

Algorithm Based Tool to Determine First Distant Recurrence Pattern in Breast Cancer Patients after Curative Surgery

Varshinee Velayudha¹, Anagha Vasista², K Tejasri³, Dr. Madhura Gangaiah⁴

^{1,2,3}Student, EIE Department, RNSIT

⁴Assistant Professor, EIE Department, RNSIT

Bangalore, India

Email : varshinee.velayudha@gmail.com, anaghavasista@gmail.com, tejasri.ias@gmail.com , madhuragangaiah@gmail.com

Abstract: In this paper fuzzy logic is used to predict first sight of distant relapse pattern in breast cancer patients after curative surgery with data. The overly expressed proteins, their structures and inhibitors are listed. Few of the proteins chosen are COX2, PgR, Nestin, SNAI 1, CK 5 and GATA3. Three primary sites of metastasis that is bone, brain and skin are chosen. The data elements are grouped into clusters and membership functions are thus obtained. A system of neurons, either artificial or organic in nature is trained with these relationship equations to estimate the first metastasis site. This network is tested for various available combinations of proteins. The table hence developed is defuzzified using lambda cut sets. These results are then compared to the ones obtained with the neuro fuzzy model.

Keywords: Metastasis, proteins, inhibitors, artificial neural networks (ANN), fuzzy arithmetic, defuzzification.

I. INTRODUCTION

Breast cancer is the second most common cancer spotted in women, first being skin cancer. The number of mortalities associated with breast cancer is firmly diminishing, due to factors such as prior detection and a better awareness of the disease. Tumor metastasis is one of the main reasons for the high mortality, i.e., about 90% and it was also seen that 20% to 30% of patients with early-stage breast cancer suffered distant metastases. These cancer cells metastasize to specific organs, known as “organotrophic metastasis”, which is regulated by subtypes of breast cancer, host organ microenvironment, and cancer cells-organ interaction. Having comprehensive understanding about the molecular mechanisms of organotrophic metastasis becomes tremendously vital for prediction centred on biomarkers, development of advanced technological strategy, and subsequent improvement of patient results.

ANFIS stands for Adaptive Neural Fuzzy Interference System. It is an artificial neural network which combines both neural network and fuzzy logic. It has a special highlight/feature which merges the gains of both the above-mentioned systems into a single framework, that is, it has the advantages of fuzzy logic theory and the learning ability of the artificial neuron network in a single system. It is also called as a universal eliminator

and does not need a sigmoid function. The architecture of ANFIS is as shown in Figure 1. Five layers are employed to build this model. Each layer comprises of numerous nodes defined by the node function. Adaptive nodes which are denoted by squares, represent the parameter sets that are flexible in these nodes. On the contrary, fixed nodes, denoted by circles, correspond to the parameter sets that are fixed in the model. The first layer has input values and governs membership functions, second layer is called the rule layer that generates firing strengths, the third layer normalises these strengths, in the fourth layer, the normalised values are given as the input to the defuzzing step and the fifth layer is the output of the defuzzing step.

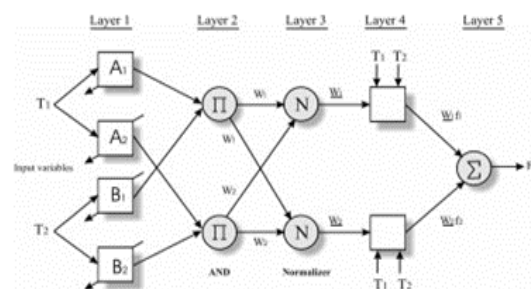


Figure 1: Basic structure of ANFIS

II. BLOCK DIAGRAM

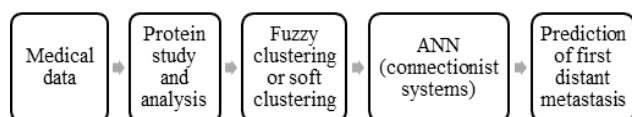


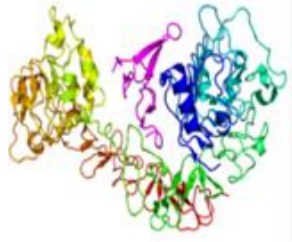
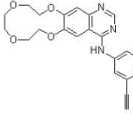
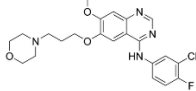
Figure 2: Block illustration of the project

Clinical records were collected and reviewed from several sites for this project which indicated that out of 2,032 cases that were known, 234 patients had distant recurrence in an average of approximately 2.7 years and

the total number of sites where it occurred was 321[1]. From the acquired statistics, concentration of the various involved proteins and the intensity of these in the tumour cell for each kind of metastasis were itemized (mainly the ones that were excessively expressed). The clustering of these proteins was performed and using the artificial neural network as the computing approach, the first distant relapse site was predicted.

Table 1: Protein structures and their inhibitors

Proteins	Structure	Features	Inhibitors	Disadvantages
COX2		Cyclooxygenase (COX) is an enzyme that is inducible and is associated with inflammatory diseases and carcinogenesis and is a primary suspect in case of angiogenesis and for the invasion of tumours.	1. Nimesulide 2. Celecoxib 	Due to concerns about the risk of hepatotoxicity, nimesulide is not popularly used. It exposes patients to fatal liver damage. Celecoxib can increase the risk of fatal heart attack or stroke and cause intestinal and stomach bleeding.[5]
PgR		Progesterone receptor is a protein found inside cells and is activated by the steroid hormone progesterone. PgR along with estrogen receptor are found in breast cancer cells and are mainly dependent on progesterone and related hormones like oestrogen to grow.	1. Anastrozole 2. Letrozole 	Due to concerns of vaginal bleeding and sleep patterns along with decreasing the density of bone, anastrozole and letrozole should only be taken on being prescribed by a doctor[5]. These drugs also increase the level of cholesterol in the body which has its own side effects.[9]

<p>Epidermal growth factor receptor</p>		<p>The epidermal growth factor receptor plays essential role in regulating cell proliferation, survival, differentiation, and migration. It is a primary Factor in epithelial malignancies and enhances tumour invasion and metastasis.</p>	<p>1.Icotinib</p>  <p>2.Gefitinib</p> 	<p>The major adverse effects of icotinib include skin and stomach problems. Gefitinib in certain cases cause interstitial lung diseases (pneumonia or inflammation in lungs without infection) along with elevation in liver function tests.</p>
------------------------------------------------	-----------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

III. FUZZY CLUSTERING

A. Basic terms

- i. **Data:** Data are characteristics or information, usually in the form of numerical value but could also be categorical or in the form of text, that are collected through observation [3]. When this data is represented as a matrix, it contains subjects in various subunits, such as time and value.
- ii. **Degree of membership:** The probability that a given data set belongs to centres is referred to as degree of membership. A membership function for a fuzzy set A on the universe of discourse X is defined as $\mu_A: X \rightarrow [0,1]$ where each element of X is mapped to a value between 0 and 1. This implies that a value closer to 1 indicates that an element belongs to a set whereas a value closer to 0 it does not.
- iii. **Clustering:** Clustering or cluster analysis is the process of grouping a set of objects in a way that those in the same group are more related each other than any other object found in a different group. A cluster need not always be a group of objects but can also be information ,i.e., data pointers, people that are located close together .When these series of numbers are plotted on a graph then it can be seen that several dots are gathered together and this is referred to as a cluster.

There are 2 types of clustering:

- Soft clustering
- Hard clustering

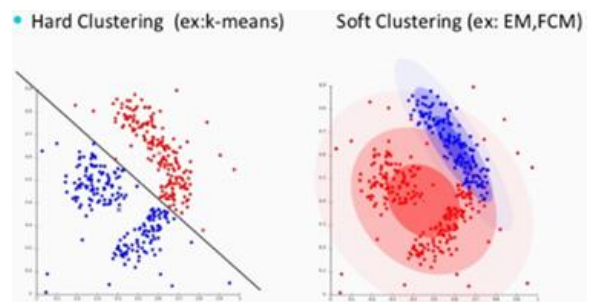


Figure 3: Clustering

Fuzzy c-means clustering (FCM): It is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. It is based on minimization of objective function. Common applications of fuzzy clustering are in marketing, biology, city planning and earthquake studies. [2]

B. Membership functions

Membership level is defined as the association of the data points to each of the clusters. It indicates the strength of association between the data elements and cluster. In fuzzy mathematics, the membership function of a fuzzy set is a simplification of the indicator function for standard sets. In fuzzy logic, it signifies the degree of truth as an extension of valuation.

Fuzzy clustering of proteins versus the first distant metastatic site is

$$\mu_{COX2} = \frac{0.435}{bone} + \frac{0.0648}{brain} + \frac{0.101}{skin}$$

$$\mu_{PgR} = \frac{0.584}{bone} + \frac{0.0094}{brain} + \frac{0.103}{skin}$$

$$\mu_{EFGR} = \frac{0.444}{bone} + \frac{0.0833}{brain} + \frac{0.166}{skin}$$

In this project the COX2 is the protein biomarker chosen and it is present in the tumour cell of bone, brain, and skin metastasis. Its membership level is different for bone, brain, and skin. It is allotted a membership function of 0.435, 0.0648 and 0.101 for bone, brain and skin respectively. In a similar method fuzzy relationship functions are created for 2 more protein biomarkers.

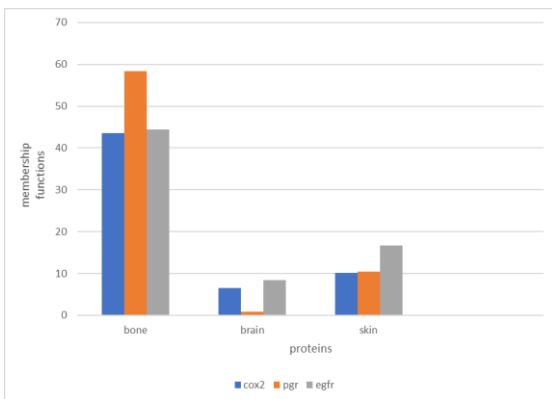


Figure 4: Bar Graph

C. ANFIS Algorithm

a) Get the input matrix ‘f’ in MATLAB for cluster heads of various metastasis sites employing the medical records taken from breast cancer patients.

$$f = \begin{bmatrix} 0.435 & 0.584 & 0.444 & 1 \\ 0.0648 & 0.0094 & 0.083 & 0 \\ 0.101 & 0.103 & 0.166 & 0 \end{bmatrix}$$

Figure 5: Input matrix

In the above matrix, the initial three columns show the proteins intensity and the last column shows the output. The proteins represented in the columns are COX2, PgR and EGFR respectively and the cluster head for bone, brain and skin are exhibited in the three rows. The entry ‘1’ can be made for any one row of the last column indicating the presence of cancer of the organ. The entry ‘0’ in the other two rows will henceforth shows that cancer will not prevail in the other two considered organs. For instance, here entry ‘1’ in the fourth column and first row of the above matrix indicates the existence

of bone cancer whereas the entry ‘0’ in the other two rows of the same column shows the absence of brain and skin cancer. This input matrix is used to train the ANFIS model to recognize bone metastasis in the absence of brain and skin cancer.

b) The network is trained using ANFIS in MATLAB with a parallel training matrix for different metastasis. On being trained, a training error is detected, and corresponding alterations are brought about in the input parameters of the ANFIS model to lessen the inaccuracies.

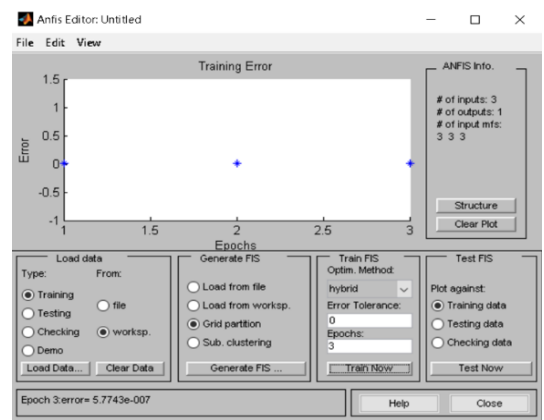


Figure 6: Training window for ‘f’ matrix

c) After the errors are curtailed and the output is observed to be adequate, the performance of FIS is tested. The neuro fuzzy model thus built is tested to obtain the finest pattern of proteins which predict the required first metastatic site with more correctness.

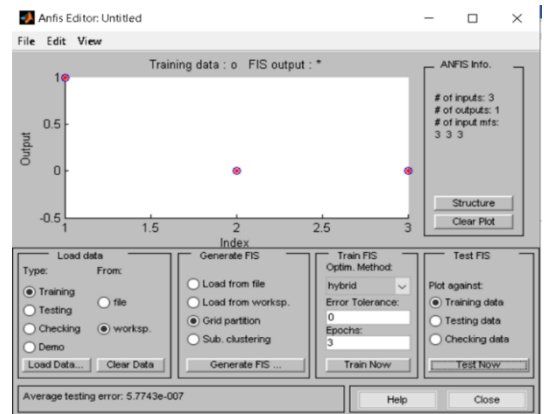


Figure 7: Testing window for ‘f’ matrix

d) The test matrix ‘r’ is produced with various combination of proteins. The neuro fuzzy structure is trained and tested with matrix r.

$$r = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Figure 8: Testing matrix

The columns in the matrix 'r' correspond to COX2, PgR and EGFR proteins. The entry '1' characterizes the presence of a protein whereas '0' indicates the absence of the same. Replacing the '1's and '0's in the matrix with the actual concentration of the proteins will give better predictions. Train the matrix.

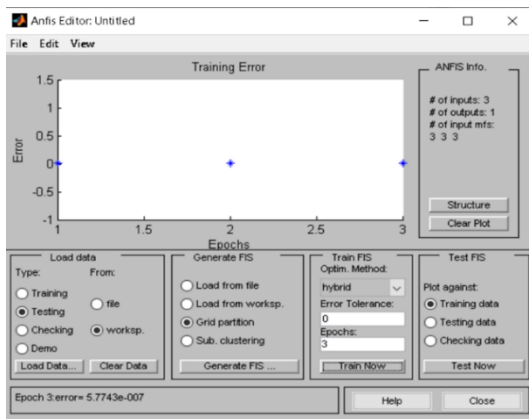


Figure 9: Training window for 'r' matrix

e) Test the matrix.

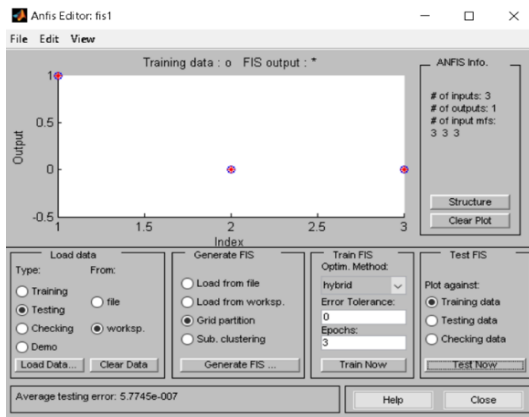


Figure 10: Testing/output window for 'r' with training data

f) Save the matrix with any name of interest, here saved as 'fis1'. An instruction

$$x = \text{readfis}('fis1')$$

is executed. Following the execution, the command 'evalfis' is used to gauge the FIS generated. The syntax for the latter command is

$$\text{output} = \text{evalfis}(f, x)$$

The output is viewed in the command window.

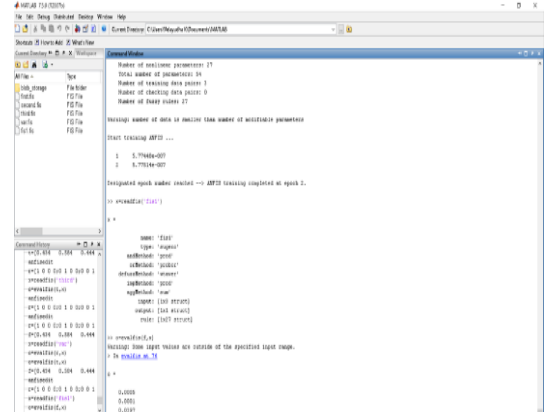


Figure 11: Output window

$$O = \begin{bmatrix} 0.0005 \\ 0.0001 \\ 0.0397 \end{bmatrix}$$

Figure 12: Output Matrix

g) ANFIS structure of the organ, here bone is viewed. This structure has five layers, namely, input, input membership, the 27 rules that are used, output membership, the obtained output.

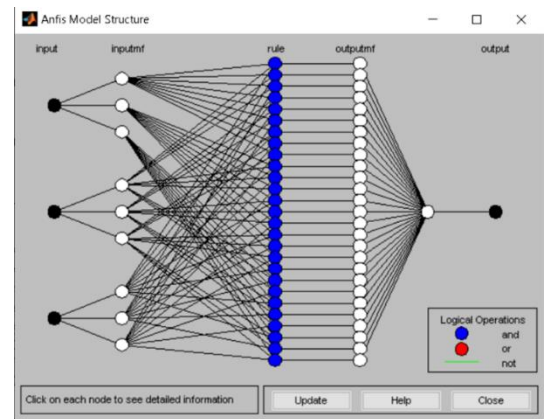


Figure 13: ANFIS structure

Inference: From the matrix 'O' the following is inferred. 3.97% of patients with only EGFR had bone as first distant metastasis. 0.01% of patients with only PgR had the first distant metastasis in bone and 0.05% of patients with only COX2 had bone metastasis. Amongst the three selected proteins, PgR has the slightest amount of impact on bone metastasis. Even though the ANFIS is optimised to give good quality results, it is bulky and thus can be employed only to a limited number of proteins and needs MATLAB command.

IV. FUZZY ARITHMETIC

The procedure consists of four steps:

Step 1: The value of proteins corresponding to first distant metastasis is calculated by taking the ratio of number of patients corresponding to the organ to number of first metastatic sites [4]. These values are presented in table 2.

Step 2: The value of 'x' is processed for normalization. This 'x' is achieved by taking the ratio of number of breast cancer patients with protein expression to the number of breast cancer patients assessed. These calculated values are shown in table 3.

Step 3: Normalization is performed by multiplying the values in table 2 with 'x' values in table 3. The normalized values are shown in table 4.

Step 4: The maximum and minimum values in each column of table 3 is selected. A lambda value is chosen, which is then multiplied with the maximum value to get lambda cut set. When the cancer is in second or third stage, the lambda value chosen must be less as the risks of metastasis is high in advanced stages. Lesser value of lambda results in the presence of a greater number of proteins triggering metastasis and therefore more inhibitors are required to restrain these proteins. The lambda value is calculated by dividing 1 by the number of variables chosen to form the membership function. In this project, the number of variables is 3 and hence $\lambda = 0.33$ [6]. The value of lambda cut set is obtained by multiplying the highest normalised value in table 4 with the calculated lambda value, i.e., 0.375×0.33 . It is then contrasted with each value in column of table 3. If the corresponding value is greater than lambda cut set level, then this value is defuzzified to 1 or if not to 0. This is tabulated in table 4.

Table 3: normalization value X for different proteins

PROTEINS	X
COX2	0.4655
PgR	0.46
EGFR	0.1836
HER2	0.2812
ER	0.612
CK 5	0.1014
Nestin	0.0923
Prominin-1	0.0833
SMA	0.041
SNAI1	0.0677
SNAI2	0.2657
CK 18	0.9684
E-Cadherin	0.8957
GATA3	0.6875

Table 4: Normalized protein values

	BONE	SKIN	BRAIN	LIVER	LUNG	OTHERS
COX2	0.1457	0.0341	0.0217	0.0745	0.0993	0.09
PgR	0.1994	0.0354	0.0032	0.0772	0.0418	0.1029
EGFR	0.0459	0.0171	0.0086	0.0315	0.043	0.0373
HER2	0.0894	0.0276	0.0028	0.0556	0.0556	0.0494
ER	0.2709	0.0539	0.0287	0.0988	0.0637	0.1148
CK5	0.0304	0.0034	0.0135	0.0101	0.0236	0.0203
Nestin	0.0342	0.0068	0.001	0.0102	0.0137	0.0171
Prominin-1	0.0217	0.0036	0.0109	0.0108	0.0145	0.0217
SMA	0.0126	0	0.0063	0.0063	0.0063	0.0095
SNAI1	0.0294	0.0058	0.0017	0.0095	0.0074	0.0139
SNAI2	0.1081	0.0098	0.0098	0.0426	0.0426	0.0525
CK 18	0.3754	0.0717	0.0329	0.1745	0.1284	0.1844
E-Cadherin	0.3592	0.0553	0.033	0.1591	0.1175	0.1756
GATA3	0.2894	0.0526	0.0158	0.125	0.0688	0.1348

Table 2: Protein values corresponding to the first distant metastasis

	BONE	SKIN	BRAIN	LIVER	LUNG	OTHERS
COX2	0.313	0.0733	0.0466	0.16	0.2133	0.1933
PgR	0.4335	0.0769	0.0069	0.1678	0.0909	0.2237
EGFR	0.25	0.093	0.04687	0.1718	0.2343	0.2031
HER2	0.318	0.098	0.01	0.1978	0.1978	0.1758
ER	0.4427	0.088	0.04687	0.1614	0.1041	0.1875
CK 5	0.3	0.0333	0.1333	0.1	0.233	0.2
Nestin	0.37	0.074	0.0111	0.111	0.1481	0.1851
Prominin-1	0.26	0.043	0.1304	0.13	0.1739	0.2608
SMA	0.3076	0	0.1538	0.1538	0.1538	0.2307
SNAI1	0.435	0.085	0.025	0.14	0.11	0.205
SNAI2	0.407	0.037	0.037	0.1604	0.1604	0.1975
CK 18	0.3877	0.074	0.034	0.1802	0.1326	0.1904
E-Cadherin	0.401	0.0617	0.0368	0.1776	0.1312	0.196
GATA3	0.421	0.0765	0.023	0.1818	0.1	0.196

Table 5: Defuzzification results

	BONE	SKIN	BRAIN
COX2	1	0	0
PgR	1	0	0
EGFR	0	0	0
HER2	0	0	0
ER	1	0	0
CK 5	0	0	0
Nestin	0	0	0
Prominin-1	0	0	0
SMA	0	0	0
SNAI1	1	0	0
SNAI2	0	0	0
CK 18	1	0	0
E-Cadherin	1	0	0
GATA3	1	0	0

Inference: It can be inferred from table 5 that COX2, PgR, ER, SNAI1, CK 18, E-Cadherin and GATA 3 are expressed only in bone metastasis whereas none of the proteins were expressed either in skin or brain metastasis. Hence for the combination of proteins and organs that have been chosen in this paper, inhibitor that binds to the proteins that have been overly expressed needs to be chosen to prevent metastasis. In case a protein contributes to more than one organ cancer, inhibitor that can bind to all those proteins will have to be considered.

V. CONCLUSION

This is a paper has been created as a prophecy model for the first distant site metastasis of breast cancer after curative surgical procedure. Membership functions have been established for proteins that are expressed in excessive quantity. The ANN vector has then been trained and tested for different responses based on the informant analysis that had been collected prior to the starting of the project. Use of manually evaluated outcomes are restricted due to the involved errors. The method used here for output involves ANFIS and lambda cut set setups. ANFIS is basically a branch of artificial intelligence which involved fuzzification, defuzzification and rule base and lambda cut set method was used to get crisp values. The results thus obtained where reasonably near to the real time results and helped in therapy.

VI. FUTURE WORK

For healthier and more precise results, a greater number of membership functions can be developed making different permutations and combinations of the proteins and the organs involved. Also, newer and alternate software such as image processing, python, MACHINA, which stands for metastatic and clonal history integrative analysis can be jointly used to verify the results.

REFERENCES

- [1] Sihto, Harri, Johan Lundin, Mikael Lundin, Tiina Lehtimäki, Ari Ristimäki, Kaija Holli, Liisa Sailas et al. "Breast cancer biological subtypes and protein expression predict for the preferential distant metastasis sites: a nationwide cohort study." *Breast Cancer Research* 13, no. 5 (2011): R87.
- [2] https://www.slideshare.net/aydinayanzadeh/fuzzy-clusteringcmeans-kmeans?qid=c0324b16-0f34-4342-9188-47e71c1c1043&v=&b=&from_search=3
- [3] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system." *IEEE transactions on systems, man, and cybernetics* 23, no. 3 (1993): 665-685.
- [4] <https://www.drugbank.ca/drugs/DB04743>
- [5] <https://in.mathworks.com/matlabcentral/answers/index>
- [6] <https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470#:~:text=Breast%20cancer%20is%20cancer%20that,far%20more%20common%20in%20women.>
- [7] <https://drsusanloverresearch.org/estrogen-receptor-and-progesterone-receptor-pr-positive-breast-cancer/>