May 2022, doi: 10.3390/curroncol29050236.
[22] G. J. Lindeman and J. E. Visvader, "Targeting BCL-2 in breast cancer: exploiting a tumor lifeline to deliver a mortal blow?," *Breast Cancer Manag*, vol. 2, no. 1, pp. 1–4, Jan. 2013, doi: 10.2217/bmt.12.60.
[23] S. Yoshimine *et al.*, "Prognostic significance of Bcl-xL expression and efficacy of Bcl-xL targeting therapy in urothelial carcinoma," *Br J Cancer*, vol. 108, no. 11, pp. 2312–2320, Jun. 2013, doi: 10.1038/bjc.2013.216.
[24] L. Gandhi *et al.*, "Phase I study of navitoclax (ABT-263), a novel bcl-2 family inhibitor, in patients with small-cell lung cancer and other solid tumors," *Journal of Clinical Oncology*, vol. 29, no. 7, pp. 909–916, Mar. 2011, doi: 10.1200/JCO.2010.31.6208.
[25] N. J. Chew *et al.*, "Evaluation of FGFR targeting in breast cancer through interrogation of patient-derived models," *Breast Cancer Research*, vol. 23, no. 1, Dec. 2021, doi: 10.1186/s13058-021-01461-4.
[26] S. De Keukeleire, D. De Maeseneer, C. Jacobs, and S. Rottey, "Targeting FGFR in bladder cancer: ready for clinical practice?," *Acta Clinica Belgica: International Journal of Clinical and Laboratory Medicine*, vol. 75, no. 1. Taylor and Francis Ltd., pp. 49–56, Jan. 02, 2020. doi: 10.1080/17843286.2019.1685738.
[27] L. Pacini, A. D. Jenks, N. C. Lima, and P. H. Huang, "Targeting the fibroblast growth factor receptor (Fgfr) family in lung cancer," *Cells*, vol. 10, no. 5, 2021, doi: 10.3390/cells10051154.
[28] N. Anastasov *et al.*, "Mek1 inhibitor combined with irradiation reduces migration of breast cancer cells including mir-221 and zeb1 emt marker expression," *Cancers (Basel)*, vol. 12, no. 12, pp. 1–20, Dec. 2020, doi: 10.3390/cancers12123760.
[29] H. Hirata, Y. Hinoda, K. Ueno, V. Shahryari, L. Tabatabai, and R. Dahiya, "MicroRNA-1826 targets VEGFC, beta-catenin (CTNNB1) and MEK1 (MAP2K1) in human bladder cancer," *Carcinogenesis*, vol. 33, no. 1, pp. 41–48, Jan. 2012, doi: 10.1093/carcin/bgr239.
[30] D. Kobelt *et al.*, "The newly identified MEK1 tyrosine phosphorylation target MACC1 is druggable by approved MEK1 inhibitors to restrict colorectal cancer metastasis," *Oncogene*, vol. 40, no. 34, pp. 5286–5301, Aug. 2021, doi: 10.1038/s41388-021-01917-z.
[31] J. Zhang, T. M. Roberts, and R. A. Shivdasani, "Targeting PI3K signaling as a therapeutic approach for colorectal cancer," *Gastroenterology*, vol. 141, no. 1. W.B. Saunders, pp. 50–61, 2011. doi: 10.1053/j.gastro.2011.05.010.
[32] J. Yang, J. Nie, X. Ma, Y. Wei, Y. Peng, and X. Wei, "Targeting PI3K in cancer: Mechanisms and advances in clinical trials," *Molecular Cancer*, vol. 18, no. 1. BioMed Central Ltd., Feb. 19, 2019. doi: 10.1186/s12943-019-0954-x.

SYARIFAH AINI received her Bachelor's Degree in Informatics from Sriwijaya University in 2016. Now, she is a master's student in the Department of Computer Science at IPB University. This paper is a part of her master's research. She can be contacted at email: ainisyarifah@apps.ipb.ac.id.

WISNU ANANTA KUSUMA received his master's degree in Informatics from Bandung Institute of Technology in 1999. He received his Ph.D. in computer science from the Tokyo Institute of Technology in 2012. Now, he is an Associate Professor at the Department of Computer Science, IPB University, and an Executive Secretary at Tropical Biopharmaca Research Center, IPB University. He is also the coordinator of the Bioinformatics Working Group, Faculty of Mathematics and Natural Science, IPB University, and Coordinator of Bioinformatics and High-Performance Computing Research Group, Advanced Laboratory, IPB University. His current research interest is machine learning, high-performance computing, and bioinformatics. Currently, he is conducting collaborative research with the Indonesian Medical Education and Research Institute, University of Indonesia, on a drug repurposing project for COVID-19. The task is to build a prediction model based on machine learning and network pharmacology for screening herbal compounds. He can be contacted at email: ananta@apps.ipb.ac.id.

MEDRIA KUSUMA DEWI HARDHIENATA is a lecturer at the Department of Computer Science, IPB University, Indonesia. She completed her Bachelor's Degree in Computer Science at the Department of Computer Science at IPB University. In 2015, she completed her Ph.D. in Computer Science at the University of New South Wales (UNSW), Canberra, Australia. Her Ph.D. thesis investigates the role of otivation in solving optimization problems with application to task allocation. In 2015/2016, she was a Research Associate at UNSW Canberra and conducted a research for developing, implementing, and testing an agent-based simulation in strategic decision-making and information warfare games. Her research interests include multi-agent systems, computational intelligence, optimization and agent-based simulation. She can be contacted at email: medria.hardhienata@apps.ipb.ac.id.

MUSHTHOFA is a lecturer at the Department of Computer Science, IPB University, Indonesia. In 2009, he completed his Master's in Computational Logic at the Vienna University of Technology, Austria. In 2018, he received his Ph.D. in Computer Science and Bioinformatics at Ghent University, Belgium. His research interests include artificial intelligence, machine learning, data science, knowledge representation and reasoning, and bioinformatics. He can be contacted at email: mush@apps.ipb.ac.id.