Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.038578
Article

Developing a Breast Cancer Resistance Protein Substrate Prediction System Using Deep Features and LDA

Mehdi Hassan[1,2], Safdar Ali[3], Jin Young Kim[2,*], Muhammad Sanaullah[4], Hani Alquhayz[5] and Khushbakht Safdar[6]

[1]Department of Computer Science, Air University, Sector E-9, PAF Complex, Islamabad, 44000, Pakistan
[2]Department of ICT Convergence System Engineering, Chonnam National University, Gwangju, Korea
[3]Directorate of National Repository, Islamabad, Pakistan
[4] Department of Computer Science, Bahauddin Zakariya University, Multan, 60000, Pakistan
[5]Department of Computer Science and Information, College of Science in Zulfi, Majmaah University, Al- Majmaah, 11952, Saudi Arabia
[6]Pakistan Atomic Energy Commission, General Hospital, Islamabad, Pakistan
*Corresponding Author: Jin Young Kim. Email: beyondi@chonnam.ac.kr
Received: 19 December 2022; Accepted: 10 April 2023; Published: 30 August 2023

Abstract: Breast cancer resistance protein (BCRP) is an important resistance protein that significantly impacts anticancer drug discovery, treatment, and rehabilitation. Early identification of BCRP substrates is quite a challenging task. This study aims to predict early substrate structure, which can help to optimize anticancer drug development and clinical diagnosis. For this study, a novel intelligent approach-based methodology is developed by modifying the ResNet101 model using transfer learning (TL) for automatic deep feature (DF) extraction followed by classification with linear discriminant analysis algorithm (TLRNDF-LDA). This study utilized structural fingerprints, which are exploited by DF contrary to conventional molecular descriptors. The proposed in silico model achieved an outstanding accuracy performance of 98.56% on test data compared to other state-of-the-art approaches using standard quality measures. Furthermore, the model's efficacy is validated via a statistical analysis ANOVA test. It is demonstrated that the developed model can be used effectively for early prediction of the substrate structure. The pipeline of this study is flexible and can be extended for in vitro assessment efficacy of anticancer drug response, identification of BCRP functions in transport experiments, and prediction of prostate or lung cancer cell lines.

Keywords: BCRP; drug response; deep learning; transfer learning; LDA; *In silico*

1 Introduction

Human breast cancer resistance protein (BCRP) was discovered in 1998 by cloning the human breast cancer cell line [1]. Specifically, BCRP is a part of subfamily G of the large human ABC

(adenosine triphosphate (ATP) binding cassette) transporter superfamily with gene symbol ABCG2 (2nd member of subfamily G). It is an ABC efflux drug transporter and imposes resistance to several chemotherapeutic agents which differ in structure and chemical properties. In addition, it can be distributed and expressed in normal tissues of the small intestine, brain endothelium, liver, and placenta. One of the essential functions of BCRP is its usage as a xenobiotic efflux pump to export antineoplastic drugs from cells [2]. The ABC efflux transporter protein is vital for discovering multidrug resistance, which is used to treat different cancers. It is beneficial for the absorption, distribution, and elimination of drugs. In the last two decades, significant knowledge concerning BCRPs has been obtained regarding expressions, functions, and molecular structures to target the development of anticancer drugs.

Recently, the substrates of BCRP have been rapidly growing to contain chemotherapeutics, non-chemotherapeutics, and physiological drugs. Besides, they contain many compounds ranging from highly lipophilic anticancer to hydrophilic organic anion agents [3]. Moreover, BCRP is an important constituent and an integral part of the blood-tissue barrier, which carries out the function of absorption and excretion. Therefore, BCRP performs vital tasks for drug distribution, absorption, metabolism, toxicity, and excretion [4]. In addition, the food and drug administration authority (FDA) has considered BCRP one of the important drug transporters that may use for drug disposition [5]. Several drugs, such as topotecan and mitoxantrone, are identified as BCRP substrates, and their inhibitors can enhance drug bioavailability. The discovery of BCRP has significantly improved cancer rehabilitation, treatment, and clinical management [6]. Before the drug development process, an efficient and automatic in silico model is necessary to predict BCRP substrates. This model may aid drug discovery and its interactions with other drugs. A list of frequently used acronyms is provided in Table 1.

Table 1: A list of frequently used acronyms

Abbreviation	Definition
BCRP	Breast cancer resistance protein
TL	Transfer learning
DL	Deep learning
LDA	Linear discriminant analysis
DF	Deep features
DNN	Deep neural networks
CNN	Convolutional neural network
ANOVA	Analysis of variance
ResNet101	Residual network of 101 layers
SVM	Support vector machine
AUC	Area under the curve
ROC	Receiver operating characteristics
LR	Logistic regression
NB	Naïve Bayes
RF	Random forest
KNN	K-nearest neighbor
QSAR	Quantitative structure-activity relationship

2 Related Studies

In silico methods are quickly gaining attention from several researchers to target developing and discovering various types of anticancer drugs. Moreover, these methods are useful for selecting compounds from large databases and libraries. Specific to BCRP substrate prediction, most of the proposed techniques are based on conventional machine learning and quantitative structure-activity relationship (QSAR). For example, Zhong et al. [7] used genetic algorithms (GA), conjugate gradient (CG), and support vector machines (SVM) for the prediction of BCRP substrate and non-substrate structures. They calculated 1000 molecular descriptors (features) to facilitate classification tasks and obtained training and testing accuracy of 91.3% and 85.0%, respectively. Hazai et al. [8] employed the SVM model to predict BCRP substrates and reported an accuracy of 73.0%. They also extracted 3250 molecular descriptors using the Dragon tool for classification problems. Sasahara et al. [9] applied machine learning models, such as the XGBoost and Naïve Bayes based architectures, and reported an accuracy of 90.0% for drug design classification. In another study, Sasahara et al. [10] proposed a drug metabolism prediction model by adopting different conventional classification techniques and reported an accuracy of 79.9% on the test dataset. Hassan et al. [11] proposed a drug response prediction model for in-vitro human liver cancer cell lines using reduced features and a quadratic discriminant analysis (QDA) algorithm. The study reported an accuracy of 98.0% on the test dataset for HepG2 cancer cell line classification.

Furthermore, Kato [12] reported an *in silico* prediction of cytochrome P450 used for drug development and discovery. Similarly, Rácz et al. [13] conducted a comprehensive survey regarding the machine-learning models being adopted for drug discovery, delivery, and safety. Ambe et al. [14] proposed an in silico-based approach for predicting the Hepatocellular Hypertrophy using various descriptors and obtained an accuracy value of 76.0%. In another study, Jian et al. [2] reported BCRP inhibitors and non-inhibitors prediction models using ensemble machine learning and deep learning models and achieved the Mathew correlation coefficient (MCC) and area under the curve (AUC) values of 81.20% and 95.80%, respectively. The authors extracted several conventional descriptors and employed different machine learning algorithms for classification. Even more, Sammut et al. [15] proposed a breast cancer therapy response prediction system using ensemble learning approach and reported an AUC value of 87.0%.

In literature, most of the methods reported for BCRP substrate classification are based on conventional quantitative structure active relationship (QSAR) and machine learning approaches [7,16,17]. Moreover, these methods require different kinds of chemical molecular descriptors. Generally, these traditional descriptors are used for the quantitative description of various compound properties, which may vary from different studies and are employed as inputs for modeling. More precisely, these chemical properties, including topological, geometrical, physicochemical, and electronic characteristics, are extracted using different software programs, such as Dragon, which extract more than 5000 descriptors. Before the modeling process, a preprocessing step is needed to eliminate redundant and irrelevant descriptor properties. Owing to the diversity of chemical structures, these approaches are less generalized, computationally expensive, costly, and prone to user dependence.

Given this, an urgent need exists for an automatic, cost-effective, and optimized feature extraction strategy that requires minimum or no user intervention for the precise classification of BCRP substrates. Contrary to conventional methods, a new approach is proposed for this goal, which exploits structural fingerprints of chemical compounds to develop a customized ResNet101 deep neural network (DNN) model via transfer learning (TL) technique for deep feature extraction. Subsequently, these extracted deep features (DF) are analyzed using linear discriminant analysis

(LDA) to classify BCRP substrate structures appropriately. This novel proposed approach, TLRNDF-LDA, utilized images of the molecular structure of different chemical compounds as input for automatic classification. In particular, the customized ResNet101 model can automatically generate complex and optimized feature maps at network's low, middle, and high-level layers. In this way, it overcomes conventional descriptor/feature extraction strategies by minimizing user intervention. To the authors' knowledge, no existing research has exploited structural fingerprints of substrates and non-substrates chemical compounds for DF extraction, followed by LDA classification. The pipeline of the proposed approach is flexible and can be extended for other types of classification tasks, such as inhibitors and non-inhibitors; much more can aid in examining drug discovery and delivery. The major contributions of this work are as follows:

- Customized ResNet101 CNN model is trained using TL approach on appropriate sample size selected by statistical power method.
- Encoder is introduced to accept large input-size images instead of the ResNet101 default input to preserve the important structural information for the classification task.
- LDA algorithm is opted for classification and trained on extracted optimized DF at the 'Pool_5' layer of the TL-ResNet101 model. The developed model is compared with state-of-the-art techniques, such as support vector machines (SVM), logistic regression (LR), naïve Bayes (NB), random forest (RF), and k-nearest neighbor (KNN) on the same substrate and non-substrate DF for appropriate BCRP substrate prediction.
- Statistical analysis is performed using the ANOVA test to evaluate the efficacy of the proposed model.

The remainder of this paper is as follows: Section 3 explains the material and methods of the proposed approach. Section 4 presents the result of the proposed approach and its comparison with state-of-the-art methods. Section 5 discusses the study, and Section 6, consists of a comprehensive conclusion of the study.

3 Material and Methods

3.1 Material

3.1.1 Sample Size Selection

In experimental research studies, selecting an appropriate sample size for proper model development and validation of results is crucial. For DL model training and testing, an enormous amount of annotated data is required due to the very large network size of the model. However, only a small amount of annotated data is available for medical diagnosis. Therefore, it is recommended to use certain statistical approaches to estimate sample size for study justification and experimentation [11]. For this purpose, G∗Power statistical tool is adopted to estimate the sample size for DL model training and validation in the context of TL concept. In hypothesis testing, α and β are the probability of Type I error (incorrectly rejecting the null hypothesis) and Type II error or false negative rate (incorrectly failing to reject the null hypothesis), respectively. Power is the probability of not making a Type II error. Mathematically, it can be given by $(1 - \beta)$. Specifically, G∗Power statistical tool requires various input parameters, such as effect size $|\rho|$ and α-value for adequate sample size estimation [18,19]. The parameters are set empirically at $|\rho| = 0.3$, $\alpha = 5\%$ level of significance and a one-tailed t-test. It is deduced from Fig. 1 that a minimum of 110 samples are required at given parameter values to adequately conduct the study for experimentation of BCRP substrates prediction. However, a dataset of 332 BCRP substrates and non-substrates chemical structures is collected from other literature to develop the proposed model. The next section explains the dataset in detail.

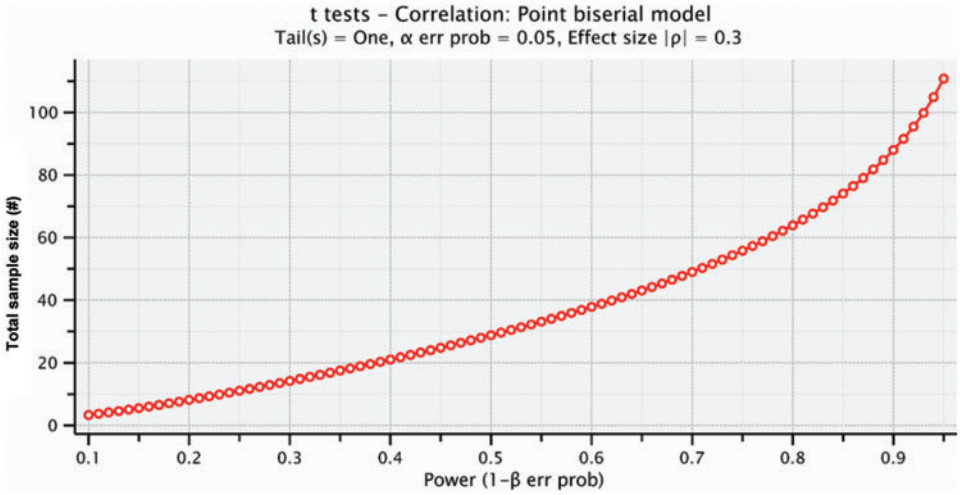


Figure 1: Sample size selection (at Power $1 - \beta = 95\%$) using the G*Power statistical tool for model training

3.1.2 Dataset of BCRP Substrates and Non-Substrates

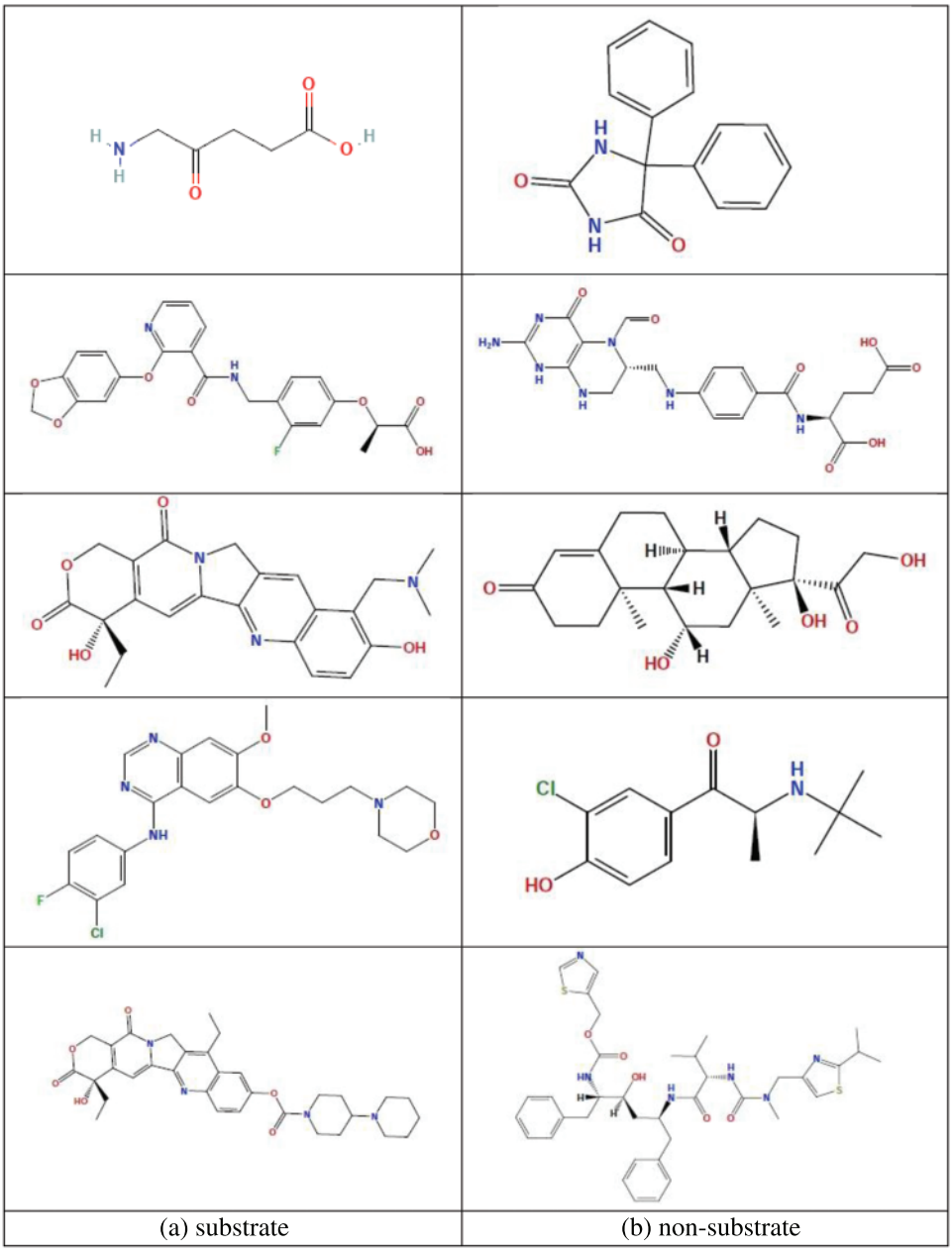
A minimum sample size of 110 is needed to adequately conduct this research study for statistically acceptable prediction models (Section 3.1.1). However, 332 diverse chemical structure-based compound images of BCRP substrates and non-substrates are obtained from the open-access literature. The dataset consists of 230 'substrate' and 102 'non-substrate' compounds of human wildtype BCRP. These molecular structures are confirmed from the public repository https://pubchem.ncbi.nlm.nih.gov/. The structurally diverse 332 compounds are given in supplementary material. Illustrative sample dataset images of the chemical structures of BCRP substrates and non-substrates are provided in Fig. 2. The left column of Fig. 2 shows the BSCRP substrates, whereas the right column includes non-substrates chemical structures. It is evident from Fig. 2 that there is diversity in the molecule, which has simple and complex structures. It is hard to differentiate the structures by utilizing conventional machine learning algorithms. The colors shown in Fig. 2 represent the atom of various types and their bond with others and have no impact on the classification tasks.

The dataset is randomly divided using an 80%:20% ratio for model training and testing. In addition, a rigorous 5-fold cross-validation technique is applied for model development, which offers generalization and avoids overfitting. The experiments are conducted using the Intel Xeon E-2246 G, 3.6 GHz processor, 16 GB RAM, and are equipped with NVIDIA GeForce GTX-1050Ti GPU and Matlab 2020(a).

3.2 Methods

Previously reported methods for BCRP substrate classification use conventional QSAR and machine learning algorithms. These methods typically require different kinds of traditional quantitative chemical molecular descriptors to describe various properties of a compound, such as topological, geometrical, physicochemical, and electronic characteristics. These methods are less generalized, computationally expensive, and prone to user dependence due to the diversity of chemical structures. In this context, contrary to conventional methods, a novel in silico intelligent method is proposed, which exploits structural fingerprints of chemical compounds to develop a customized ResNet101

deep neural network model via the TL technique for DF extraction. Afterward, extracted DF is fed to the LDA algorithm, which will predict the cancer protein substrate structures in the human breast. The framework of the proposed intelligent system is shown in Fig. 3. The details of each component of the proposed methodology are given as follows.



(a) substrate (b) non-substrate

Figure 2: Samples of BCRP substrate and non-substrate structures of chemical compounds

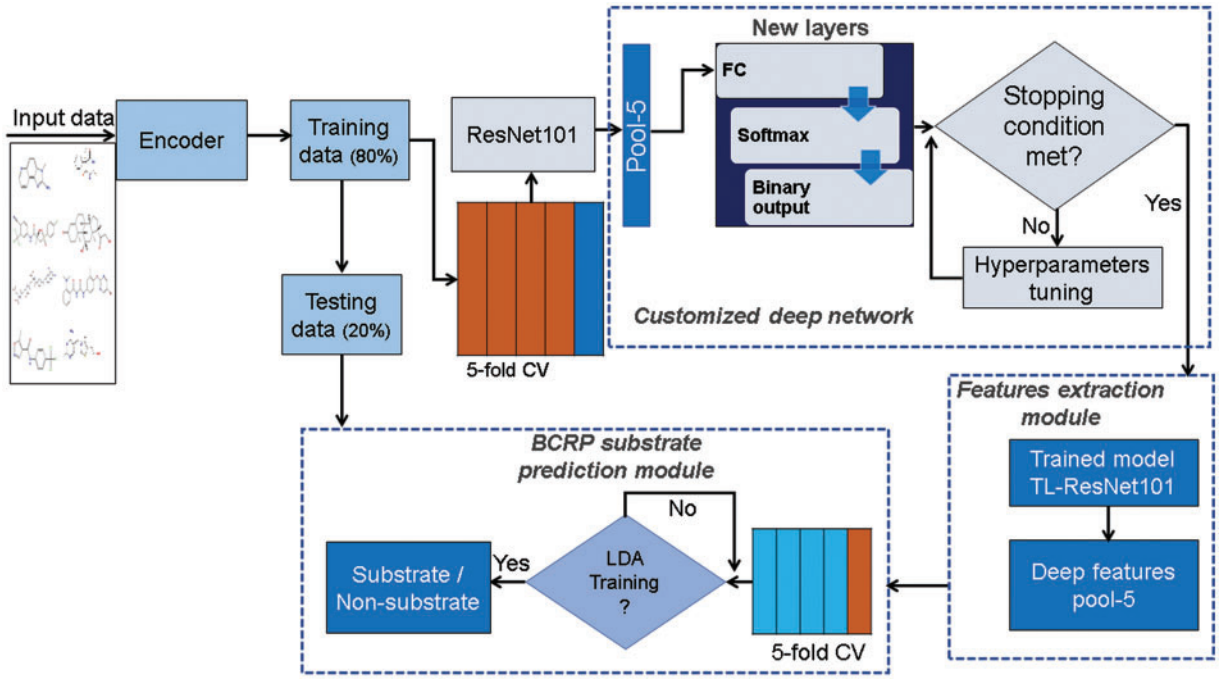


Figure 3: The framework of the BCRP substrate prediction system

3.2.1 The Encoder

The encoder is introduced to accommodate the larger input size BCRP substrate and non-substrate molecular structure images. Although the original ResNet101 DNN model accepts fixed size input ($224 \times 224 \times 3$) images, the dataset used in this study contains different image sizes. Due to the sensitive and important information of substrate structures, it is highly desirable to use larger input layers to preserve all structural details for accurate prediction. Therefore, the first layer is replaced by deploying an encoder to accept the input size ($270 \times 350 \times 3$) of BCRP substrate and non-substrate molecular structure images. Interestingly, the impact of the encoder has sufficiently enhanced the performance of the proposed models by retaining all structural details contained in input molecular compound images.

3.2.2 Customized ResNet101 Deep Learning Model

He et al. [20] first introduced the residual network (ResNet) model–the winner of the ImageNet challenge for the classification task of 1000 classes. The ResNet model has several variants, such as ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152 adapted to solve real-world problems. The basic architecture of these networks is the same, with varying depths. The increasing network depth enables them to solve complex problems with improved classification accuracy but at a relatively higher computational cost [20]. The ResNet model architecture contains stacking of several convolutional and pooling layers to perform image feature extraction followed by classification layers. Moreover, the ResNet deep learning architecture has a skip connection or residual learning capability, enabling the network to explore and learn hidden image patterns better than other networks, such as AlexNet, GoogLeNet, and VGG [11,20,21]. In this study, ResNet50, ResNet101, InceptionV3, GoogLeNet, and MobileNetV2 models are customized and trained to determine the best network for the current

classification task. Interestingly, among these DL models, customized ResNet101 demonstrates the highest performance (see Section 4). For this study, developing a customized ResNet101 network is further explored as a baseline model to solve the problem of BCRP substrates classification.

Generally, deep learning models require sufficient annotated data for training and validation. Moreover, due to inheriting limited BCRP substrates and non-substrates data, using the existing ResNet101 network is convenient via a TL approach. TL is a modern method for resolving new problems by utilizing the weight-sharing mechanism of existing DNN models. Here, ResNet101 is customized by replacing the input layer with the encoder and the last three layers (FC_1000, FC_Softmax, and Class_output) with three new layers, namely FC_2, FC_Softmax, and Binary_classoutput. Excluding these four layers, the rest of the network's learned weights are used to classify the substrates and non-substrates. The modified architecture of TL-ResNet101 is shown in Fig. 4. Each convolutional block consists of several filters mentioned in the squared (yellow) block with several stacked blocks in small rectangles (blue). For instance, Conv2_x block contains 1 × 1, 3 × 3, and 1 × 1 with 64, 64, and 256 filters, along with three stacked blocks. The stacked layers for each residual block are 3, 4, 23, and 3, which are shown in Fig. 4. Average pool operation is performed at the 'pool_5' layer of the network. The remaining three layers are introduced to solve the new problem of BCRP substrate structure prediction problem.

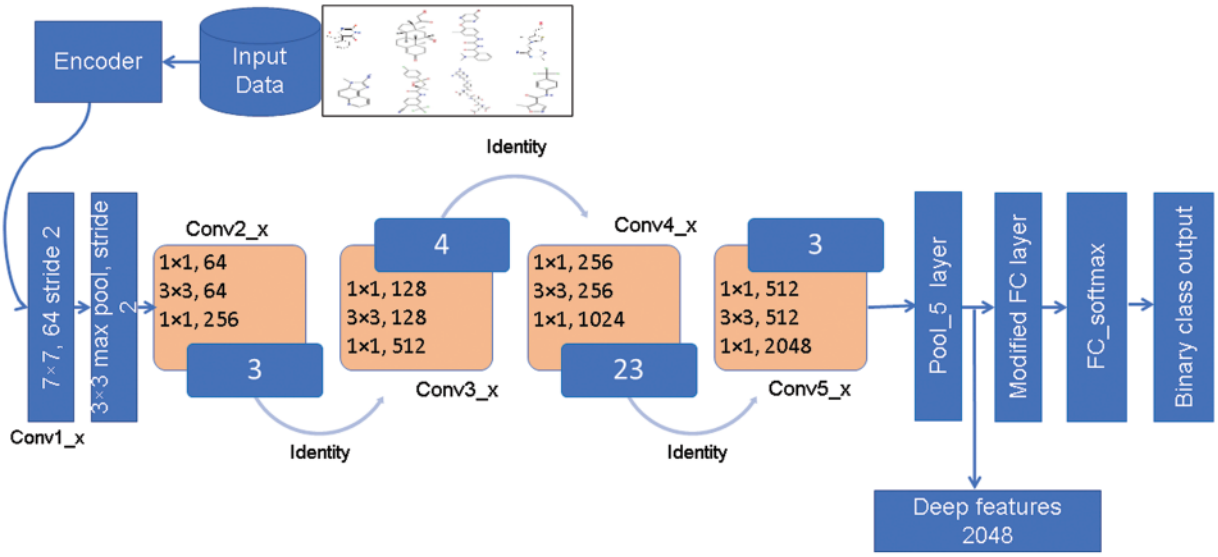


Figure 4: The customized TL-ResNet101 architecture for BCRP substrates and non-substrates classification

Furthermore, during the training process of the TL-ResNet101 network, the network learns the weights of new layers while the rest of the network weights are frozen. For better generalization, TL-ResNet101 hyperparameters are tuned to obtain a precise classification. As indicated in Fig. 4, the DF is obtained at the 'Pool_5' layer of the network and fed to LDA for classification. The main components of ResNet101 are explained below.

3.2.3 The Convolution Layers

Like other deep convolutional neural networks, ResNet101 contains several cascading convolutional layers with varying kernel sizes. These layers and kernel sizes are shown in Fig. 4. These layers

aim to learn and extract hidden patterns, which are further used for the classification task. The input image is passed to the network during learning to learn weights for newly replaced layers. In this way, the network can classify the new dataset of BCRP substrates and non-substrates. The input $M^{n}_{x\_input}$ feature map to the convolution layer at an arbitrary index n is defined as follows:

$$M^{n}_{x\_input} = \frac{M^{n-1}_{x\_input} - K^{n}_{x\_input}}{S^{n}_{x\_input} + 1} + 1 \tag{1}$$

The following equation obtains the output $M^{n}_{y\_output}$ convolution feature.

$$M^{n}_{y\_output} = \frac{M^{n-1}_{y\_output} - K^{n}_{y\_output}}{S^{n}_{y\_output} + 1} + 1 \tag{2}$$

where S, K_{x_input} and K_{y_output}denote stride, input, and output Kernel sizes, respectively.

3.2.4 The Pooling Layers

The pooling layers are introduced in ResNet101 architecture to reduce image size and enhance computational efficiency. Another objective is incorporating invariants considering scale, rotation, and translation of the objects in input images. The modified TL-ResNet101 architecture uses the max pooling operation expressed by the following equation.

$$y_i = \max_{1 \le j \le M \times M} (x_j), x_j \in \mathbf{X}_i \tag{3}$$

in which $X = [X_0, X_1, .., X_n]$ shows input image regions, and the i^{th} region is given by $X_i = (x_1, x_2, .., x_{M \times M})$ with a size of M. A kernel size 2×2 with stride $S = 2$ is used in pooling operation. Therefore, the computational complexity is significantly reduced via efficient reduction of input image size and much more, exploring the discriminant and hidden features of the image.

In addition, TL-ResNet101 was used to assess the network's weight learning validation iteratively. This study adopts a binary cross-entropy loss function during model training. The modified loss function for the classification of BCRP substrates is given below.

$$LF_{Binary} = -z_i . \log(p(z_i)) + (1 - z_i) . \log(1 - p(z_i)) \tag{4}$$

where $p(z)$ is the probability of TL-ResNet101, which helps to predict class z.

The class probabilities of 'substrate' and 'non-substrate' are computed from the input of the modified fully connected (FC_2) layer using an activation function. As illustrated in Fig. 4, the modified FC_2 layer received optimized DF from the 'Pool_5' layer. The Softmax activation function is used to find the class probabilities. The mathematical representation of the Softmax function is given as follows;

$$softmax\ \left(\widehat{Y}_i\right) = \frac{\exp(y_i)}{\sum_{j=1}^{2} exp(y_j)} \tag{5}$$

where $\widehat{Y}$ is the class probability obtained from the i^{th} input vector y.

3.2.5 The ResNet101 Transfer Learning

The pre-trained ResNet101 is customized via the TL approach to solving a new challenge regarding the BCRP substrate and non-substrate classification. Specifically, it is practical and useful approach when the dataset size is relatively small. In TL, the weight-sharing mechanism is achieved by freezing the weights of some layers in the existing network and then retraining for newly introduced

layers. This procedure is successfully being adopted to solve several problems, such as human liver cancer drug response, HBV, skin cancer diagnosis, and plant disease predictions [11,22–25].

This study used the ResNet101 deep learning model trained on the ImageNet dataset with 1000 classes and employed the TL concept to classify substrates and non-substrates binary class problems. As depicted in Fig. 3, the partial weights of the ImageNet are used, and three new layers (FC_2, FC_softmax, and Binary_classoutput) weights are learned by customized ResNet101 architecture for prediction of 'substrate' and 'non-substrate'. Furthermore, for better TL-ResNet101 training and validation, hyperparameter tunings, such as batch size 20, learning rate 0.0001, data augmentation of range (−5, +5), and a total number of epochs is set to 1000. These parameters are set empirically and result in improved performance.

3.2.6 Deep TL-ResNet101 BCRP Feature Extraction

This study has exploited deep features contrary to conventional molecular descriptors to accurately predict BCRP substrates and non-substrates. More precisely, the customized TL-ResNet101 model automatically learns and generates complex feature maps at low, middle, and high-level layers. Thus, it outperforms conventional descriptor/feature extraction strategy by minimizing user intervention. After TL-ResNet101 deep learning model is trained with substrates and non-substrates molecular structure images, a set of DF at the 'Pool_5' layer is extracted, as shown in Fig. 4. These DFs are further used for LDA classification. Similarly, several real-world problems are solved by the TL-based deep feature extraction model followed by classification [11,21]. A feature vector $x_{DF_1}, x_{DF_2}, \ldots, x_{DF_N}$ ($x_{DF_i} \in \Re^n$) of $n = 2048$ dimensions is obtained for every image after the fifth residual block of TL-ResNet101 model. Moreover, unlike the conventional feature extraction strategy, DF is extracted without user intervention and parameter optimization. The developed TL-ResNet101 model is trained based on randomly selected 80% of the data via a 5-fold cross-validation technique. The trained model is evaluated on the remaining 20% of unseen data. Particularly, TL-ResNet101 outperforms other convolutional networks, such as GoogLeNet and InceptionV3, because of its residual learning capability. Besides, it has a lower probability of overfitting and better generalization than other DNN models [26,27]. The extracted DF is fed to LDA to classify BCRP substrates and non-substrates.

3.2.7 Linear Discriminant Analysis (LDA) Classification

Typically, convolutional neural network (CNN) models utilize the Softmax activation function, a generalized logistic regression (LR) form, to classify input feature vectors. However, the LDA algorithm is employed instead of Softmax for efficient substrates classification due to the complex nature of BCRP substrates and non-substrates molecular structure data. The LDA provides better performance and classifies feature vectors more precisely. In the existing literature, it has been reported that LDA offers better performance with various complex data. For instance, the re-identification of persons is performed by the LDA classifier [28]. In another study, LDA is adopted for multiclass problems of plant disease detection and offers superior performance [24]. Similarly, LDA algorithm is used to classify DF for a medical image modality classification task [21]. To the authors' knowledge, DF extraction followed by LDA classification of substrates and non-substrates has yet to be explored.

The new classification module consists of DF of the dataset $\mathrm{DF_{TLNN}} = [\mathrm{df_1}, \mathrm{df_2}, \ldots, \mathrm{df_n}]$, along with two 'substrate', and 'non-substrate' classes of size 'm × n', where m is the number of data samples. The LDA model is trained on 80% of DF dataset and randomly selected using 5-fold cross-validation. The remaining 20% of the unseen DF dataset is fed to the trained model for testing and model evaluation.

The DF_{TLNN} is fed to LDA to get an LD score matrix Z.

$$\boldsymbol{Z} = \boldsymbol{DF}_{\boldsymbol{TLNN}} \times \boldsymbol{W} \tag{6}$$

The objective of linear combination is to optimally draw a decision boundary for BCRP substrates and non-substrates datasets. Specifically, it searches the optimal weight vector $W = [w_1, w_2, \ldots, w_l]$, with l solutions that maximize the rate of inter and intra-class scatter. Between class scatter CS_{bcs} is represented as follows.

$$CS_{bcs} = \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T \tag{7}$$

and CS_{Ics} (intra-class scatter) is defined as:

$$CS_{Ics} = \sum_{i=1}^{c} \sum_{j=1}^{m_j} (\mu_j - \mu_i)(\mu_j - \mu_i)^T \tag{8}$$

where μ_i is the average value of ith class, m_j is the sum of observations of ith respective class, μ_j is an observation instance, and T is a transpose symbol.

The objective function of LDA, $J(\boldsymbol{W})$ was evaluated by employing inter-and intra-class scatter from Eqs. (7) and (8):

$$J(\boldsymbol{W}) = \frac{\boldsymbol{W}^T CS_{bcs} \boldsymbol{W}}{\boldsymbol{W}^T CS_{Ics} \boldsymbol{W}} \tag{9}$$

Specifically, it searches for the optimal $\boldsymbol{W}^*$ weight vector is associated with the discriminant function of the variables, such that function J is maximized. More so, LD scores "Z" matrix shows a compact format of original deep features, "$\mathbf{DF}_{\mathbf{TLNN}}$," which efficiently differentiates 'substrate' and 'non-substrate' classes. The specifics of LDA can be found in [29,30].

Interestingly, testing of the TLRNDF-LDA is straightforward. Specifically, the input chemical structure of compounds is fed to the trained model to get DF, followed by LDA for accurate prediction of the 'substrate' or 'non-substrate' class.

3.2.8 Model Performance Evaluation

The performance of the developed models is examined using standard evaluation parameters, such as accuracy, sensitivity, specificity, F-score, Mathew correlation coefficient (MCC), receiver operating characteristics (ROC) curve, and area under the curve (AUC). These parameters are also used to compare the proposed method with other contemporary methods. Besides, evaluation is performed on model training and testing data. The detail of these parameters can be found in [21].

4 Experimental Results

For BCRP substrate prediction, extensive experiments are conducted to evaluate the performance of the proposed TLRNDF-LDA approach. First, various DL models, such as ResNet50, ResNet101, InceptionV3, GoogLeNet, and MobileNetV2 are trained by employing the TL approach and are used for the classification of the dataset. All developed models use similar hypermeter tuning strategies and are trained and tested on 80%:20% randomly split datasets, respectively. Among these DL models, due to their impressive performance, ResNet101 model was selected for DF extraction prior to the LDA classification task. Moreover, several other classification algorithms, such as SVM, LR, Naïve

Bayes (NB), random forest (RF), and k-nearest neighbors (KNN) are trained on extracted DF for the classification of the substrate and non-substrate chemical compound structures.

The visualization of DF extracted from the customized TL-ResNet101 model is depicted in Fig. 5. Interestingly, the features scatter plot shows the distinction of the classes. Fig. 6 shows the training accuracy and loss curves of the modified TL-ResNet101 model. During the training process, the optimal model is obtained against 1000 epochs (iterations) and extracted most useful DF, which are further fed to LDA for the classification tasks. The training and testing performance of various DL models for the classification of substrates is shown in Table 2. Similarly, the performance of the developed models on the unseen 20% test dataset is presented in Table 3.

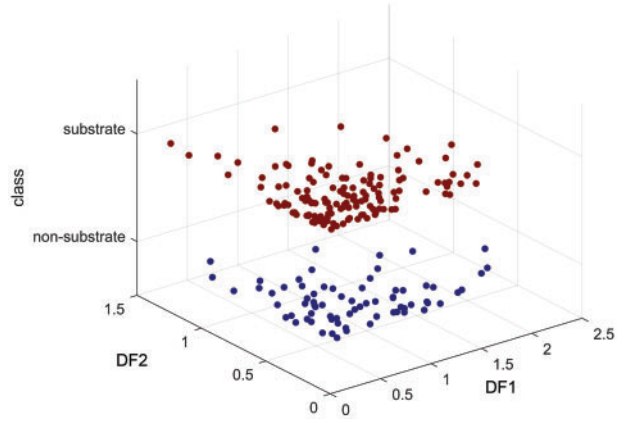


Figure 5: The scatter plot of TL-ResNet101 deep features (DF1 and DF2) for substrate and non-substrate datasets

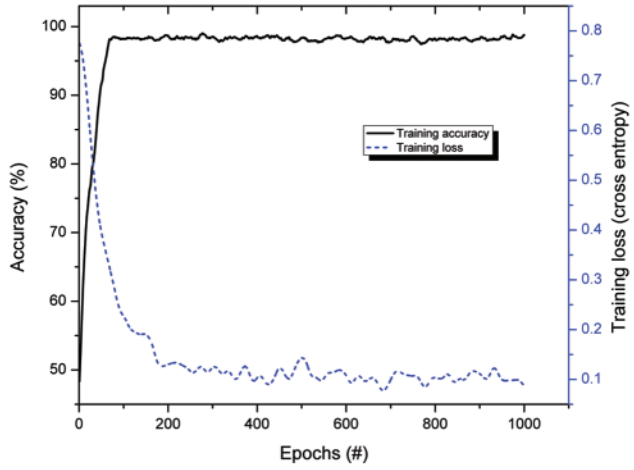


Figure 6: Training accuracy and loss curves of the proposed baseline customized TL-ResNet101 model

Table 2: Training performance comparison of various DL models by employing TL to select baseline model for prediction of substrates and non-substrates

Models	Accuracy	Sensitivity	Specificity	F-score	MCC
TL-ResNet50	0.9561	0.9643	0.9412	0.9677	0.9021
TL-InceptionV3	0.7231	0.8372	0.5000	0.8000	0.3566
TL-GoogLeNet	0.8458	0.824	0.7692	0.8856	0.6490
TL-MobileNetV2	0.7179	0.9535	0.2576	0.8173	0.9096
The proposed TL-ResNet101 (baseline)	**0.9808**	**0.9853**	**0.9722**	**0.9853**	**0.9575**

Table 3: Performance comparison of developed DL models by employing TL on test dataset to select baseline model for prediction of substrates and non-substrates

Models	Accuracy	Sensitivity	Specificity	F-score	MCC
TL-ResNet50	0.8056	0.8750	0.6667	0.8571	0.5543
TL-InceptionV3	0.6327	0.7188	0.4706	0.7188	0.1893
TL-GoogLeNet	0.6170	0.6250	0.6000	0.6897	0.2110
TL-MobileNetV2	0.5714	0.6250	0.4706	0.6557	0.0925
The proposed TL-ResNet101 (baseline)	**0.9631**	**0.9691**	**0.9512**	**0.9721**	**0.9176**

Among several developed DL models, TL-ResNet101 indicated well performance on DF, and thus, it is further used for classification. More so, a statistical ANOVA test is performed with the null hypothesis that all TLRNDF-based models have similar average performance. Statistically, rejection of the null hypothesis has been observed from the ANOVA test, and results are shown in Table 4. The performance of the proposed TLRNDF-LDA and other approaches are given in Table 5. It may be inferred from Table 5 that the proposed approach provides superior performance at all quality measures compared to other developed models on the same dataset.

Table 4: The ANOVA test on the developed approaches using DF, such as LDA, SVM, LR, NB, RF, and KNN classifiers

Source	SS	DF	MS	F	*p*-value
Models	0.2492	5	0.0498	10.28	0.0000
Error	0.1454	30	0.0048		
Total	**0.3946**	**35**			

Table 5: Performance comparison of the proposed TLRNDF-LDA and other approaches for prediction of BCRP substrates and non-substrates

Models	Accuracy	Sensitivity	Specificity	F-score	MCC	AUC
TLRNDF-SVM (linear)	0.9712	0.9714	0.9706	0.9784	0.9352	0.9375
TLRNDF-SVM (Quadratic)	0.9712	0.9712	0.9710	0.9783	0.9356	0.9380
TLRNDF-LR	0.7740	0.8214	0.6765	0.8303	0.4926	0.7708
TLRNDF-NB	0.9153	0.9220	0.8955	0.9353	0.8057	0.8958
TLRNDF-RF	0.9038	0.8824	0.9636	0.9310	0.7869	0.8750
TLRNDF-KNN (k = 3)	0.8592	0.8544	0.8727	0.9000	0.6751	0.8541
The proposed TLRNDF-LDA	**0.9856**	**0.9926**	**0.9722**	**0.9890**	**0.9681**	**0.9583**

In medical diagnostics, ROC curves are one of the important performance evaluation measures. Fig. 7 compares ROC curves of the proposed TLRNDF-LDA and other approaches. The ROC curve of the proposed approach is close to the vertical axis, which indicates its effectiveness. More so, comparison in terms of AUC of the proposed approach and other developed models are shown in Table 5. These AUC values are computed from ROC curves shown in Fig. 7. A high AUC value confirmed that the proposed model could identify BCRP substrates successfully and, subsequently, validated from the maximum area covered by respective ROC curve.

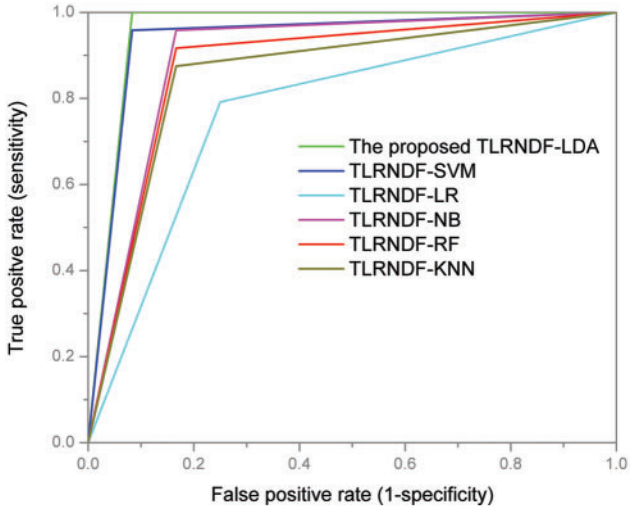


Figure 7: ROC curves of the proposed TLRNDF-LDA method and other approaches

The AUC error plot limits, as shown in Fig. 8, indicate that the proposed TLRNDF-LDA has a high value of AUC with fewer error limits than other developed models. Similarly, ANOVA boxplots (multi-comparison) of the proposed model and other models are shown in Fig. 9. The proposed model offers high-performance values with low scatter compared to other models on the same dataset. Fig. 9 validates that the proposed approach, TLRNDF-LDA, outperformed all other developed approaches using SVM, LR, NB, RF, and KNN classifiers.

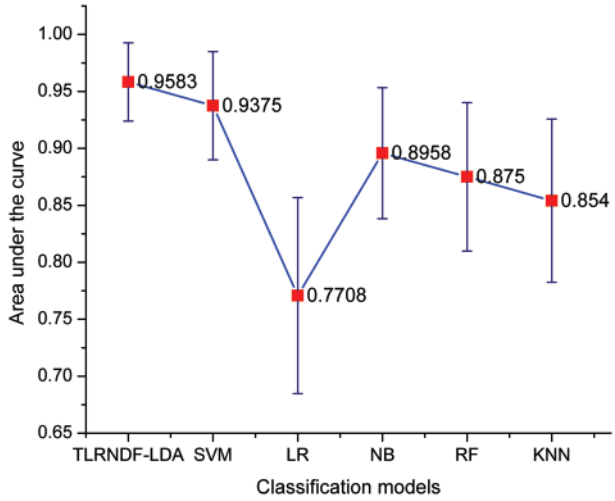


Figure 8: Comparison of AUC error plots of the proposed TLRNDF-LDA method with other developed models using DF, such as SVM, LR, NB, RF, and KNN classifiers

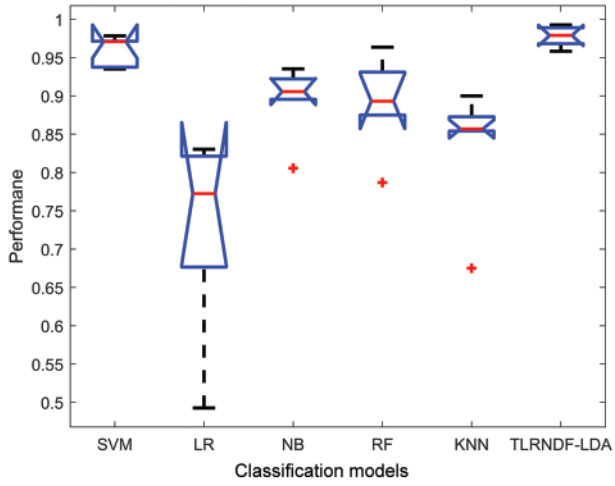


Figure 9: The ANOVA boxplots of the proposed TLRNDF-LDA and other developed models using DF, such as SVM, LR, NB, RF, and KNN classifiers

5 Discussion

Breast cancer resistance protein (BCRP) is one of the vital drug transporters required for clinical drug discovery and disposition. It belongs to adenosine triphosphate (ATP) binding cassette (ABC) efflux transporter proteins and is considered an important multidrug resistance protein for the therapy of different kinds of cancer. It also aids the absorption, distribution, and elimination of drugs. However, the early identification of BCRP substrates to inform the appropriate optimization of anticancer drug discovery, treatment, and rehabilitation process, is quite challenging. For this study, a new intelligent approach is developed to solve the challenging prediction problem of BCRP substrate and non-substrate. It is effective for cancer drug development, treatment, and rehabilitation.

It is observed from previous studies that conventional approaches, like QSAR and machine learning algorithms for BCRP substrate classification, require traditional quantitative chemical molecular descriptors, such as topological, geometrical, physicochemical, electronic characteristics, etc. The traditional approaches are less generalized, computationally expensive, and prone to user dependence due to the diverse nature of chemical structures. Specifically, several issues are associated with conventional approaches, such as the number of features, types, and selection of the most discriminant features. These methods are highly dependent on the expertise and the availability of computational resources. Therefore, the current prediction problem of BCRP substrate and non-substrate is solved by exploiting structural fingerprints of chemical compounds; and much more utilized a customized ResNet101 model by employing the transfer learning (TL) approach for deep feature extraction followed by LDA classification. Contrary to conventional approaches, the proposed method is an automatic, cost-effective, and optimized feature extraction strategy with minimum or no user intervention for precise classification of BCRP substrates.

First, an encoder is introduced to incorporate a large image input size of $270 \times 350 \times 3$ instead of the default image size ($224 \times 224 \times 3$) of ResNet101. In this way, the model effectively preserves the minute details contained in BCRP substrate and non-substrate structure images, significantly contributing to the prediction improvement. In the case of adopting default input image size, there is the possibility that some details of image structure may vanish. Thus the model may lose useful information for adequate classification.

In the medical diagnosis field, annotated data is inherently small, whereas training of DL models requires enough data. Therefore, an important TL weight-sharing strategy is employed, which was used to solve the new problem by partially utilizing existing network weights. However, the minimum sample size required to apply the TL concept is another challenge that was addressed statistically. More precisely, the statistical G∗Power tool was employed to select a suitable sample size for model training. As illustrated in Fig. 1, at least 110 samples are required for customized TL-ResNet101 model training. Meanwhile, a dataset contains 332 substrate and non-substrate structure images to satisfy this minimal sample size requirement.

Furthermore, to make the pre-trained ResNet101 model compatible with the present binary classification problem, the model is modified by replacing the first and last three layers with new layers of the network for BCRP substrates classification. Specifically, the first 'input_layer' was replaced with an encoder, and the last three layers 'FC_1000', 'FC_Softmax', and 'FC_Classoutput', were replaced with new FC_2, FC_Softmax, and Binary_classoutput layers, respectively for the binary classification problem. The customized architecture of TL-ResNet101 is shown in Fig. 4. A DF vector of size 2048 at the 'Pool_5' layer is obtained after training the model and fed to LDA classifier for final prediction (Fig. 3). The performance of the TL-ResNet101 model on both training and testing datasets is shown in Tables 2 and 3, respectively. More so, the customized model provides high training performance values of 98.08%, 98.53%, 97.22%, 98.53%, and 95.75% for accuracy, sensitivity, specificity, F-Score, and MCC measures, respectively. Similarly, the developed model offers performance on unseen test datasets with values of 96.31%, 96.91%, 95.21%, 91.76%, and 91.76% for accuracy, sensitivity, specificity, F-score, and MCC measures, respectively. Meanwhile, other DL models, such as TL-ResNet50, TL-GoogLeNet, TL-InceptionV3, and TL-MobileNetV2 were also trained and tested on the same dataset and with the same hyperparameter tuning as used for the TL-ResNet101 model. It is observed that TL-ResNet101 provides superior performance in all measures compared to all other models. Regarding accuracy, the TL-ResNet101 improves 15.75% on the test dataset (Table 2) compared to the second-best TL-ResNet50 model. It is understood that the impressive performance

of TL-ResNet101 model is due to its depth and residual learning mechanism. Besides, although TL-GoogLeNet, TL-InceptionV3, and TL-MobileNetV2 models are very deep, the lack of residual weight learning capability during training results in low performance compared to ResNet-based models. Given the high performance of the TL-ResNet101 model, it is used as a baseline model for DF extraction.

It has been observed from Tables 2 and 3 that a sufficient difference exists between training and testing performance in all performance assessment measures. For instance, the customized TL-ResNet101 model has a 1.77% difference in accuracy value, whereas there is a 4% difference in MCC measure between training and testing performance, respectively. Comparing Table 1 (training) with Table 3 (testing) shows that the difference in pairwise performance is sufficiently high, exceeding 15% in accuracy value for all models except TL-ResNet101. Among the developed DL models, the TL-ResNet101 difference in accuracy measure was 1.77%, which shows that TL-ResNet101 models provide better generalization compared to the other developed DL models on the complex BCRP substrates dataset. This fact suggested that TL-ResNet101 can be adequately used as a baseline model for extracting DF to achieve enhanced prediction (see Table 5).

Furthermore, conventional classification techniques utilize traditional molecular descriptors-based feature extraction strategies. However, in this study, structural fingerprints are used by exploiting DF contrary to conventional molecular descriptors for the classification of BCRP 'substrate' and 'non-substrate'. This strategy minimizes user intervention and sufficiently enhances the prediction performance of the classifier. Specifically, the developed TL-ResNet101 model automatically extracted DF at the 'Pool_5' layer of the network (Fig. 4). The substrate and non-substrate classification based on DF are shown in Table 5. The proposed TLRNDF-LDA model provides improved performance values of 98.56%, 99.26%, 97.22%, 98.90%, 96.81%, and 95.83% for accuracy, sensitivity, specificity, F-score, MCC, and AUC, respectively. It is observed from Table 4 that the proposed approach outperforms the other DF-based developed models, such as SVM (linear), SVM (quadratic), LR, NB, RF, and KNN, in terms of all quality measures. Thus, the proposed approach provides a minimum and maximum accuracy improvement of 1.44% and 21.16% concerning TLRNDF-SVM, and TLRNDF-LR models, respectively. In terms of AUC, it provides a minimum and maximum improvement of 2.03% and 18.75% for TLRNDF-SVM, and TLRNDF-LR models, respectively. Similar trends can be observed in Table 5 for all other quality measures.

The ROC curve is another important parameter for classification measurements, especially in medical diagnosis. It evaluates the average sensitivity test over the possible specificity range and vice versa. The comparison of ROC curves of the proposed TLRNDF-LDA with other approaches is shown in Fig. 7. It is observed that the curve of the proposed approach is very close to the vertical axis and covers a maximum of 95.83% AUC (Table 5), which shows its effectiveness. Moreover, it is important to observe ($\pm\sigma$) error range of AUC values, as shown in Fig. 8. It was revealed that the proposed model error range is minimum compared to other models, which validates the preciseness of the proposed approach for BCRP substrates prediction.

Performance of the proposed and other developed models are evaluated statistically by employing the ANOVA test for the null hypothesis – that all models developed on DF have no significant difference in average performance. The alternative hypothesis of having a significant difference in average performance is tested at a 5% level of significance. Table 4 shows ANOVA test results for all developed models (Table 4). It can be observed from Table 4 that p-value $<$ F-static, which supports the rejection of the null hypothesis and acceptance of the alternative hypothesis. It is demonstrated that the proposed TLRFDN-LDA model has significant difference in performance. Fig. 9 shows a

multi-comparison boxplot of the proposed and other models. It can be observed that the proposed model's performance is higher compared to other models. It is also observed from Table 5 that LDA and SVM performance is relatively better compared to other models because the DF of BCRP data is linearly separable.

The proposed TLRNDF-LDA approach is compared with other reported studies in the literature to predict BCRP substrate and non-substrate compound structures. Zhong et al. [7] proposed a BCRP classification approach using genetic algorithm (GA), conjugate gradient (CG), and support vector machines (SVM) techniques. They have reported training and testing accuracy values of 91.30% and 85.0%, respectively. In contrast, the proposed TLRNDF-LDA approach provides training and testing accuracy values of 98.08% and 96.31%, which are 6.78% and 11.31% higher than the reported approach on the same type of dataset. In addition, Garg et al. [4] also used SVM to classify substrate and non-substrate and reported 95.0%, 97.0%, 90.0%, and 89.0% values of accuracy, sensitivity, specificity, and MCC, respectively. The proposed model for this study has a performance of 3.56%, 2.26%, 7.22%, and 8.90% higher than [4]. Hence, the proposed model demonstrated better performance with full automation and, thus, does not require any user intervention in selecting features.

The proposed pipeline is developed using DL models by employing the TL concept. This pipeline can be utilized in other interesting areas and is expected to obtain better results over state-of-the-art approaches. These areas include biomedical and optoelectronics sensor data for monitoring health conditions, especially in ICUs and pandemic-like environments. For instance, Masud et al. developed highly sensitive biomedical optoelectronics sensors [31]. Moreover, Masud et al. developed an interesting two-mode biomedical sensor [32]. In another study, Masud et al. developed a dual-mode spectroscopic biomedical sensor using the Gabor expansion model [33]. Moreover, the data generated by these sensors can be fed to DL models, such as ResNet101, for real-time monitoring of critically ill patients. Class labels for model development can be generated using autoencoders and three sigma techniques.

Overall, it is inferred that the TLRNDF-LDA model provides superior performance in standard quantitative measures. More precisely, it is an intelligent, efficient, automated, and cost-effective approach for identifying BCRP substrates. The results are statistically validated and compared with other approaches, demonstrating their efficacy. Furthermore, the approach can be extended and used for other types of cancer for in silico classification. Moreover, the computer code and data are available for the researchers and can be provided for further experimentation.

6 Conclusion

Precisely identifying substrate structure is an important and challenging task due to the diversity of chemical structures. Conventional approaches utilize traditional quantitative chemical molecular descriptors-based feature extraction strategies to develop BCRP substrates classification models. These approaches are prone to a user dependent and less generalized. In this scenario, a novel deep-learning approach has been developed by modifying the ResNet101 DL model for automatic DF extraction and classification using the LDA algorithm. This study utilized structural fingerprints, which are exploited by DF instead of conventional molecular descriptors. The other contribution of this approach is an automatic, cost-effective, and optimized feature extraction strategy with minimum or no user intervention for better classification of BCRP substrates. Moreover, it offered the highest accuracy performance of 98.56% on test data compared to other state-of-the-art approaches. The proposed approach (TLRNDF-LDA) provides high classification performance compared to TLRNDF-SVM, TLRNDF-LR, TLRNDF-RF, TLRNDF-NB, and TLRNDF-KNN models on

different standard quality measures. For instance, the ROC curve of the developed model covers a maximum of 95.83% AUC which is higher than other models, indicating its effectiveness. In addition, the ANOVA test confirms the validity of the proposed approach by rejecting the null hypothesis at α 5% level. Finally, the proposed approach for this study can be extended for in vitro assessment efficacy of anticancer drug response prediction of prostate or lung cancer cell lines.

Acknowledgement: This research was supported by the BK21 FOUR Program (Fostering Outstanding Universities for Research, 5199991714138) funded by the Ministry of Education (MOE, Korea) and the National Research Foundation of Korea (NRF).

Funding Statement: The authors received no specific funding for this study.

Availability of Data and Materials: Dataset, computer code, and trained models are available and can be obtained by sending e-mail to the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. A. Doyle, W. Yang, L. V. Abruzzo, T. Krogmann, Y. Gao *et al.,* "A multidrug resistance transporter from human MCF-7 breast cancer cells," *Proc. of the National Academy of Sciences of the United States of America*, vol. 95, no. 26, pp. 15665–15670, 1998.

[2] D. Jiang, T. Lei, Z. Wang, C. Shen, D. Cao *et al.,* "ADMET evaluation in drug discovery. 20. Prediction of breast cancer resistance protein inhibition through machine learning," *Journal of Cheminformatics*, vol. 12, no. 1, pp. 16, 2020.

[3] H. Saito, H. Hirano, H. Nakagawa, T. Fukami, K. Oosumi *et al.,* "A new strategy of high-speed screening and quantitative structure-activity relationship analysis to evaluate human ATP-binding cassette transporter ABCG2-drug interactions," *The Journal of Pharmacology and Experimental Therapeutics*, vol. 317, no. 3, pp. 1114–1124, 2006.

[4] P. Garg, R. Dhakne and V. Belekar, "Role of breast cancer resistance protein (BCRP) as active efflux transporter on blood-brain barrier (BBB) permeability," *Molecular Diversity*, vol. 19, no. 1, pp. 163–172, 2015.

[5] Q. Mao and J. D. Unadkat, "Role of the breast cancer resistance protein (BCRP/ABCG2) in drug transport–An update," *The AAPS Journal*, vol. 17, no. 1, pp. 65–82, 2015.

[6] Y. Pan, P. P. Chothe and P. W. Swaan, "Identification of novel breast cancer resistance protein (BCRP) inhibitors by virtual screening," *Molecular Pharmaceutics*, vol. 10, no. 4, pp. 1236–1248, 2013.

[7] L. Zhong, C. Y. Ma, H. Zhang, L. J. Yang, H. L. Wan *et al.,* "A prediction model of substrates and non-substrates of breast cancer resistance protein (BCRP) developed by GA-CG–SVM method," *Computers in Biology and Medicine*, vol. 41, no. 11, pp. 1006–1013, 2011.

[8] E. Hazai, I. Hazai, I. R-Majlessi, S. P. Chung, Z. Bikadi *et al.,* "Predicting substrates of the human breast cancer resistance protein using a support vector machine method," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–7, 2013.

[9] K. Sasahara, M. Shibata, H. Sasabe, T. Suzuki, K. Takeuchi *et al.,* "Feature importance of machine learning prediction models shows structurally active part and important physicochemical features in drug design," *Drug Metabolism and Pharmacokinetics*, vol. 39, no. 100401, pp. 1–9, 2021.

[10] K. Sasahara, M. Shibata, H. Sasabe, T. Suzuki, K. Takeuchi *et al.,* "Predicting drug metabolism and pharmacokinetics features of in-house compounds by a hybrid machine-learning model," *Drug Metabolism and Pharmacokinetics*, vol. 39, no. 100395, pp. 1–10, 2021.

[11] M. Hassan, S. Ali, H. Alquhayz, J. Y. Kim and M. Sanaullah, "Developing liver cancer drug response prediction system using late fusion of reduced deep features," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 10, pp. 8122–8135, 2022.
[12] H. Kato, "Computational prediction of cytochrome P450 inhibition and induction," *Drug Metabolism and Pharmacokinetics*, vol. 35, no. 1, pp. 30–44, 2020.
[13] A. Rácz, D. Bajusz, R. A. Miranda-Quintana and K. Héberger, "Machine learning models for classification tasks related to drug safety," *Molecular Diversity*, vol. 25, no. 3, pp. 1409–1424, 2021.
[14] K. Ambe, K. Ishihara, T. Ochibe, K. Ohya, S. Tamura *et al.,* "In silico prediction of chemical-induced hepatocellular hypertrophy using molecular descriptors," *Toxicological Sciences: An Official Journal of the Society of Toxicology*, vol. 162, no. 2, pp. 667–675, 2018.
[15] S. J. Sammut, M. Crispin-Ortuzar, S. F. Chin, E. Provenzano, H. A. Bardwell *et al.,* "Multi-omic machine learning predictor of breast cancer therapy response," *Nature*, vol. 601, no. 7894, pp. 623–629, 2022.
[16] R. Parthasarathi and A. Dhawan, "Chapter 5–In silico approaches for predictive toxicology," In: A. Dhawan and S. Kwonpp (Eds.), *In Vitro Toxicology,* pp. 91–109, Cambridge, Massachusetts, USA: Academic Press, 2018.
[17] K. T. Rim, "In silico prediction of toxicity and its applications for chemicals at work," *Toxicology and Environmental Health Sciences*, vol. 12, no. 3, pp. 191–202, 2020.
[18] J. Cohen, "A power primer," *Psychological Bulletin*, vol. 112, no. 1, pp. 155–159, 1992.
[19] F. Faul, E. Erdfelder, A. G. Lang and A. Buchner, "G∗Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, vol. 39, no. 2, pp. 175–191, 2007.
[20] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770–778, 2016.
[21] M. Hassan, S. Ali, H. Alquhayz and K. Safdar, "Developing intelligent medical image modality classification system using deep transfer learning and LDA," *Scientific Reports*, vol. 10, no. 12868, pp. 1– 14, 2020.
[22] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter *et al.,* "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
[23] S. Ali, M. Hassan, M. Saleem and S. F. Tahir, "Deep transfer learning based hepatitis B virus diagnosis using spectroscopic images," *International Journal of Imaging Systems and Technology*, vol. 31, no. 1, pp. 94–105, 2021.
[24] S. Ali, M. Hassan, J. Y. Kim, M. I. Farid, M. Sanaullah *et al.,* "FF-PCA-LDA: Intelligent feature fusion based PCA-LDA classification system for plant leaf diseases," *Applied Sciences*, vol. 12, no. 7, pp. 1–15, 2022.
[25] A. S. Podda, R. Balia, S. Barra, S. Carta, G. Fenu *et al.,* "Fully-automated deep learning pipeline for segmentation and classification of breast ultrasound images," *Journal of Computational Science*, vol. 63, no. 101816, pp. 1–14, 2022.
[26] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. of 3rd Int. Conf. on Learning Representations*, San Diego, California, USA, pp. 1–14, 2015.
[28] L. Wu, C. Shen and A. Van Den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, no. 1, pp. 238–250, 2017.
[29] A. Amin, N. Ghouri, S. Ali, M. Ahmed, M. Saleem *et al.,* "Identification of new spectral signatures associated with dengue virus infected sera," *Journal of Raman Spectroscopy*, vol. 48, no. 5, pp. 705–710, 2017.
[30] S. Duraipandian, D. Traynor, P. Kearney, C. Martin, J. J. O'Leary *et al.,* "Raman spectroscopic detection of high-grade cervical cytology: Using morphologically normal appearing cells," *Scientific Reports*, vol. 8, no. 15048, pp. 1–8, 2018.

[31] U. Masud, F. Jeribi, A. Zeeshan, A. Tahir and M. Ali, "Highly sensitive microsensor based on absorption spectroscopy: Design considerations for optical receiver," *IEEE Access*, vol. 8, no. 2996973, pp. 100212–100225, 2020.
[32] U. Masud, F. Jeribi, M. Alhameed, F. Akram, A. Tahir *et al.,* "Two-mode biomedical sensor build-up: Characterization of optical amplifier," *Computers, Materials & Continua*, vol. 70, no. 3, pp. 5487–5489, 2021.
[33] U. Masud, M. Ali, F. Qamar, A. Zeeshan and M. Ikram, "Dual mode spectroscopic biomedical sensor: Technical considerations for the wireless testbed," *Physica Scripta*, vol. 95, no. 10, pp. 1–11, 2020.