St. Francis Xavier University

Department of Computer Science

CSCI 546: BIOMEDICAL COMPUTATION

Fall 2022

FINAL REPORT

PREDICTING THE EFFICIENCY OF ANTI-CANCER DRUG

Team Scorpion: Jayanth Kottamasu (x2020dxy), Sai Teja Reddy Palam (x2021gpy)

ABSTRACT:

According to World Health Organization (WHO) [1] cancer is the leading cause of death around the world. In 2020, cancer will be responsible for nearly 10 million deaths. Most people suffer from breast and lung cancers, mainly caused by tobacco intake, high body mass index, and alcohol intake [2]. Many scientists and researchers around the world are working to find a cure. Still, the issue is precision medicine, which uses data about a person's genes or proteins to treat, diagnose, or prevent disease [3]. Consequently, we must foresee the effectiveness of anti-cancer drug activity. In this article, we present a graph in the form of a Structured Data File (SDF) that depicts the compound's chemical structure. Each data sample details the atoms and their interactions within the molecule. In this case, the atoms and connections are the features. Nodes and edges describe the atoms and their connections, respectively. The previous methods have predicted the efficiency of the drugs using Manifold Learning, Machine Learning Clustering Techniques, and based on gene expression inference of cancer drug sensitivity [9]. The proposed model is based on the Graph Neural Network, which calculates the probability of the output class. The message-passing methods can be implemented distinctly using different methods. Overall, we conclude that RGCN, as the message-passing class for the balanced dataset, gave us the best accuracy of 81%.

KEYWORDS:

World Health Organization, Cancer, Structured Data File (SDF), Graph Neural Network

INTRODUCTION:

According to estimates, cancer kills around one in four Canadians [20]. About 65% of cancer patients survived the disease in the past five years. Even though the survival rates have improved, population growth also leads to more cases being detected yearly. Relapsing Cancer is one of the factors contributing to the rise in cases. Recurrence or recurrent cancer [21] are phrases used in medicine to describe the disease's occurring again in a patient. Some cancer cells become resistant to the treatment while being treated for their cancer. These cancer cells are resistant to treatment; therefore, it might take weeks, months, or even years to multiply. One of the biggest obstacles to treating cancer is the condition of recurrence.

There are numerous cancer treatment options, with medication therapy being one of the most successful. Additionally, there are three types of medication therapy: combination drug therapy, tailored monotherapy, and one-size-fit [4]. In a one-size-fits-all treatment method, patients with a similar type of cancer receive the same drug regimen. With personalized monotherapy, the patient receives a prescription for a targeted medicine based on the type of cancer mutations present. Because of the variety of tumors, people with the same type of cancer may react differently to the same drugs. Due to the ineffective anti-cancer medications, some of the usual side effects include fatigue, mouth discomfort, nausea, vomiting, lack of appetite, and hair loss. As a result, it is crucial to develop computational approaches that can help researchers understand how genomic information and medication sensitivity relate to one another.

MATERIALS:

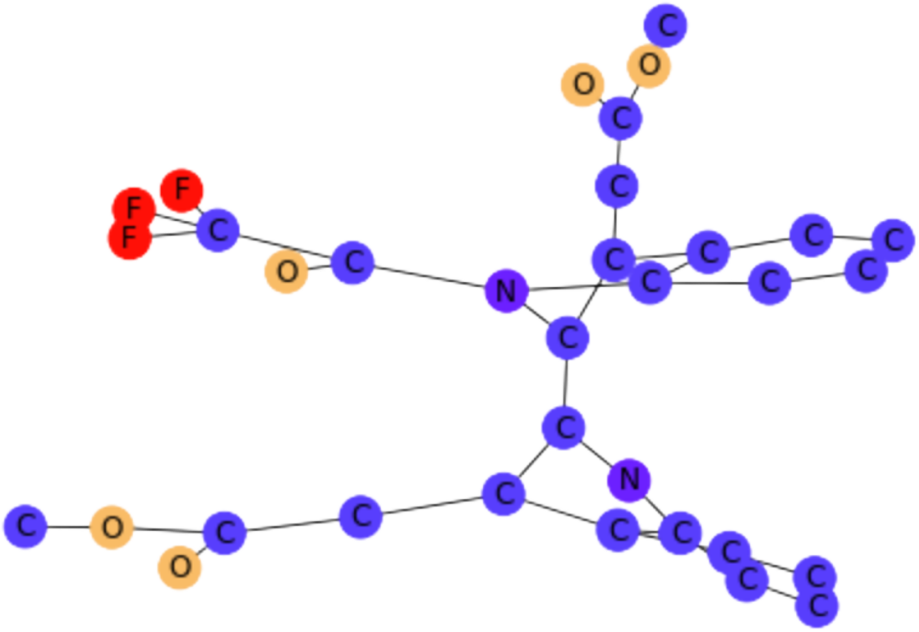
The information is presented as a graph that reveals the drug's chemical composition. The atoms and their interactions inside each data sample are described in detail. As a result, in this scenario, the atoms and connections represent the traits of the features. The project's input file is the Structured Data File (SDF) [5].

```
(['O', 'O', 'N', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C'], array([[ 0,  4],
       [ 0,  5],
       [ 1,  4],
       [ 1, 15],
       [ 2,  3],
       [ 2,  5],
       [ 2,  7],
       [ 3,  4],
       [ 3,  6],
       [ 5,  8],
       [ 6,  9],
       [ 6, 10],
       [ 7, 11],
       [ 8, 12],
       [ 9, 13],
       [10, 14],
       [11, 12],
       [13, 16],
       [14, 16],
       [15, 17]]), 0)
```

It provides information about the molecule's chemical composition. The relationships between individual atoms and their locations within chemical compounds are recorded in the SDF file. Each sample or molecule starts with a header that details the chemical's name or title. In this

project, information about the molecule will be gathered and stored as edges and nodes using atoms and their relationships as building blocks.

After viewing and examining a sample, the training set provides data for each molecule. Three elements are present in each sample array. The first part contains textual information about the atoms, the second element contains data about connections, and the third element describes the label for each molecule.



The idea of “personalized medicine” states that different treatment options are given for specific individuals based on various features that can be learned. The absence of precedent examples, such as patient gene expression data and well-known treatment procedures' confirmed efficacy is a challenge for many machine learning algorithms. To assess the efficiency of various anti-cancer drugs in slowing down cell proliferation, thousands of different cell lines have been treated, and the gene expression profiles of all of these cell line cultures have been analyzed. The clinical effectiveness of anti-cancer drugs for specific patients can be predicted using a cutting-edge machine-learning technique that transfers attributes from the cell lines based on expression data [6].

METHODS:

By looking at different research studies, we learned that lung cancer cell lines could be put into groups using simple K-means, Filtered Clustering, and figuring out which drugs work best on each cell line. The K-means clustering technique can determine which lung cancer cell lines are sensitive to a given concentration of medication. These chemicals come from natural sources, like apple leaves, pepper, pepper fruits, sheep intestine, sea sponge, and pepper. Anti-drug chemicals like Neopeltolide, Parbendazole, Phloretin, and Piperlongumine made all 91 cell lines work better at different concentrations (p-value 0.001) [7]. This examination of the published experimental findings showed that certain chemicals are more sensitive than others.

ADRML [3], a model for predicting the effectiveness of anticancer drugs, integrates cell line data with manifold learning and drug knowledge to develop precise good drugs. This model shows how to put the drug response matrix into the lower-rank spaces, which gives new information about the drugs and the cell lines. Using the low-rank method [9], the drug response for a novel cell line-drug pair is calculated. The analysis of ADRML's efficiency on numerous cell types and drug data is featured. Additionally, the comparisons with earlier suggested techniques demonstrate that ADRML offers precise and solid forecasts. More investigation into the link between drug response and pathway activity ratings is needed to give light on the underlying pharmacological process. The results suggest that ADRML can accurately predict and impute anticancer therapy response.

In the proposed method, SDF files keep track of links between atoms and information about their positions inside chemical compounds. The delimiter $$$$ distinguishes several compounds. Each sample or molecule begins with a header that provides information about the compound's name or title. Information on the Atom count, the version number, connections, etc., can be found in other sections. The atom block describes the elements of the compound. The bond block describes the bonding structure of the chemical. These two blocks are used in this method to collect information about the compound and store it as edges and nodes [8]. The labels are assigned as 1 and 0 if present as 1.0 and -1.0 in the dataset after collecting the information and storing the data in the form of edges and nodes. We will split the training set into training and validation sets. We will perform the train_test_split method again on the train data to split it into

training and testing data sets. Here training set and validation set are used to train the dataset, whereas the testing set is used to set the model to know how well the model is performing.

The nodes of the chemical molecule are present in the data in tokenized form. Each compound's nodes are retrieved and tokenized with the help of a tokenizer. Tokenization in Python generally means breaking up a longer text into smaller lines or words or making up terms for languages other than English. Then padded, the dataset was using the pad sequence technique. Each batch has a shape [batch_size, max_len_nodes], where batch_size is the total number of samples, and max_len_nodes is the length of tokenized nodes following padding. Edge is the input tensor that contains data about atom connections. An edge has the following shape: [sum_of_all_edges, 2]. The sum_of_all_edges represents the number of edges in each sample of the batch size. For instance, the edge tensor's size would be [81, 2] in a batch of three samples where sample 1 has 21, sample 2 has 20, and sample 3 has 40 edges.

The input tensor node2graph [8] provides details about segmented ids used for segmented mean. The gnn_out has the following dimensions: [batch_size_node_dimension, hidden_layer], where batch_size_node_dimension is the dimension of the input data. It displays the model's aggregate output for each hidden layer.

The final result indicates whether or not a chemical component is likely to be active for cancer cells. As a result, the final output for each sample is a number that reflects the chance that each chemical molecule will be active. There are 2 message-passing methods used, which are RGCN (Relational graph convolution layers) [14] and RGAT (Relational graph attention networks) [15]. Using multiple GNNs, the graph's complexity can be considered, and a better model can be made.

RESULTS:

UNBALANCED DATA:

Specialized neural network types, such as “graph neural networks,” operate on a graph data structure. Predictive tasks involving nodes, edges, and graphs are performed with GNNs. Here, we use all the default parameters by get_default_hyperparameters()[19].

For Model 1, we used the message passing method as RGAT with all the default hyperparameters and made some changes, such as epochs where it is set to 40, the batch size is set to 32, and hidden layers are set to 64. Here we used two accuracy methods: accuracy score and balanced accuracy score from sklearn. We got a result of 94% accuracy and 54% balanced accuracy for this model.

For Model 2, we used the message passing method as RGCN with all the default hyperparameters and made some changes, such as epochs where it is set to 40, the batch size is set to 32, and hidden layers are set to 64. Here we used two accuracy methods: accuracy score and balanced accuracy score from sklearn. We got a result of 94% accuracy and 54% balanced accuracy for this model. The table below will clearly describe how the parameters are varied.

Parameters	**Model-1**	**Model-2**
message_calculation_class	RGAT	RGCN
epochs	40	40
batch size	32	32
hidden layers	32	32

BALANCED DATA:

For balancing the data, we have used the resample method to balance the minority class. In general, a resampling approach involves repeatedly selecting samples from a dataset and analyzing the statistics and metrics of each sample to learn more about something; in the context of machine learning, this item is the performance of a model. The table below is driven from the code, which shows the difference between the classes before and after upsampling[16].

```
Shape of Unbalanced Data before Upsampling:
DATA WITH LABEL 0: (14550, 3)
DATA WITH LABEL 1: (714, 3)


New class counts after Upsampling:
0    14550
1    14550
Name: Labels, dtype: int64
```

For Model 1 on balanced data, we used the message passing method as RGAT with all the default hyperparameters and made some changes, such as epochs set to 40, batch size set to 32, and hidden layers set to 64. Here we used two accuracy methods: accuracy score and balanced accuracy score from sklearn. We got a result of 54% accuracy and 54% balanced accuracy for this model.

For Model 2 on balanced, we used the message passing method as RGCN with all the default hyperparameters and made some changes, such as epochs set to 25, the batch size set to 32, and hidden layers set to 64. Here we used two accuracy methods: accuracy score [17] and balanced accuracy score [18] from sklearn. We got a result of 81% and 75% balanced accuracy for this model. The table below will clearly describe how the parameters are varied.

Parameters	Model-1	Model-2
message_calculation_class	RGAT	RGCN
epochs	40	20
batch size	32	32
hidden layers	32	64

CONCLUSION:

To sum up, we have reviewed earlier research on the effectiveness of anti-cancer drugs. Then, using scores and accuracy rates, we intend to deploy Graph Convolution Neural Network to predict the most active anti-cancer medications. For better outcomes, we also intend to concentrate on hyperparameter optimization. Consequently, the primary objective of this report is to develop a model with significantly specific outcomes and lower time complexity. We ended up with RGCN as the message passing class gave us the best accuracy of 81% and the best-balanced accuracy of 75%. For this model, we used a balanced accuracy score as it works well on the imbalanced dataset using the average of recall obtained on each class [18].

MODEL	Un-Balanced Dataset		Balanced Dataset	
	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy
RGAT	0.95	0.54	0.54	0.54
RGCN	0.95	0.54	0.81	0.75

GitHub REPOSITORY:

https://github.com/palamsaiteja333/PREDICTING-THE-EFFICIENCY-OF-ANTI-CANCER-DRUG

REFERENCES:

1. World Health Organization (WHO). https://www.who.int/health-topics/cancer#tab=tab_1
2. Anand, Preetha, Ajaikumar B. Kunnumakara, Chitra Sundaram, Kuzhuvelil B. Harikumar, Sheeja T. Tharakan, Oiki S. Lai, Bokyung Sung, and Bharat B. Aggarwal. "Cancer is a preventable disease that requires major lifestyle changes." Pharmaceutical research 25, no. 9 (2008): 2097-2116.
3. Ahmadi Moughari, Fatemeh, and Changiz Eslahchi. "ADRML: anticancer drug response prediction using manifold learning." *Scientific reports* 10, no. 1 (2020): 1-18.
4. An Intelligent Approach to Predict Drug Combination towards More Effective Treatment of Cancer by Eshitha Reddy Chitla | Supervised by Othman Soufan.
5. GitHub link for the Structured Data File (SDF) https://github.com/hendmuhmd99/Anti-Cancer-Drug-Activity-Prediction
6. Borisov, Nicolas, Victor Tkachev, Maria Suntsova, Olga Kovalchuk, Alex Zhavoronkov, Ilya Muchnik, and Anton Buzdin. "A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency." *Cell Cycle* 17, no. 4 (2018): 486-491.
7. Wanigasooriya, Chandi S., Malka N. Halgamuge, and Azeem Mohammad. "The analysis of anticancer drug sensitivity of lung cancer cell lines by using machine learning clustering techniques." *International Journal of Advanced Computer Science and Applications* 8, no. 9 (2017).
8. Anti-Cancer_Drug_Activity_Prediction. https://github.com/ahmedsallem/Anti-Cancer_Drug_Activity_Prediction
9. Chawla, Smriti, Anja Rockstroh, Melanie Lehman, Ellca Ratther, Atishay Jain, Anuneet Anand, Apoorva Gupta et al. "Gene expression based inference of cancer drug sensitivity." *Nature communications* 13, no. 1 (2022): 1-15.
10. Wang, Lin, Xiaozhong Li, Louxin Zhang, and Qiang Gao. "Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization." BMC cancer 17, no. 1 (2017): 1-12.

11. Perumal, Prasath Palanisamy1 P., K. Thangavel, and R. Manavalan. "Informative Gene Selection for Leukemia Cancer Using Weighted K-Means Clustering."
12. https://docs.python.org/3/library/tokenize.html
13. https://stackoverflow.com/questions/42943291/what-does-keras-io-preprocessing-sequence-pad-sequences-do#:~:text=pad_sequences%20is%20used%20to%20ensure,length%20as%20the%20longest%20sequence.
14. Thanapalasingam, Thiviyan, Lucas van Berkel, Peter Bloem, and Paul Groth. "Relational graph convolutional networks: a closer look." PeerJ Computer Science 8 (2022): e1073.
15. Busbridge, D., Sherburn, D., Cavallo, P., & Hammerla, N. Y. (2019). Relational Graph Attention Networks. *arXiv*. https://doi.org/10.48550/arXiv.1904.05811
16. https://en.wikipedia.org/wiki/Resampling_(statistics)
17. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
18. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html
19. Zhou, Jie, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. "Graph neural networks: A review of methods and applications." *AI Open* 1 (2020): 57-81.
20. https://cancer.ca/en/research/cancer-statistics/cancer-statistics-at-a-glance#:~:text=About%201%20out%20of%204,expected%20to%20die%20from%20cancer.
21. https://www.cancer.gov/types/recurrent-cancer#:~:text=Local%20recurrence%20means%20that%20the,far%20from%20the%20original%20cancer.