Machine Learning approach for Chemotherapy Suitability Prediction using Genomic Data

Krutik Bajariya
K.J. Somaiya Institute of Engineering and Information Technology Sion, Mumbai
University of Mumbai, India
krutik.bajriya@somaiya.edu

Yash Deshpande
K.J. Somaiya Institute of Engineering and Information Technology Sion, Mumbai
University of Mumbai, India
deshpande.y@somaiya.edu

Pruthil Gandhi
K.J. Somaiya Institute of Engineering and Information Technology Sion, Mumbai
University of Mumbai, India
pruthil.g@somaiya.edu

Reena Lokare
K.J. Somaiya Institute of Engineering and Information Technology Sion, Mumbai
University of Mumbai, India
reena.l@somaiya.edu

Abstract*—Chemotherapy uses anti-cancer drugs that may be injected or asked to consume to kill cancer cells. For Breast Cancer, chemotherapy may be given before or after surgery. Many of the times, chemotherapy can over treat a small tumour or be less effective in certain cases. Also Chemotherapy can cause harmful side effects that can even be life threatening. A proper system to predict whether chemotherapy will be suitable to a breast cancer patient or not is required. For that gene expression data of the patients is utilised. Several Machine Learning models can be used for classifying the treatment suitable and unsuitable for patients. A system which takes certain gene expressions of patients and tells the suitability status of chemotherapy is implemented. This system aims to help doctors and researchers improve the overall cancer treatment scenario.

***Keywords*—Chemotherapy, Genomic Data, Breast Cancer, Support Vector Machine, Random Forest, Decision Tree, Explainable AI**

I. INTRODUCTION

Chemotherapy for Breast Cancer carries a risk of side effects, some of them are temporary and mild and some of them are dangerous and permanent. Breast Cancer is caused by certain genetic mutations in women. Majorly changes in the genes BRCA1 and BRCA2 are highly responsible for developing this tumour. Modern Technology is changing the way treatment is carried out. Earlier, only clinical features were taken into consideration while deciding a treatment. Clinical and Pathological analysis of breast cancer tissue and axillary lymph nodes were done. Further, the type of breast tumour was also taken into account. But this method was not enough for successful treatment and cure. New methods are needed for this optimization of breast cancer treatment. Gene expression profiling is one method which researchers and doctors have been studying for a long time. It is a genetic microarray analysis of genetic transcriptional variations between normal and malignant cells. Various gene expression datasets are available open source. Supervised learning algorithms like Support Vector Machine, Decision Tree, Random Forest are used to classify the treatment suitability in patients. This classification is done on the basis, whether there is pathological complete response or residual disease.

Pathological Complete Response is defined as the disappearance of all invasive cancer in the breast after neoadjuvant chemotherapy. Residual disease means that cancer cells are still present even after attempts to remove them.

Gene Expression Profiling is the measure of activity of thousands of genes at a time to make a universal picture of cellular obligation. These profiles can help distinguish between the cells that are actively dividing or show how the cells react to a particular treatment. There are gene profiling tests that help to determine the right treatment for the right person.

A microarray database is a repository containing microarray gene expression data. It is a new line of research in machine learning. This type of data is employed to gather information from tissue and cell samples relating to genetic phenomenon variations that may be helpful for malady diagnosing or distinctive variety of tumor. Microarray, a tool that helps in recognizing numerous gene expressions at once. The microscopic slides that have a number of little spots written in specific positions are claimed to be DNA microarrays. Each spot within the microscopic slides is referred to as DNA Sequence. The DNA molecules on such slides act as probes that facilitate in detecting gene expression. These molecules also are referred to as transcriptome or RNA transcripts.

Researchers are finding ways to analyze these complex datasets. Various platforms like National Center for Biotechnology Information [15], cBioPortal for Cancer Genomics [16], etc. are providing tools to visualize and analyze the gene expressions. This makes work easier for finding differentially expressed or mutated genes for a learning like semi-supervised learning, unsupervised learning and supervised learning. Supervised learning as the name indicates, has a presence of a supervisor or teacher. In this type, the dataset already has correct output

labels. In Unsupervised learning there is no label or correct output present in the data. Semi Supervised learning consists of data which has both labelled outputs and unlabelled values. For classification, Support Vector Machine (SVM), Decision Tree, Random Forest, Naive Bayes, k-Nearest Neighbours (kNN), Logistic Regression, Linear Regression, AdaBoost are some of the frequently used algorithms. In this study some of the above mentioned algorithms were used.

II. LITERATURE SURVEY

Yu-HongQu implemented Prediction of Pathological Complete Response after Neoadjuvant Chemotherapy for breast cancer using ensemble machine learning. Deux Machine Learning Framework was used in this study which focuses on multi criteria decision-making technique known as weighted simple additive weighting (WSAW) instead of accuracy [1].

Gabriel A Brooks et al. proposed A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy. A total of 146 (9.2%) of 1579 patients from the parent cohort experienced hospitalization related to chemotherapy. Age, charlson comorbidity index (CCI), creatinine test, calcium level in blood, less than normal white blood cell and/or platelet count are some significant variables associated [2].

Ravi Bharat Parikh et al. implemented A machine learning approach to predict short-term mortality risk in patients starting chemotherapy treatment. The most common cancers were breast (23.6%), colorectal (17.6%), and lung (16.6%). 18.4% of patients died within 180 days after chemotherapy initiation. Electronic Health records data from 2004-14 was used [3].

Cai Huang et al. proposed that Machine learning predicts individual cancer patient responses to chemotherapeutic drugs with high accuracy. Here they predicted 175 cancer patient's response to a variety of chemotherapy related drugs from gene expression profiles. The accuracy of models were found to be above 80% [4].

Yu-Chiao Chiu et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. In this analysis, they trained as well as tested the model on a dataset of 622 malignancy cell lines and achieved an overall prediction performance of mean square error at 1.96 (log-scale IC50 values). The performance was superior in predicting, fallacy or steadiness than other standard methods like linear regression and support vector machines [5].

Eliseos J. Mucaki et al. implemented Predicting responses to platin chemotherapy agents with biochemically-inspired machine learning. Data from Cancer Genome Atlas[TCGA] was used here. Bladder, colorectal, ovarian cancer patients were utilized to check the cisplatin, carboplatin, and oxaliplatin signatures. Accuracy was found to be 71.0% ,54.5%, 60.2% respectively in prediction of disease recurrence. Also accuracy was found to be 59%, 72%, 61% respectively in disease remission [6].

Breast Cancer Prognosis Using a Machine Learning Approach. ML-based decision support system (DSS), combined with random optimization (RO), to extricate prognostic data from habitually collected demographic, clinical and physiological information of breast cancer patients. A DSS model was developed within a training set (n = 318). Its performance analysis within the testing set (n = 136) resulted in an accuracy of 86% [7].

Azuaje F. et al. have developed a model which tells effective therapy to the patients by analyzing their tumour characteristics. [8].

Eugene Lin, Po-Hsiu Kuo's Deep Learning Approach for Predicting Antidepressant Response in patients suffering from depression.. They used data of 455 patients who were treated with selective serotonin reuptake inhibitors (treatment-response rate = 61.0%; remission rate = 33.0%) [9].

Yukun Chen et al. implemented Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations. They explored the patterns of 1,760,846 corporal mutations identified from 230,255 cancer patients along with gene function data employing a support vector machine. Using only gene features, the accuracy of the baseline was 57%. When information of mutation and chromosome was added accuracy was improved to 62%. [10].

Siew‑Kee Low et al. Breast cancer: The translation of big genomic data to cancer precision medicine. This study encapsulates the identification of genomic alterations with high outturn screening and use of genomic info for clinical trials [11].

Ashraf Abou Tabl et al. implemented Machine Learning Approach for identifying Gene Biomarkers guiding the treatment of Breast cancer. This study focuses on patients who survived breast cancer and taking their gene expressions as a clue for finding the right treatment for future patients [12].

Diaddin Hamdan et al,. Genomics applied to the treatment of breast cancer. This study focuses on genomics and its contribution to the treatment of breast cancer. It also considers the future applications linked to the data [13].

Byung-Ju Kim et al. Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. This study finds genomic susceptibility of patients to 20 common cancer types. [14].

III. COMPARISON WITH EXISTING SYSTEM

In the current system, machine learning is used for predicting outcome in chemotherapy and radiotherapy. The outcome is both for survival after treatment and whether the treatment will be suitable or not. The datasets used here are clinical and dosimetric. Dosimetric data

refers to the amount of radiation dosage given to treat a patient. Random Forest, Support Vector Machine, LogitBoost, Neural Network, Decision Tree algorithms were used for training the model. The limitation of their study is absence of genomic input features. In the proposed system, genomic features are used which is helping in giving better results.

IV. METHODOLOGY

The main objective of the system is accurate prediction of chemotherapy suitability. Gene Expression data is used for the same. Suitable treatment if predicted in advance will reduce a lot of trouble for patients as well as doctors. If wrong treatment is started it will lead to a financial as well as a health loss of patients. So predicting optimal treatment in advance is vital. The block diagram is shown below in Fig 1.

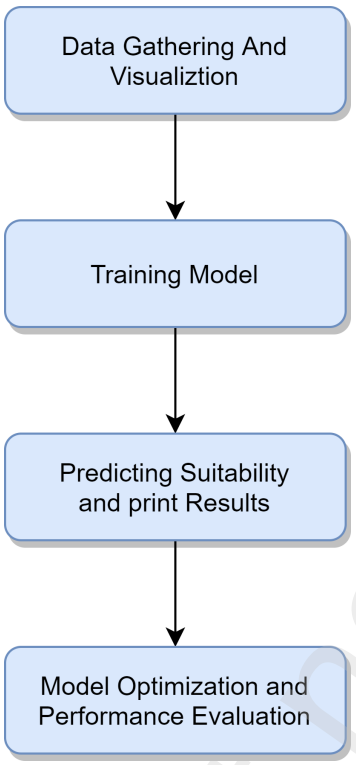

Fig. 1 - Block Diagram

First the Data Gathering is done from various open source platforms. Then the model is trained which provides certain results. Based on that Model is Optimized and then deployed for further use.

A. *Datasets Used*

Microarray Gene Expression Data was taken from National Center for Biotechnology Information (NCBI) [15]. It was available open source to all the users. It was extracted in series matrix format. The dataset had more than 20,000 Genes (Probe Id) as features. The dataset included 510 breast cancer patients out of which 101 patients had Pathological Complete Response and rest had Residual Disease.

B. *Data Preprocessing and Feature Selection*

The dataset had some samples which did not have any labels. So those samples were removed. If all 20,000 genes are selected, then overfitting will occur and proper results will never be achieved. So, the genes that are mutated on giving chemotherapy were selected. 20 such genes which were extracted are TP53, TP53TG1, NFIB, FEN1, TTK, MELK, TPX2, METRN, H2AFZ, MCM6, ESR1, TOPBP1, E2F3, CYP2B7P, MLPH, SLC7A8, RARRES1, ROPN1B, PLCH1, DEK [15]. These genes helped in better prediction of results.

C. *Model Building*

After all the data cleaning and feature selection, and testing the dataset on various tools like Weka, model building was done on Google Colab using Python 3. Model was trained using Support Vector Machine, Random Forest, Decision Tree, AdaBoost and Logistic Regression. Models were tested on a test dataset. Also various plots were plotted to visualize the results obtained. Model was optimized by tuning the parameters, using different algorithms and ensembling the model. The flow of the model building process is shown in Fig 2.

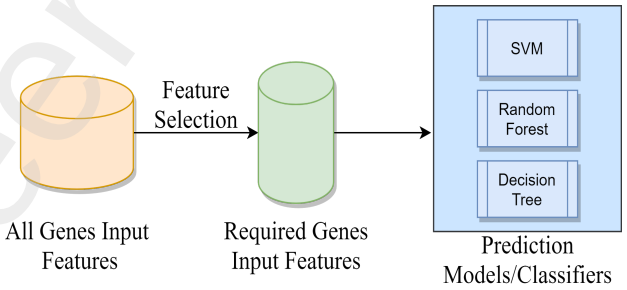

Fig. 2 - Model Building Flow

D. *Model Interpretation*

Explainable AI refers to a technique to interpret the results given by Machine Learning Model in a form that can be understood by Humans. Various algorithms like LIME (Local Interpretable Model-agnostic Explanations), SHAP (Shapely Additive Explanations), etc. are available to interpret the model. In this research Eli5 Library of Python is used. Eli5 stands for 'Explain Like I am 5 years old'. It helps to visualize and debug various machine learning models.

E. *UI Development*

Model was developed using Python 3. Flask 1.1.2 was used for model deployment. End users can enter the value of gene expressions manually, or can even upload the csv file containing patients gene-expression values to check the result given by the model.

V. RESULTS

Six Algorithms were implemented on the dataset used. The accuracies, F1 score, Log loss, AUC scores were calculated and compared.

All the implemented models achieved an accuracy of around 80 %. The reason for average accuracy is that the dataset contained more samples of patients who did not have a suitable chemotherapy and less samples of patients achieving complete response to treatment. If the dataset contained an equal number of both types of samples, then it would help in obtaining better accuracy. Below Table shows the performance of models.

Table 1 - Performance of algorithms

Sr No	Algorithm	Accuracy	AUC	F1 Score	Log Loss
1.	SVM	81.12%	0.7719	0.4637	0.44276
2.	Random Forest	80.61%	0.7826	0.3777	0.4461
3.	Decision Tree	79.59%	0.727	0.5341	3.530
4.	Gaussian NB	78.60%	0.8573	0.6577	1.6967
5	Ada Boost	77.55%	0.8035	0.4761	0.5368
6.	Logistic Regression	80.10%	0.8157	0.1702	0.4113

SVM has an accuracy of 81.12% and Random Forest has 80.61% accuracy. These two algorithms were found to have better accuracy amongst all. Confusion matrix of SVM is shown in Fig 3 below.

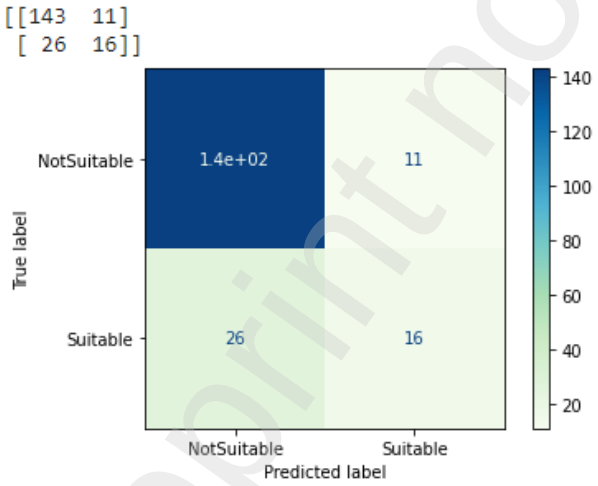

Fig. 3 - Confusion Matrix of Support Vector Machine

As shown in Fig 3,confusion matrix helps in giving a better idea about the classification model. There are 16 true positive labels and 143 true negative labels. Just measuring the accuracy of a model is not enough, real errors can be observed through the confusion matrix.

Another way of measuring the performance is precision recall curve. The precision-recall curve is the measure of prediction success especially when classes are imbalanced. In this study, as there are more labels of chemotherapy not suitable and less labels of chemotherapy suitable. Hence, this curve is an important factor here. In fig 4, the precision-recall curve is shown for Support Vector Machine Algorithm.

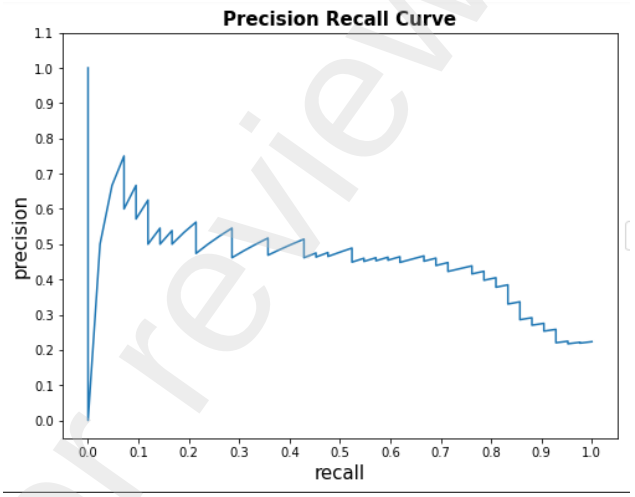

Fig. 4 - Precision-Recall Curve of SVM

Model Interpretability is observed using Eli5 which is a Python Library. It shows the features and their contribution towards the result. It is shown in Fig 5.

y=1 (probability **0.374**, score **0.003**) top features

Contribution[?]	Feature	Value
+7.316	FEN1	10.441
+7.297	TPX2	10.882
+6.319	TP53TG1	9.091
+3.762	DEK	12.325
+3.587	H2AFZ	12.424
+3.260	MCM6	10.850
+2.510	E2F3	9.999
+2.213	TOPBP1	10.570
+1.982	RARRES1	10.240
+1.803	MLPH	10.660
+0.721	PLCH1	8.237
+0.497	NFIB	13.097
+0.209	CYP2B7P	7.519
-0.204	ESR1	8.439
-0.209	ROPN1B	10.531
-2.038	METRN	8.618
-2.491	TTK	8.997
-2.721	SLC7A8	8.747
-3.051	MELK	10.170
-4.265	MCM2	11.392

Fig. 5 - Eli5 Explainable AI

In Fig 5, model interpretation is done. In this figure y=1 means that chemotherapy is suitable for that patient. The feature list is displayed in order according to the significance of that feature in the obtained result. For example, FEN1 is contributing the most in the result. In this way, it helps us to understand and explain results to the concerned people involved.

VI. CONCLUSION AND FUTURE SCOPE

From this research, it can be concluded that suitable breast cancer treatment can be predicted in advance using gene expressions data. But more work can be done on datasets as they are available in limited amounts to all users. Treatment prediction can be used for cancer as well as various other diseases which have a variety of drugs in the market. In this system, six algorithms were implemented with decent accuracy, but there is scope of improvement. Also some clinical factors like age, stage of cancer, type of breast cancer can be integrated in the dataset and may help achieve better accuracy. This step will help in making cancer treatment more precise and personalized with affordable costs.

VII. LIMITATIONS OF THE STUDY

This study is limited only to the chemotherapy treatment suitability for breast cancer patients. Also only genomic data is used in this study. Various clinical features can be used with genomic data in future to boost the accuracy.

VIII. REFERENCES

[1] Y. H. Qu, H. T. Zhu, K. Cao, X. T. Li, M. Ye, and Y. S. Sun, "Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using a deep learning (DL) method," *Thoracic Cancer*, vol. 11, no. 3, pp. 651–658, 2020.

[2] G. A. Brooks, A. J. Kansagra, S. R. Rao, J. I. Weitzman, E. A. Linden, and J. O. Jacobson, "A Clinical Prediction Model to Assess Risk for Chemotherapy-Related Hospitalization in Patients Initiating Palliative Chemotherapy," *JAMA Oncology*, vol. 1, no. 4, p. 441, 2015.

[3] R. B. Parikh, A. Elfiky, M. J. Pany, and Z. Obermeyer, "A machine learning approach to predicting short-term mortality risk for patients starting chemotherapy.," *Journal of Clinical Oncology*, vol. 35, no. 15_suppl, pp. 6538–6538, 2017.

[4] C. Huang, E. A. Clayton, L. V. Matyunina, L. D. E. McDonald, B. B. Benigno, F. Vannberg, and J. F. McDonald, "Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy," *Scientific Reports*, vol. 8, no. 1, 2018.

[5] Y.-C. Chiu, H.-I. H. Chen, T. Zhang, S. Zhang, A. Gorthi, L.-J. Wang, Y. Huang, and Y. Chen, "Predicting drug response of tumors from integrated genomic profiles by deep neural networks," *BMC Medical Genomics*, vol. 12, no. S1, 2019.

[6] E. J. Mucaki, J. Z. L. Zhao, D. Lizotte, and P. K. Rogan, "Predicting Response to Platin Chemotherapy Agents with Biochemically- inspired Machine Learning," 2017.

[7] B. R. Andjelkovic Cirkovic, "Machine learning approach for breast cancer prognosis prediction," *Computational Modeling in Bioengineering and Bioinformatics*, pp. 41–68, 2020.

[8] Azuaje F. Computational models for predicting drug responses in cancer research. Brief Bioinform. 2016; pii: bbw065 (Epub ahead of print). [PMC free article][PubMed]

[9] E. Lin, P.-H. Kuo, Y.-L. Liu, Y. W.-Y. Yu, A. C. Yang, and S.-J. Tsai, "A Deep Learning Approach for Predicting Antidepressant Response in Major Depression Using Clinical and Genetic Biomarkers," *Frontiers in Psychiatry*, vol. 9, 2018.

[10] Y. Chen, J. Sun, L.-C. Huang, H. Xu, and Z. Zhao, "Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations," *BioMed Research International*, vol. 2015, pp. 1–9, 2015.

[11] S.-K. Low, H. Zembutsu, and Y. Nakamura, "Breast cancer: The translation of big genomic data to cancer precision medicine," *Cancer Science*, vol. 109, no. 3, pp. 497–506, 2017.

[12] Abou Tabl, Ashraf & Alkhateeb, Abedalrhman & Elmaraghy, W. & Rueda, Luis & Ngom, Alioune," A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer", Frontiers in Genetics. 10. 256. 10.3389/fgene.2019.00256.

[13] D. Hamdan, T. T. Nguyen, C. Leboeuf, S. Meles, A. Janin, and G. Bousquet, "Genomics applied to the treatment of breast cancer," *Oncotarget*, vol. 10, no. 46, pp. 4786–4801, 2019.

[14] B.-J. Kim and S.-H. Kim, "Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method," *Proceedings of the National Academy of Sciences*, vol. 115, no. 6, pp. 1322–1327, 2018.

[15] "Opt-In," *Home Page - NCB Management Services, Inc.* [Online]. Available: https://www.ncbi.nlm.nih.gov/guide/genes-expression/ [Accessed: 12-Nov-2020].

[16] *cBioPortal for Cancer Genomics*. [Online]. Available: https://www.cbioportal.org/. [Accessed: 2-Dec-2020].

[17] *Towards Data Science*. [Online]. Available: https://towardsdatascience.com/. [Accessed: 9-Dec-2020].

[18] "Genomics of Drug Sensitivity in Cancer," *CancerRxGenes*. [Online]. Available: https://www.cancerrxgene.org/. [Accessed: 5-Dec-2020].

[19] "Your Machine Learning and Data Science Community," *Kaggle*. [Online]. Available: https://www.kaggle.com/ [Accessed: 10-Dec-2020].

[20] Breastcancer.org. Accesses on: Feb 13, 2020. Available: https://www.breastcancer.org/risk/factors/genetics.