# Improving Drug Combination Repositioning using Positive Unlabelled Learning and Ensemble Learning

Yashodha Ruchini Maralanda, Pathima Nusrath Hameed

*Abstract*— **Drug repositioning is a cost-effective and time-effective concept that enables the use of existing drugs/drug combinations for therapeutic effects. The number of drug combinations used for therapeutic effects is smaller than all possible drug combinations in the present drug databases. These databases consist of a smaller set of labelled positives and a majority of unlabelled drug combinations. Therefore, there is a need for determining both reliable positive and reliable negative samples to develop binary classification models. Since, we only have labelled positives, the unlabelled data has to be separated into positives and negatives by a reliable technique. This study proposes and demonstrates the significance of using Positive Unlabelled Learning, for determining reliable positive and negative drug combinations for drug repositioning. In the proposed approach, the dataset with known positives and unlabelled samples was clustered by a Deep Learning based Self Organizing Map. Then, an ensemble learning methodology was followed by employing three classification models. The proposed PUL model was compared with the frequently used approach that randomly selects negative drug pairs from unlabelled samples. A significant improvement of 19.15%, 20.56% and 20.23% in the Precision, Recall and F-measure, respectively, was observed for the proposed PUL-based ensemble learning approach. Moreover, 128 drug repositioning candidates were predicted by the proposed methodology. Further, we found literature-based evidence to support five drug combinations that may be able to be repositioned. These discoveries show our proposed PUL approach as a promising strategy that is applicable in drug combination prediction for repositioning.**

**Keywords**— **Drug repositioning, Positive Unlabelled Learning (PUL), Deep Learning, Self-Organizing Maps (SOM), Support Vector Machine (SVM)**

## I. Introduction

Introducing a new drug to the market is time consuming and costly. Nearly it takes seven to fifteen years to introduce a new drug to the market and approximately around $700-$1000 million cost for the whole process since it requires to undergo a massive experimental procedure before going to the hand of patients. [1]. Therefore, most of the pharmaceutical companies and medical research institutes are trying to find alternatives, which can be used to prevent and cure human diseases. As one of the most efficient and trust worthy approaches, repurposing or the reuse of existing drugs as treatments to some other diseases that still do not have proper treatments is an emerging topic from the last decade. This concept is known as drug repositioning or drug repurposing.

Moreover, drug combinational treatments are identified to be much efficient in avoiding drug resistance at treating complex diseases like cancer [2]. Since there exist approximately 16,000 [3] of approved drugs in the market, an extremely large number of drug combinations can be formed. However, only a very small number out of them are confirmed with experimental researches. Therefore, there is a need of an accurate and more predictive approach to infer useful drug combinations out of those millions of possible drug combinations, which remains yet unlabelled.

Existing drug combination repositioning approaches have followed binary classifications [4]–[6] as well as several other approaches such as tree based techniques [8] for repositioning of drug combination data. In the existing binary classification approaches, the unlabelled samples were considered as negatives [4]–[6]. Therefore, the results of existing studies might be unreliable, inaccurate and may cause to the loss of valuable and repositionable drug combinations.

In this study, a Positive Unlabelled Learning (PUL) based approach was proposed to address this problem. It uses a deep learning based unsupervised clustering approach followed by binary classification enabling us to select reliable negatives for binary classification. Unsupervised clustering method was based on a Deep Learning model using identified drug-drug similarities. The clusters with least significant known drug combinations were considered as the clusters with negatives. Our model has been compared using the frequently used binary classification approach that randomly selects samples as negatives from unlabelled data. Thereby, model predictions were evaluated and the significance of the PUL approach has been highlighted. To the best of our knowledge, this is the first attempt focusing on learning from positive unlabelled data for drug combination repositioning using drug-based features.

In Section II, an overview of existing literature under the domain, and their limitations are emphasized stating the need and the importance of our work. In Methodology and Materials Section, our dataset and our research workflow are explained in detail. Then, in the Results Section, we have illustrated our results that are relevant to the PUL-based ensemble learning methodology and the final predictions. Next, under Discussions Section, we have emphasized the significance of the proposed PUL approach, future work capabilities and literature based evidences about some of the predicted results. Finally, Section VI provides the concluding remarks of this study.

## II. Related work

Drug repositioning via in-silico methods has become popular and there exist many successful efforts in this

Correspondence: Yashodha Ruchini Maralanda (E-mail: yashodar95@gmail.com) Received: 24-08-2021
Revised:04-01-2023 Accepted: 11-01-2023
Yashodha Ruchini Maralanda, Pathima Nusrath Hameedis from Department of Computer Science, University of Ruhuna, Sri Lanka (yashodar95@gmail.com, nusrath@dcs.ruh.ac.lk ).

DOI: http://doi.org/10.4038/icter.v15i3.7232

domain. Majority of them are single drug repositioning approaches while a considerable number have focused on repositioning of drug combinations as novel therapeutics to diseases. Moreover, Machine Learning techniques such as Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, and Random Forest as well as Deep learning techniques including Deep Neural Networks, Convolutional Neural Networks and Deep Feed Forward Networks were employed.

Use of unlabelled data under binary classification, involves different methods. One common paradigm is the random selection of samples from unlabelled data as labelled negatives. Li et al. [4]'s study was for repositioning of drug combinations and it was a binary classification problem. Their dataset was composed of a majority of unlabelled data and a comparably smaller set of labelled positive samples. However, they have taken all the unlabelled samples as negative rather than selecting the plausible positives and negatives in unlabelled data and have further checked for overfitting by varying the positive and negative sample ratio from 1:12. Finally, they have chosen 1:1 as the most appropriate ratio since it has produced the best result. Even though their study was involved with an ensemble learning methodology, the reliability of the predictions might not be satisfiable because of the random selection of the negative sample.

Similarly, Chen et al. [5] has used this approach of random negative selection from a set of drug combinations, which do not have a proper labelling. They have carried out a binary classification of the selected labelled positives and randomly sampled negatives via Random Forest based on chemical interactions between drugs, protein interactions between drug targets and the target enrichment of KEGG pathways. Furthermore, map reduce programming model was used together with SVM and Naïve Bayes classifiers to identify novel drug combinations by Sun et al. [6]. Their negative dataset was composed of randomly paired drugs, which were belonging to the 103 single drugs that have been selected from DCDB [7].

TreeCombo [8] is another work, which has used a tree based approach to predict drug combinations with the use of physical and chemical properties of drugs together with gene expression levels of cell lines. Use of clinical side-effects to predict drug combinations has been tested by Huang et al. [9]. They have applied Logistic Regression and predicted drug combinations based on their clinical side effects. Here, they have categorized drug combinations as safe and unsafe by using three key side effects that were identified as more contributing towards model performance. NLLSS [10] was another approach that has integrated known synergistic drug combinations, unlabelled combinations, drug-target interactions and drug chemical structures to predict synergistic drug combinations. Moreover, they have followed a different method by involving Loewe Score [11] for drug combinations and they have classified data into principal drugs and adjuvant drugs based on a set of rules.

Kalantarmotamedi et al. [12] applied a Random Forest approach with Transcriptional Drug Repositioning in order to identify synergistic drug combinations against Malaria. Li et al. [13] has implemented PEA; an algorithm to model drug combinations using a Bayesian network which was also integrated with a similarity algorithm.

Shi et al. [14] have used Matrix Factorization to predict potential Drug-Drug Interactions (DDIs) between two drugs as well as between a set of drugs by using side effect information of known drugs. Moreover, they have introduced the ability of predicting the interaction between new drugs with another new drug that has no yet approved interactions.

Apart from machine learning, recently, Deep Learning has grabbed more interest in the domain of drug combination repositioning. Several studies have been carried out in order to predict novel drug candidate pairs. MatchMaker [15] is a Supervised Learning framework implemented based on a Deep Neural Network to predict drug synergy scores referred to as Loewe score. Chemical structures and untreated cell line gene expression profiles of drugs were utilized with three separate sub-networks where two of them are parallel executions for separate drugs in a pair and the third sub-network is for the whole drug pair.

DeepSynergy [16] for predicting anti-cancer drug synergy is based on a Feed Forward Neural Network which takes three inputs including chemical descriptors from two drugs and the genomic information of the cell line. The output from the network was the synergy score for the given input drugs. These synergy scores then decided whether the drug combination is positive or negative.

Lee et al. [17] have used Deep Feed-Forward Networks to predict DDI effects based on a set of drug features; structural similarity profiles, gene ontology term similarity profiles and target gene similarity profiles. In order to perform feature reductions, they have used Autoencoders, which was proven to have improved performances rates than Principle Component Analysis (PCA). Reduced profile pairs were then concatenated and fed to the network. RMSprop and Adam were used as optimizers with the Autoencoder and Deep Feed Forward Network respectively. Autoencoders were trained twice in order to predict DDI types more accurately. Li et al. [18] have presented a novel Convolutional Neural Network based model which is capable of predicting indications for new drugs by identifying the relevant lead compounds using the drug molecular structure information and disease symptom information. Under this, they have constructed similarity matrices out of the above two vectors of information and they were mapped into one grey scale image. Finally, this was used as the input to a Convolutional Neural Network model and that was executed using MATLAB software. Here, they have used stochastic gradient descent as an optimizer.

Peng et al. [19] has performed a prediction of drug-drug interactions using a deep learning model. They have taken true positive and true negative drug combinations from the dataset under first approach and true positive and sampled negative drug combinations in their second approach. Lee et al. [20] has involved drug pairs but it is not an approach for prediction of drug combinations as repositioning candidates. They have used Deep feed-forward networks to predict drug-drug interaction effects based on a set of drug features.

Zhang et al. [21] have implemented an ensemble model for DDI predictions. They have followed a semi-supervised learning approach because they wanted to identify unobserved DDIs, which might be available among other possible drug pairs. This is similar to identifying possible positive samples out of an unlabelled sample.

PUL has become an emerging topic since most of the natural data exist as positive and unlabelled data samples rather than having already defined positive and negative samples. There are several researches carried out under learning from PU data.

Sellamanickam et al. [22] have proposed a ranking based SVM model (RSVM) where the positive samples obtain higher scores than the unlabelled samples. A threshold parameter was estimated to form their final classifier. Liu et al [23] have followed a similar approach. They have introduced a novel computational framework for drug-drug interaction prediction with Dyadic PUL. They have identified the lack of a reliable method for separation of their unlabelled data into positives and negatives. Therefore, they have introduced a scoring function and assigned a certain score to each data pair. According to the assigned scores they have separated data into positive and negative by making the top scoring data pairs as positives while keeping the lower scoring data pairs as negatives. The top scoring data pairs were defined as the samples, which obtain a higher score than the average score of the unlabelled data pairs.

Further, Zhao et al. [24] have proposed a method for protein complex mining by employing SVM with the use of PUL. They have introduced an efficient sub graph searching method that can search complex sub graphs. First, they have tried to express the traditional training dataset with positive and negative samples as a non-traditional training set with positive and unlabelled samples. Then they have tried to identify the relationship between the two classifiers that were trained with those two types of training samples.

Even though, there are studies that have used PUL, to the best of our knowledge, no study was identified that has specifically focused on learning from positive unlabelled data for drug combination repositioning domain according to drug based features. Since, drug combination repositioning is one of the interesting and hot topics, and similarly PUL is an emerging field, we can identify the need of a PUL study with related to drug combination repositioning. The primary objective of our study is to introduce a new reliable computational method for PUL-based drug combination repositioning.

## III. MATERIALS AND METHODOLOGY

### A. Dataset

In order to demonstrate the effectiveness of the proposed approach, 183,315 drug combinations from 606 drugs that was collected from Li et al [4]'s study were used. Drug Target Similarity, Drug Indication Similarity, Drug Structure

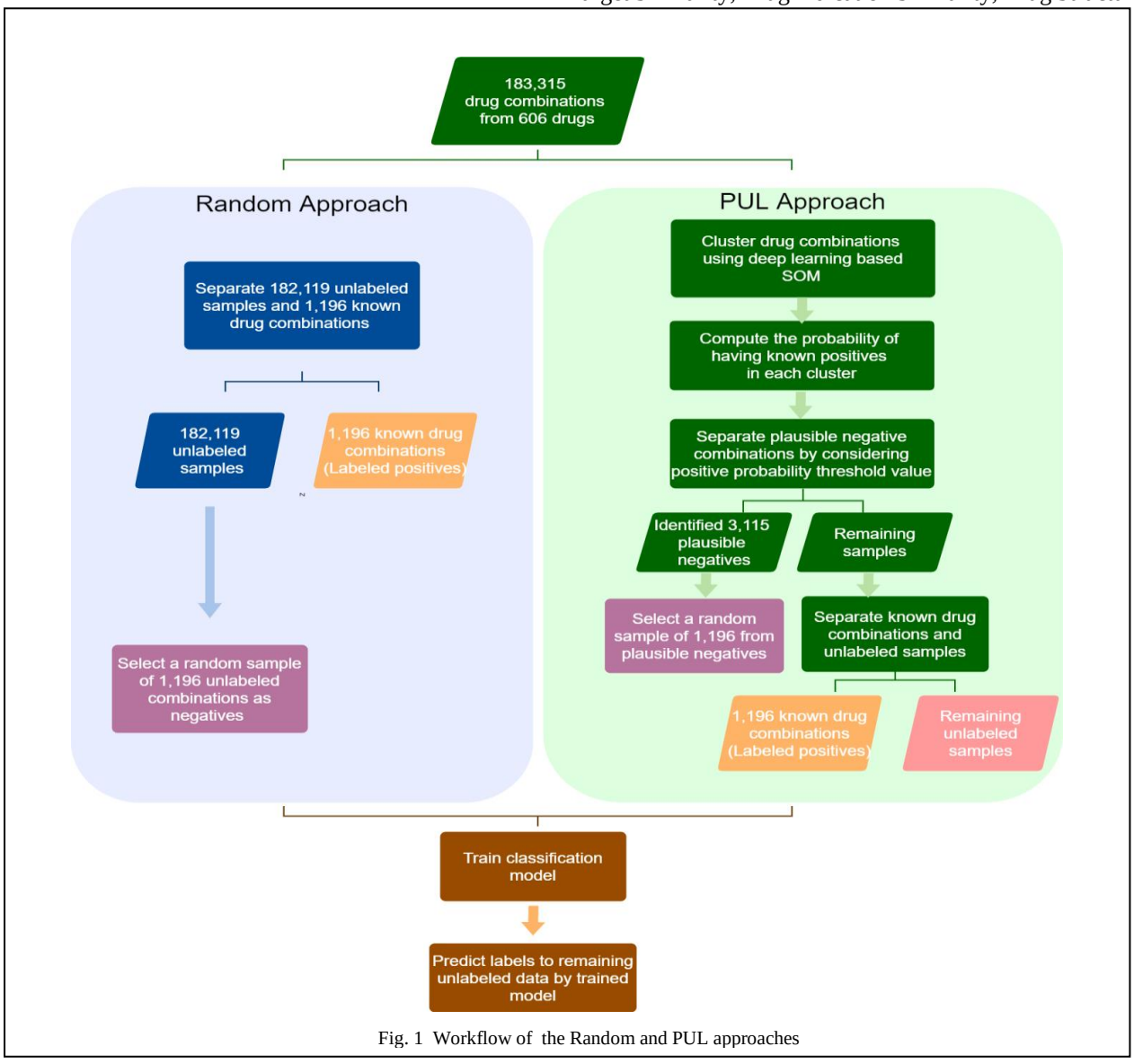


Fig. 1 Workflow of the Random and PUL approaches

Similarity, Drug Expression Similarity and Drug Module Similarity of the above drug combinations were also collected from Li et al [4]'s study. They consist of Jaccard coefficient to represent the above similarities between drug pairs. Using them, we constructed a drug combination similarity matrix with the corresponding five feature similarity scores and the file was (183,315, 5) dimensions large. Li et al.[4]'s study was composed of 1,196 labelled positive drug combinations for the 606 drugs that we are interested. After separation of labelled positives, there were 182,119 drug combinations in the unlabelled dataset (Supplementary Files S1 and S2).

### *B. Proposed Methodology*

The concept of learning from positive and unlabelled data is a setting where we have only that majority of unlabelled data and a set of already labelled positive data. Even though it is yet unlabelled, this set of unlabelled data may also contain both positive and negative samples. With this PUL technique, we are trying to identify them separately. The concept of PUL has drawn the attention of researchers due to its ability of providing reliable solutions. With the surge of this technique, it has diminished the need of having fully supervised data for computational model driven research work. With this PUL concept, it has enabled the involvement of unlabelled data for computational model driven learning processes. Many applications and research work have utilized this concept.

Unlabelled drug combinations may compose of plausible negative samples as well as repositionable drug combinations. Therefore, there is a need of a proper mechanism to identify the most probable set of negative samples to develop a reliable classification model. We have introduced a novel PUL approach for drug combination repositioning. Our proposed method enables learning from positives and unlabelled drug combinations in order to identify plausible negatives as well as plausible positives within majority of unlabelled data. We proposed PUL using a deep learning and ensemble learning methodology to predict reliable drug combinations for repositioning.

Here, we have used two approaches, which can be used to determine negative drug combinations from the unlabelled dataset. Firstly, the frequently used random selection of negatives from unlabelled data and secondly, the proposed PUL using deep learning and ensemble learning. Fig. 1 and Fig. 2 illustrates the complete workflow based on the two approaches. We demonstrate a comparison of the performance of both approaches employing Receiver Operating Curve, Precision-Recall Curve, accuracy, precision, recall and F-measure. Hence, the significance of the PUL approach for drug combination based drug repositioning is emphasized. Furthermore, we have identified a set of plausible positive drug combinations that can be repositioned for new/rare diseases. Repositioning of these predicted drug combinations need further research with laboratory experiments and other background analysis with expertise knowledge. Therefore, it needs to be carried on as a separate experiment which becomes the second phase of our research.

### *C. Random Approach*

In this approach, a randomly selected sample of unlabelled drug combinations, which is equal in size to that of the labelled positive sample was employed. Our labelled

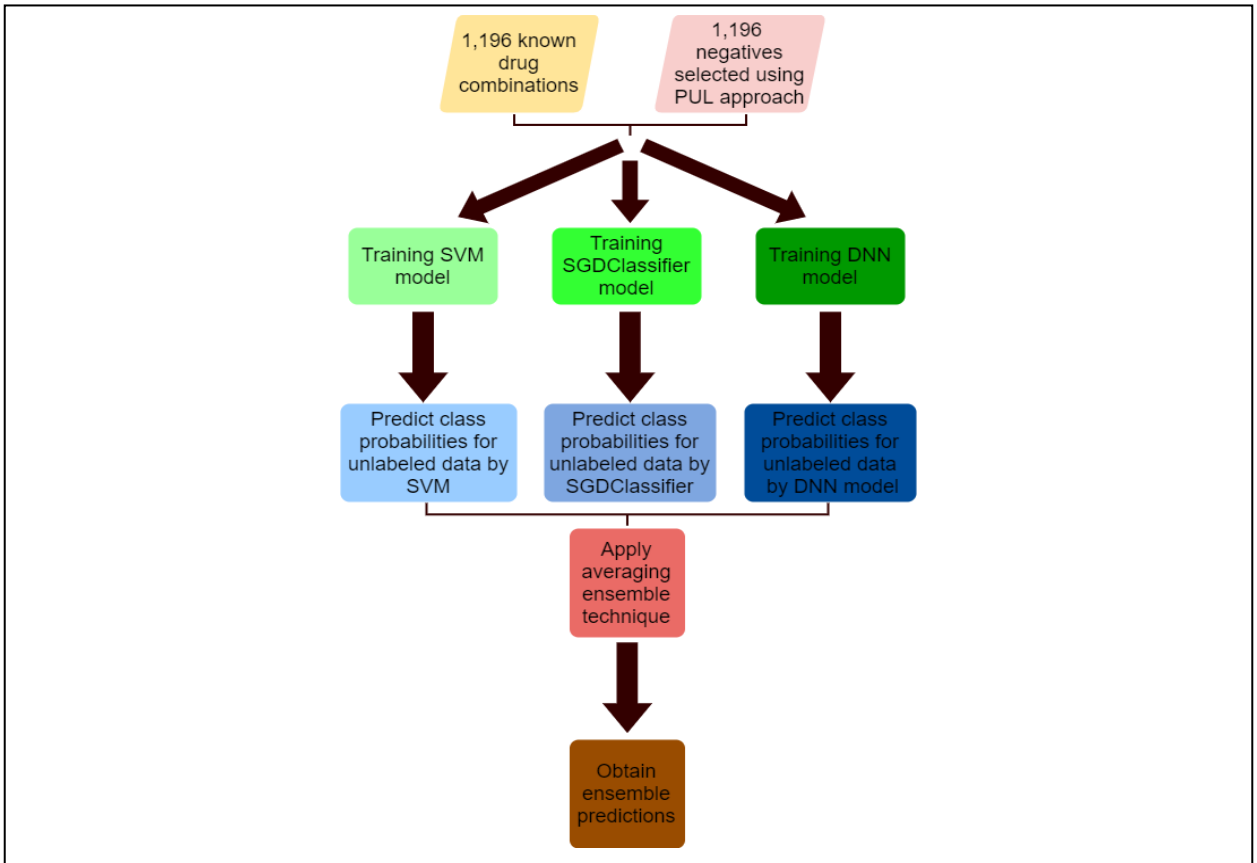


Fig. 2 Workflow of the ensemble methodology for inferring drug reositioning candidates

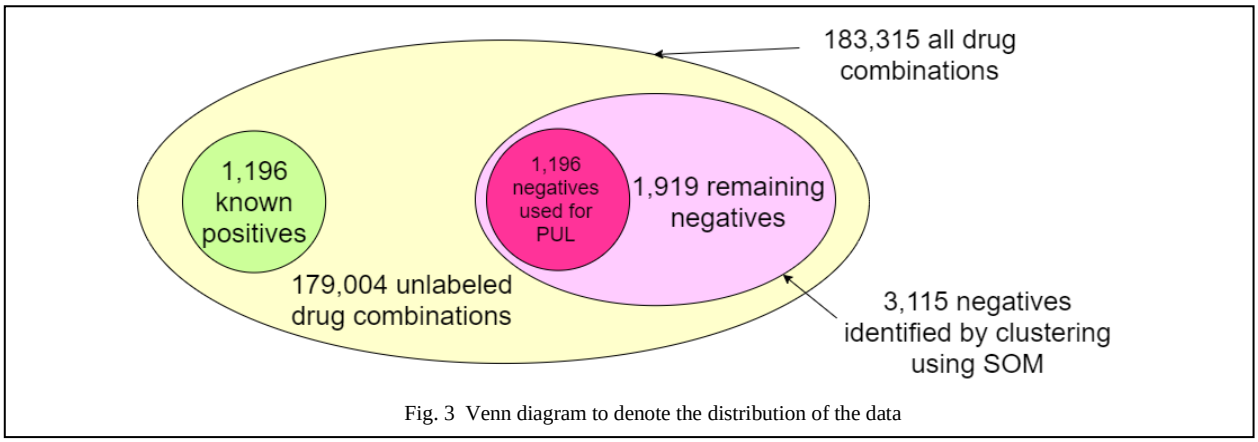


Fig. 3 Venn diagram to denote the distribution of the data

positive sample was composed of 1,196 drug combinations. Hence, we have taken a random sample of 1,196 unlabelled drug combinations as negatives. As this was a binary classification, class labels were assigned as 1 and 0, where 1 for positive and 0 for negative classes respectively. Classification was carried out using the three classifiers; SVM, Stochastic Gradient Descent-based Classifier (SGD-Classifier) and the Deep Neural Network (DNN) classifier. According to Nguyen et al. [25], we have identified that a train-test split of 70:30 is much effective with random sampling. Therefore, we decided to use the same split for both approaches. Out of the positive and negative datasets, 30% was used for model testing while the remaining 70% was taken for training the model. Implementation was carried out using python with scikit-learn library [26] for SVM and SGD-Classifier and the Keras library for the deep neural network. The accuracy, precision, recall and F1-scores were then recorded.

### *D. Positive Unlabelled Learning (PUL) Approach*

Labelled positive sample was the same as in random approach, but selection of the negative sample was carried out by learning from positive and unlabelled drug combination data. A Self-Organizing Map (SOM) was used to cluster the sample with positive and unlabelled data and then the clusters were analysed to identify plausible negative samples from unlabelled data.

For each cluster, probability of having labelled positive samples was calculated according to the *Positive Probability*. (Positive probabilities for each cluster are provided in Supplementary File S3). We defined the *Positive Probability* to be the ratio between *Known drug combinations in cluster i* and *Total number of combinations in cluster i* where *i* is the cluster ID.

Since there are 1,196 known positives, we need 1,196 reliable negatives to train the binary classifier. Therefore, we sorted each cluster based on its calculated positive probability value. The unlabelled drug pairs in the clusters with the lowest positive probability are considered as reliable negatives. Therefore, we aggregated the clusters with lower positive probability until we observe a sample size greater than or equal to 1,196. Accordingly, three clusters with the least positive probabilities were combined to get the set of least significant drug combinations. Thereby we observed 3,115 negatives by aggregating the clusters where the positive probability is less than or equal to 0.000962. Since we required balanced positive and negative samples, we randomly selected 1,196 negatives from the above-identified 3,115 negatives.

After selection of a negative sample via PUL, labelled positive and the negative sample were classified using the SVM, SGD-Classifier and the DNN model. Since, we needed to compare the performance of random and the PUL approach, we kept the model parameters fixed to the ones that were used in random approach. Similarly, 30% of data

TABLE 1
PERFORMANCE ASSESSMENT OF THE PROPOSED POSITIVE UNLABELLED LEARNING APPROACH AND RANDOM APPROACH

| | **SVM** | | **SGD-Classifier** | | **DNN Classifier** | |
|---|---|---|---|---|---|---|
| | **Random** | **PUL** | **Random** | **PUL** | **Random** | **PUL** |
| **Accuracy** | 0.6421 | 0.7925 | 0.7103 | 0.8774 | 0.7326 | 0.9721 |
| **Precision** | 0.6799 | 0.8413 | 0.8036 | 0.9564 | 0.7203 | 0.9806 |
| **Recall** | 0.5628 | 0.7454 | 0.5893 | 0.7917 | 0.7328 | 0.9646 |
| **F1 - score** | 0.6158 | 0.7904 | 0.6800 | 0.8663 | 0.7265 | 0.9725 |

taken as the testing set while remaining 70% was taken for model training. Then, accuracy, precision, recall and the F1-score given by the model were recorded.

### *E. Ensemble Learning Methodology*

Figure 2 illustrates the ensemble learning approach used in this study. In order to predict drug repositioning candidates from unlabelled drug combinations, averaging ensemble learning technique was used. First, class probabilities for the unlabelled combinations were predicted using the three individual models separately. Then the separate probabilities obtained for each drug combination to be belonged to class 0 (negative class) or class 1 (positive class) were averaged and predicted a novel probability for each drug combination. The new class probabilities were the ensemble learning based class predictions. We then predicted the best candidate drug combinations.

### *F. Clustering and Classification Models*

*1) Self-Organizing Maps (SOM):* SOM [27] is an Artificial Neural Network, which is widely used under unsupervised learning problems. The major difference of SOM with compared to other neural network models is the use of competitive learning. SOM has the capabilities of dimensionality reduction and it has the ability to identify similarities in data. It is evident that deep learning models have higher performance, with compared to machine learning approaches [28]. So, we have decided to cluster our unlabelled dataset using a minimalistic and Numpy based implementation of SOM known as MiniSom (https://github.com/JustGlowing/minisom/), which is a python library and much more adaptive with the environment where it is being used.

A two-dimensional SOM of size 9x9 was chosen as the optimal size with a learning rate of 0.09 which is trained for

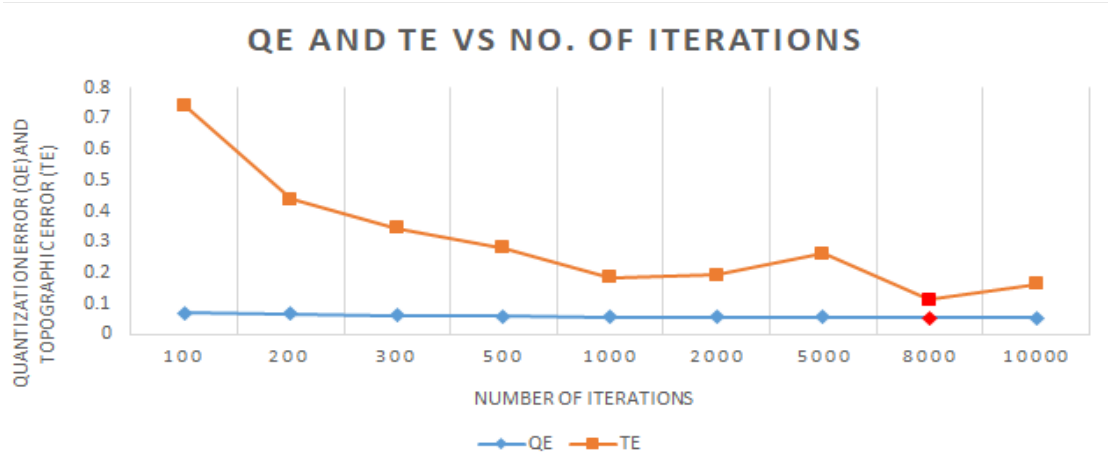


Fig. 4 Plot of Quantization Error and Topogrphic Error with a fixed learning rate of 0.5 and fixed map size of 7x7

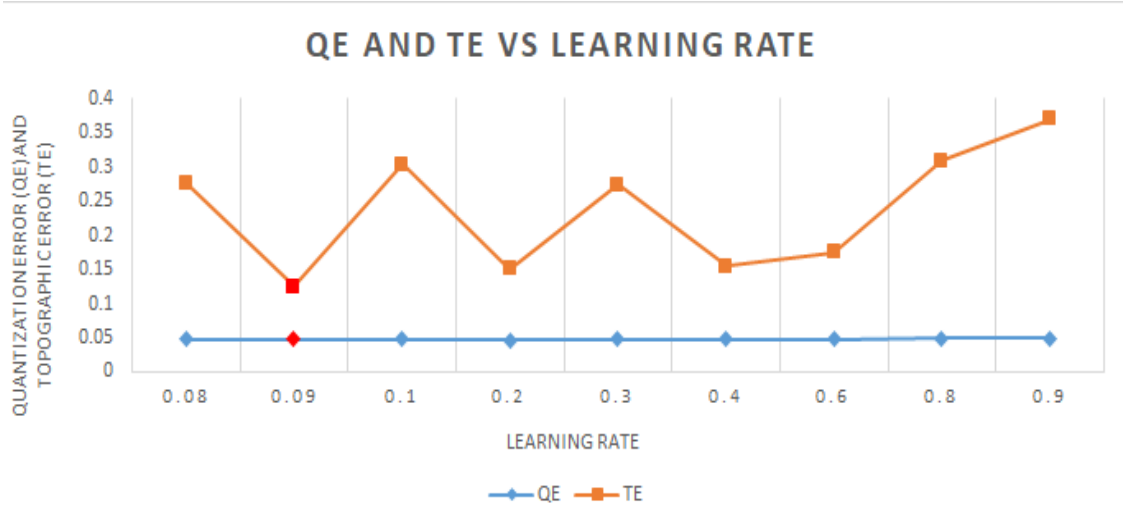


Fig. 5 Plot of Quantization Error and Topogrphic Error with a fixed map size of 9x9 and fixed number of iterations of 8000

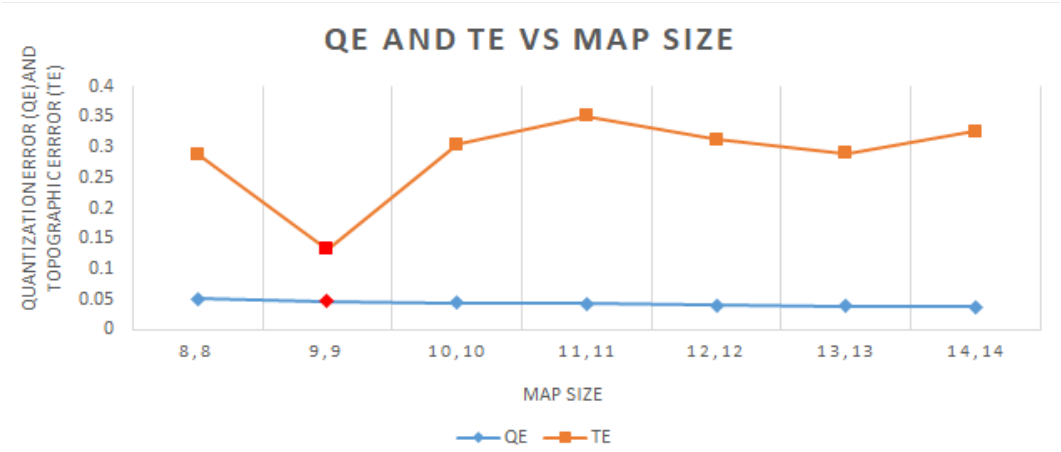


Fig. 6 Plot of Quantization Error and Topogrphic Error with a fixed learning rate of 0.5 and fixed number of iterations of 8000

8000 iterations. Selection of the optimal size; learning rate and the number of iterations were performed after calculating the quantization and topographic errors by varying their values appropriately [29], [30].

As the first step of optimal parameter identification, a set of initial parameters were needed to be determined. Hence, the learning rate of 0.5, was chosen as the initial learning rate for our model. Since a large dataset is used in this study, a considerably larger map size is required. Therefore, the map size of SOM was decided gradually increasing the dimensionality from 7x7. Hence, the initial parameters for learning rate and map size were defined as 0.5 and 7x7 respectively.

Model training was carried out multiple times with varying number of iterations, fixed learning rate and map size in order to record the Topographic Error and Quantization Error based on each experiment. Recorded error values for number of iterations that has been used in each experiment were plotted (see Fig. 4). According to the elbow technique, the experiment with 8000 iterations was chosen as the optimal value.

Since the optimal number of iterations was identified, our next experiment was followed to identify the optimal map size. We fixed the learning rate to 0.5 and number of iterations to 8000 and performed training of the model multiple times by gradually increasing the map size at each experiment. At a map size of 9x9, we could observe a clear deduction in Topographic and Quantization Error, which then again shows an increase in error values (see Fig. 5). Therefore, we determined 9x9 as the optimal map size.

After that, we used the above identified map size and the number of iterations to determine the optimal learning rate. We set the number of iterations and map size to 8000 and 9x9, respectively. The training process was performed multiple times for different learning rates. Finally, an experiment of the error values corresponding to a learning rate of 0.09 was determined as the optimal learning rate in our problem.

*2) Support Vector Machine (SVM):* SVM [31] is an algorithm, which always finds a hyperplane in an n-dimensional space where the number of dimensions is equal to the number of features used in the dataset. This can be applied for both binary classification as well as multi-class classification problems. Since this is an algorithm that has been widely used because of its higher prediction capabilities, we have decided to use it as a binary classifier in our work. The employed SVM model was followed by a sigmoid kernel, since sigmoid kernel is the most appropriate for binary classification problems.

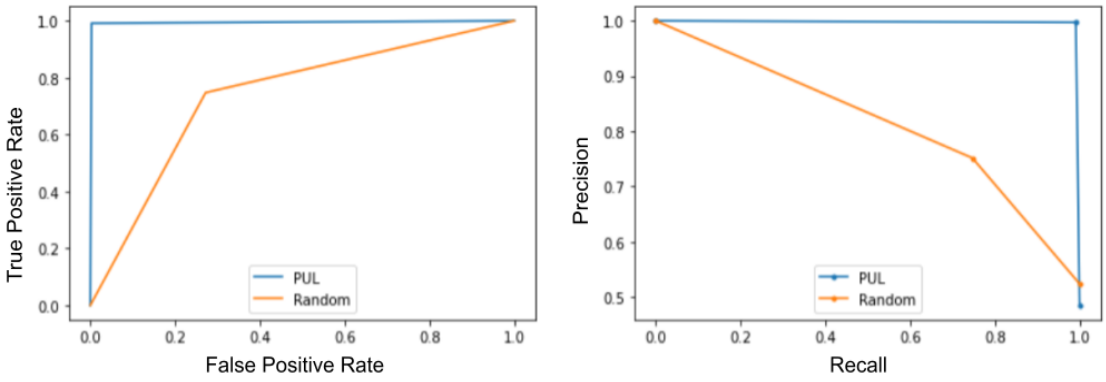


Fig. 7 Receiver Operating Curve and Precision Recall Curve demonstrating the performance of Deep Neural Network classifier for Random and PUL approache

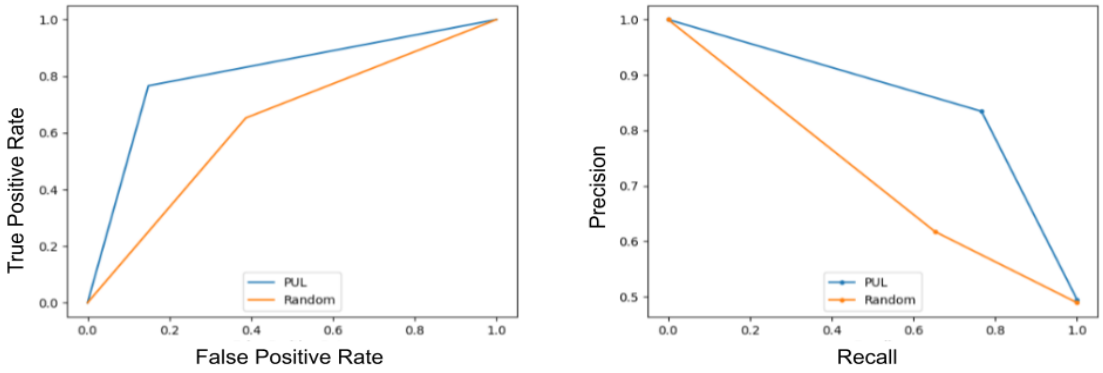


Fig. 8 Receiver Operating Curve and Precision Recall Curve demonstrating the performance of Support Vector Machine classifier for Random and PUL approaches

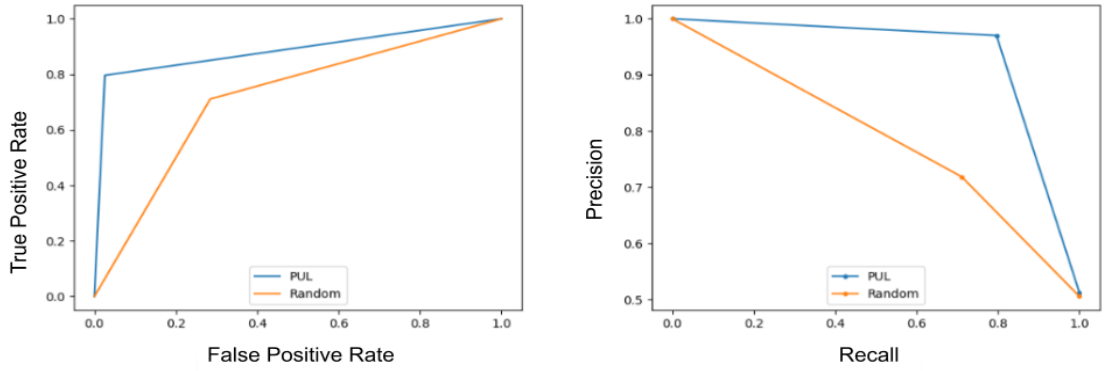


Fig. 9 Receiver Operating Curve and Precision Recall Curve demonstrating the performance of Stochastic Gradient Discent-based classifier for Random and PUL approaches

*3) Stochastic Gradient Descent based Classifier (SGD-Classifier):* This is a linear classifier that is emphasized in Scikit-learn [26], that has been optimized using Stochastic Gradient Descent (SGD). It supports loss functions and penalties that are used in classification purposes. Further, this is capable of minimizing/maximizing the loss function defined by the model. Here, we have used the log los function, and with that, our model acts similar to logistic regression (LR). However, importance of using SGD-Classifier with log loss apart from direct LR model is that, even if LR is not capable of directly calculating the minimum value of its loss function, with the use of SGD-Classifier we can easily perform it. Therefore, the performance is comparably better and so that we have used SGD-Classifier for classification purposes in our work. Even though both log loss and modified_huber loss for the loss parameter in SGD-Classifier enables to predict class probabilities for data, log loss has given the best performance in our case. Therefore, we employed a SGD-Classifier model followed by a log loss function.

*4) Deep Neural Network (DNN) Classifier:* The DNN model that was implemented using Keras library (http://github.com/keras-team/keras) was composed of a fully connected network with three layers. Since, ReLu activation function shows better performance when referring to a majority of current researches, it was used in the first two layers and sigmoid activation function was used in the output layer since this is a binary classification problem. The dimensions of the layers were selected as 5, 12, 5 and 1 for the input layer, two hidden layers and the output layer respectively such that it gives a better model for the classification of our dataset. We have set the loss parameter as binary_crossentrphy as it is specifically designed for binary classification problems in Keras. Further, we have employed the Adam optimizer as it is well suited for the instances where there are large datasets. Since our prediction dataset is large, we have involved Adam optimizer to improve the accuracy of predictions.

### *G. Evaluation Metrics*

We have divided our dataset into training and testing sets in order to validate the implemented model performances. 70% of the dataset was used for training and 30% was used for testing. Common validation measures including accuracy, precision, recall and F1-scores from the random and PUL approaches were calculated using below equations where, TP – True Positive, FP – False Positive, TN – True Negative and FN – False Negative.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$\text{F1 Score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

Furthermore, Receiver Operating Curve (ROC) is an important measure at binary classification problems, which

plots false positive rate versus true positive rate. Precision-Recall (PR) Curve provides more information by plotting the precision and recall for different thresholds. Therefore, we have observed the ROC and PR curves for our two approaches.

## IV. Results

In comparison to the random approach for negative sample selection, our proposed PUL approach demonstrates a significant improvement in the performance. (See Table 1). The accuracy, precision, recall, and F1-score for the PUL approach based on the three classifiers SVM, SGD-Classifier and the DNN classifier shows higher accuracies than the values recorded with random approach. For instance, F1-score has improved by 17.46%, 18.63% and 24.60% for SVM, SGD-Classifier and the DNN classifier respectively when the PUL approach is used.

When comparing the performance of three classifiers based on accuracy, precision, recall and F1-score, DNN classifier shows relatively higher performance for both random as well as the PUL approach. (See Table 1) SGD-Classifier shows the second-best performance while SVM has relatively lower performance with compared to the other two classifiers.

A comparison of the ROC and PR curves for random and the PUL approaches based on the three models also emphasize the higher skill of the model that was trained under the PUL approach (See Fig. 7, Fig. 8, and Fig. 9).

The ROC and PR curves are drawn in blue and orange colours for PUL, random approaches respectively. The x-axis represents false positive rate. If this rate is closer to zero, our model predicts only a few false positives. Similarly, the y-axis shows true positive rate. If this rate is closer to one, the model predicts a majority of the true positives. Therefore, an ROC curve that has bowed much towards the (0, 1) coordinate of the plot is considered to have higher skill compared to others. The blue coloured ROC plot based on each classifier has bowed towards the (0, 1) coordinate of the plot more than the orange coloured plot of random approach. Hence, the ROC curves emphasize the higher skill of the models that are trained using PUL approach.

The x-axis of PR curve represents recall. If recall gives a value that is closer to one, our model predicts only a few false negatives. Similarly, the y-axis shows precision. If precision is closer to one, the model predicts only a few false positives. Therefore, a PR curve that has bowed much towards the (1, 1) coordinate of the plot is considered to have higher skill compared to others.

TABLE 2
Performance assessment of Ensemble learning

| | **Random** | **PUL** |
|---|---|---|
| **Accuracy** | 0.6950 | 0.8807 |
| **Precision** | 0.7346 | 0.9261 |
| **Recall** | 0.6283 | 0.8339 |
| **F1 - score** | 0.6741 | 0.8764 |

The blue coloured plots of PUL approach have bowed towards (1, 1) coordinate more than the orange coloured plots of the random approach. This further emphasizes the higher skill of the models that are trained using proposed PUL approach.

A further comparison between the three ROC curves emphasize that DNN classifier gives the highest skilled model out of the three classifiers. The reason is that the ROC curve of DNN classifier is bowed the most towards the (0, 1) coordinate of the plot. The PR curve of DNN classifier is bowed the most towards (1, 1) coordinate showing the least number of false negatives and false positives. This further proves the higher skill of the DNN classifier.

We built the classifiers using SVM, SGD-Classifier and DNN and then we combined their individual predictions to obtain the final prediction. This may reduce the variance of the final outputs. Table 2 summarizes the performance assessment of the ensemble learning approach where the performance measures of the three classifiers are averaged. The evaluation metrics derived by the ensemble learning method has shown an improvement of 20.23% in the F1 – score for the PUL approach over the random approach. Hence, the proposed PUL approach outperforms the frequently used random approach and it enables predicting reliable repositioning candidates.

It should be noted that since we have identified 1,916 known positives [3] and 3,115 negatives by clustering, there are 179,004 remaining unlabelled drug combinations for predictions (See Fig. 3). We employed the proposed PUL-based three classification models as base predictors of the ensemble learning methodology to classify the unlabelled samples. Averaging ensemble learning technique was employed. Thereby we could infer 128 drug combinations with the highest posterior probabilities greater than 0.99. We infer this set of 128 drug combinations as potential candidates for drug repositioning. (See Supplementary File S4)

Furthermore, we have employed the proposed PUL approach using the three classification models to classify the 1,919 remaining negatives identified by clustering (not used to train the classification models; see Fig. 3). We assessed the predicted probabilities greater than 0.5 for class 0 (negative class) for those 1919 drug pairs. We observed 91.40%, 95.73%, and 98.59% accuracy of being predicted as a negative drug combination using SVM, SGD-Classifier, and DNN classifier, respectively. Similarly, we have observed that accuracy is 98.44% when the ensemble averaging technique is applied. Moreover, it is relatively higher than that of the SVM and SGD-classifiers. These observations confirm the accuracy of the used negatives, and on the other hand, it depicts the high accuracy of the prediction models based on the proposed PUL approach. Further, it clearly depicts the significance of the ensemble learning methodology.

## V. discussions

Most of the real world data exist as positive and unlabelled samples. It is the same in pharmaceutical domain. Several drug combination repositioning studies have used binary classification based approaches to build novel drug repositioning models. Since there exist only labelled positives and no labelled negatives, researchers use different approaches to define their own negative samples. However,

directly taking unlabelled samples as negative data might not provide accurate results since unlabelled data may contain unidentified positive samples within it. This will cause the model to provide wrong predictions. The problem of not having an exact method for identifying the most probable set of negative samples from drug combination related unlabelled data is yet not experimented. So, in this study, that gap is being addressed.

We have used balanced samples of positives and negatives for both random and PUL approaches to train the three classification models because a balanced sample ratio reduces the bias of the model predictions [4]. Since we observed a significant improvement when the PUL approach is used, it is employed to infer plausible drug combinations. We have predicted the probability of each drug combination to have a positive or a negative class label by using the averaging ensemble learning technique and thereby the label of the highest probability was assigned to the drug combination. Carrying out further experiments is essential to validate the effectiveness of the predicted 128 drug combinations (see Supplementary File S4) so that some drug combinations out of the above prediction can be experimentally proved as repositionable drug combinations.

One limitation involved with our approach is that, it only involves one clustering technique to cluster the drug combinations. Another limitation with this study is that we haven't kept any bench mark dataset to verify the model performances so that, we would have verified our results and findings. Furthermore, as a future directive, we will involve side effects associated with the drugs, so that we can filter out the drug combinations, which are free of harmful side effects and it will further improve the reliability and accuracy of the predictions. However, in the current experiment, we did not take side effects associated with the drugs into consideration.

### *A. Literature-based evidence for predicted drug combinations*

Out of the 128 predicted candidates, we found literature-based evidence to support that five drug combinations as already experimentally proven as co-administered drugs. The non-steroidal anti-inflammatory drug, Tenoxicam was experimentally identified by Moser et al. [32] as a treatment for chronic painful inflammatory conditions that occur with degenerative and extra-articular rheumatic diseases of musculo-skeletal system. This was identified to be as effective as Piroxicam. Similarly, the ratio of the compounds, Nortriptyline to Amitriptyline in the plasma of patients who were treated with Amitriptyline is identified to be useful in treating patients with depression [33]. Terazosin and Doxazosin is a drug combination that was predicted by our ensemble methodology and they have shown experimental efficacy in treatment to symptomatic benign prostatic hyperplasia in normotensive men [34]. Ofloxacin and Norfloxacin is a drug combination that is belonging to Fluoroquinolones family and able to be used as antibacterial agents. Murillo et al. [35] has tested the resolution of this drug combination as a binary mixture. Diltiazem and Betaxolol is another drug combination that has been predicted as effective in our study. Koh et al. [36] has experimentally proven that Diltiazem and Betaxolol both are effective in controlling ventricular rate in chronic atrial fibrillation when combined with digoxin.

## VI. CONCLUSION

Drug combination repositioning is an emerging research focus that gained attention of pharmaceutical and computational researchers. Moreover, computational-based approaches have showed a significant contribution for the development and improvement of drug repositioning. Since the number of known drug combinations is significantly low with compared to the number of possible drug combinations, we proposed a Positive Unlabelled Learning based ensemble learning approach to infer reliable plausible drug combinations as repositioning candidates. The ensemble learning approach enables aggregating the classification results of SVM, SGD-Classifier and DNN classification model to minimize the variance of the final predictions. Further, we have shown the applicability of proposed PUL approach in predicting drug repositioning candidates. The literature-based evidence shows the clinical significance of the proposed approach.

## REFERENCES


[1] Wouters O. J., McKee M., and Luyten J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, *JAMA - Journal of the American Medical Association*, **323**(9), 844–853

[2] DeVita V. T. & Schein, P. S. (1973). The use of drugs in combination for the treatment of cancer: rationale and results. *The New England journal of medicine*, **288**(19), 998–1006.

[3] Wishart D. et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, **46**(D1), D1074–D1082.

[4] Li J., Tong X. Y., Zhu L. D., Zhang H. Y. (2020). A Machine Learning Method for Drug Combination Prediction. *Frontiers in genetics*, **11**, 1000.

[5] Chen L., Li B. Q., Zheng M. Y., Zhang J., Feng K. Y., Cai Y. D. (2013). Prediction of effective drug combinations by chemical interaction, protein interaction and target enrichment of KEGG pathways. *BioMed research international*, **2013**, 723780.

[6] Sun Y., Xiong Y., Xu Q., Wei D. (2014). A Hadoop-based method to predict potential effective drug combination. *BioMed research international*, **2014,** 196858.

[7] Liu Y., Hu B., Fu C., Chen X. (2010). DCDB: Drug combination database, *Bioinformatics (Oxford, England)*, **26**(4), 587–588.

[8] Janizek J., Celik S., Lee S. (2018). Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine. *bioRxiv*.

[9] Huang H., Zhang P., Qu A., Sanseau, P., Yang, L. (2014). Systematic prediction of drug combinations based on clinical side-effects. *Scientific reports*, **4**.

[10] Chen X., Ren B., Chen M., Wang Q., Zhang L., Yan, G. (2016). NLLSS: Predicting Synergistic Drug Combinations Based on Semi-supervised Learning. *PLoS computational biology*, **12**(7), 1-23.

[11] LOEWE S. (1953). The problem of synergism and antagonism of combined drugs. *Arzneimittel-Forschung*, **3**(6), 285–290.

[12] KalantarMotamedi Y., Eastman R.T., Guha R., Bender A. (2018). A systematic and prospectively validated approach for identifying synergistic drug combinations against malaria. *Malaria Journal*, **17**(1), 1-15.

[13] Li P et al. (2015). Large-scale exploration and analysis of drug combinations. *Bioinformatics (Oxford, England)*, **31**(12), 2007–2016.

[14] Shi J. Y. et al. (2018). TMFUF: A triple matrix factorization-based unified framework for predicting comprehensive drug-drug interactions of new drugs. *BMC Bioinformatics*, **19** (14)

[15] Kuru H. I., Tastan O., Cicek E. (2021). MatchMaker: A Deep Learning Framework for Drug Synergy Prediction. *IEEE/ACM transactions on computational biology and bioinformatics*.

[16] Preuer K., Lewis R., Hochreiter S., Bender A., Bulusu K. C., Klambauer G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. *Bioinformatics (Oxford, England)*, **34**(9), 1538–1546.

[17] Lee G., Park C., and Ahn J. (2019). Novel deep learning model for more accurate prediction of drug-drug interaction effects. *BMC Bioinformatics*, **20** (1), 1–8.

[18] Li Z. et al. (2020). Identification of Drug-Disease Associations Using Information of Molecular Structures and Clinical Symptoms via Deep Convolutional Neural Network. *Frontiers in Chemistry*, **7**.

[19] Peng B. and Ning X. (2019). Deep learning for high-order drug-drug interaction prediction. *ACM-BCB 2019 - Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 197–206.

[20] Lee G., Park C., and Ahn J. (2019). Novel deep learning model for more accurate prediction of drug-drug interaction effects. *BMC Bioinformatics*, **20**, (1), 1–8.

[21] Zhang W., Chen Y., Liu F., Luo F., Tian G., and Li X. (2017). Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data, *BMC Bioinformatics*, **18** (1), 1–12.

[22] Sellamanickam S., Garg P., and Selvaraj S. K. (2011). A pairwise ranking based approach to learning with positive and unlabeled examples. *International Conference on Information and Knowledge Management, Proceedings*, 663–672.

[23] Liu Y. et al. (2017). Computational drug discovery with dyadic positive-unlabeled learning. *Proceedings of the 17th SIAM International Conference on Data Mining, SDM*, 45–53.

[24] Zhao J., Liang X., Wang Y., Xu Z., and Liu Y. (2016). Protein complexes prediction via positive and unlabeled learning of the PPI networks, *13th International Conference on Service Systems and Service Management, ICSSSM*.

[25] Nguyen, Q.H., Ly, H., Ho, L.S., Al-Ansari, N., Le, H.V., Tran, V.Q., Prakash, I., & Pham, B.T. (2021). Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. *Mathematical Problems in Engineering*, **2021**, 1-15.

[26] Pedregosa F. et al. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*. **12**: 2825–2830.

[27] Kohonen T. (1990). The self-organizing map. *Proceedings of the IEEE*. **78**(9), 1464-1480.

[28] Aliper A., Plis S., Artemov A., Ulloa A., Mamoshina P., Zhavoronkov A. (2016). Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular pharmaceutics*, **13**(7), 2524–2530.

[29] Kiviluoto K. (1996). Topology preservation in self-organizing maps. *Proceedings of IEEE International Conference on Neural Networks (ICNN'96)*. **1**, 294-299.

[30] Pölzlbauer G. (2004). Survey and Comparison of Quality Measures for Self-Organizing Maps. *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*, 67—82.

[31] Cortes C., Vapnik V. (1995). Support-vector networks. *Machine Learning*. **20**, 273–297.

[32] Moser, U., Waldburger, H., Schwarz, H. A., & Gobelet, C. A. (1989). A double-blind randomised multicentre study with tenoxicam, piroxicam and diclofenac sodium retard in the treatment of ambulant patients with osteoarthritis and extra-articular rheumatism. *Scandinavian Journal of Rheumatology*, **18**(S80), 71–80.

[33] Jungkunz, G., Kuß, H., & Nortriptylin-arnitriptylin-, Z. (1980). On the Relationship of Nortriptyline : Amitriptyline Ratio to Clinical Improvement of Amitriptyline Treated Depressive Patients. *Pharmakopsychiatrie, Neuro-Psychopharmakologie*. **13**, 111–116.

[34] Kaplan, S. A., Soldo, K. A., & Olsson, C. A. (1995). Terazosin and doxazosin in normotensive men with symptomatic prostatism: A pilot study to determine the effect of dosing regimen on efficacy and safety. *European Urology*. **28**(3), 223–228.

[35] Murillo J. A., Alañón M. A., Muñoz De La P. A., Durán M.I., & Jiménez G. A. (2007). Resolution of ofloxacin-ciprofloxacin and ofloxacin-norfloxacin binary mixtures by flow-injection chemiluminescence in combination with partial least squares multivariate calibration. *Journal of Fluorescence*. **17**(5), 481–491.

[36] Koh, K. K., Song, J. H., Kwon, K. S., Park, H. B., Baik, S. H., Park, Y. S., In, H. H., Moon, T. H., Park, G. S., Cho, S. K., & Kim, S. S. (1995). Comparative study of efficacy and safety of low-dose diltiazem or betaxolol in combination with digoxin to control ventricular rate in chronic atrial fibrillation: randomized crossover study. *International Journal of Cardiology*. **52**(2), 167–174.