Discriminating the type of Tumor based on its Features.

Final Project Report – Spring 2023

Submitted by:

Shivam Ranjan [20017743]

Sai Mounish Ramisetti [20018010]

Javeed Shaik [20018566]

Zhecheng Qin

BIA 650 – PROCESS ANALYTICS AND OPTIMIZATION

Prof. Edward A. Stohr

Table of Contents

Discriminating the type of Tumor based on its Features.

Abstract:

Cancer is a complex disease caused by a variety of genetic and environmental variables. Accurately identifying the type of tumor is critical for precise diagnosis, prognosis, and therapy plan. Conventional cancer diagnosis methods, such as histological examination and immunohistochemistry, are time-consuming, intrusive, and necessitate the use of highly skilled professionals. Recent advances in molecular biology and computational technologies, on the other hand, have enabled the creation of non-invasive and extremely precise approaches for diagnosing cancer in patients.

Accurate classification of tumors plays a crucial role in medical research and clinical practice. Discriminating between different types of tumors based on their features can inform diagnosis, guide treatment decisions, and predict patient outcomes. This abstract provides an overview of the methodology, steps, and potential applications of discriminant analysis in tumor classification, highlighting its significance in advancing cancer diagnostics and treatment.

Machine learning and artificial intelligence (AI) algorithms have also been employed for the identification of cancer types in patients. These algorithms can analyze large amounts of genomic and clinical data to identify patterns and features that are associated with different types of cancer. For example, a recent study demonstrated that an AI algorithm trained on gene expression data from over 2,000 breast tumors could accurately classify breast cancer subtypes with over 90% accuracy.

In our project, we will manage to look for a good way to detect the cancer type by Genetic Algorithm and Discriminant Analysis. And as we know, it’s especially serious when a malignant tumor predicted as beginning type, because of the risk of delayed treatment, we will try to find a more reasonable way to cut off this type of error.

Discriminating the type of Tumor based on its Features.

Introduction:

Identifying and understanding the characteristics of tumors is a critical task in the field of medical research and healthcare. The ability to accurately discriminate between different types of tumors based on their features plays a crucial role in diagnosis, treatment planning, and patient prognosis. In recent years, various statistical and machine learning techniques have been employed to develop effective methods for tumor classification. One such method is discriminant analysis, which enables researchers to analyze the features of tumors and determine the factors that distinguish one type from another.

Discriminant analysis, also known as classification analysis or supervised learning, is a statistical approach that aims to find a discriminant function to differentiate between predefined groups. In the context of tumor classification, discriminant analysis utilizes a set of features, such as size, shape, texture, and genetic markers, to identify the specific type of tumor. By examining these features, the discriminant function assigns weights or coefficients to each variable, indicating their relative importance in distinguishing between tumor types.

The process of discriminating tumor types based on their features begins with the collection of relevant data, including clinical observations, imaging scans, histopathology reports, and molecular profiles. These data provide valuable insights into the unique characteristics exhibited by different tumor types. Discriminant analysis is then applied to this dataset to develop a classification model. The model is trained using labeled data, where each tumor instance is already assigned to a specific type. The discriminant function is derived from this training process and is subsequently used to predict the type of new, unlabeled tumors.

The accurate classification of tumors based on their features has numerous clinical implications. It can guide treatment decisions by identifying the most appropriate therapeutic interventions for a specific tumor type. For example, certain tumor types may respond better to specific medications or therapies, and precise classification enables personalized treatment plans. Additionally, tumor classification can contribute to the understanding of disease progression, enabling the identification of prognostic factors and the development of targeted interventions.

As the field of medical research continues to advance, discriminant analysis offers a valuable tool in the quest for improved tumor classification. By leveraging the power of statistical methods, this approach helps unravel the complex relationships between tumor features and their types. Consequently, it facilitates more accurate diagnoses, individualized treatment strategies, and ultimately, better patient outcomes.

In the subsequent sections, we will delve further into the methodology, steps, and applications of discriminant analysis in tumor classification, exploring its potential to revolutionize cancer diagnostics and treatment.

Discriminating the type of Tumor based on its Features.

Types of Tumors:

There are various types of tumors that can develop in the human body. But based on our objective of this project, our concern is about the broad classification i.e, Benign and Malignant.

1. **Benign Tumors**: Benign tumors are non-cancerous growths that do not invade nearby tissues or spread to other parts of the body. They tend to grow slowly and have well-defined boundaries. Examples of benign tumors include:

a. Adenomas: These tumors develop in the glandular tissues, such as the breast, colon, or thyroid gland.

b. Lipomas: Lipomas are composed of fat cells and commonly occur just beneath the skin.

c. Fibroids: Fibroids are benign tumors that form in the muscular wall of the uterus.

d. Meningiomas: Meningiomas are tumors that arise from the meninges, the protective membranes surrounding the brain and spinal cord.

2. **Malignant Tumors**: Malignant tumors, also known as cancerous tumors, are more aggressive and have the potential to invade nearby tissues and spread to other parts of body through a process called metastasis. Examples of Malignant tumors:

a. Carcinomas: Carcinomas are cancers that develop in the epithelial cells, which line the surfaces and organs of the body. The most common types of carcinoma include breast, lung, prostate, and colorectal cancer.

b. Sarcomas: Sarcomas arise from connective tissues, such as bones, muscles, and blood vessels. Examples include osteosarcoma, rhabdomyosarcoma, and angiosarcoma.

c. Lymphomas: Lymphomas are cancers of the lymphatic system, which is part of the immune system. They can be further classified into Hodgkin lymphoma and non-Hodgkin lymphoma.

d. Leukemias: Leukemias are cancers of the blood-forming tissues, particularly the bone marrow and blood. They result in the overproduction of abnormal white blood cells.

It's important to note that these are just a few examples, and there are many other specific types of tumors that can occur in different organs and tissues throughout the body. Each type of tumor has its own characteristics, behavior, and treatment options. Accurate classification of tumors based on their specific type is crucial for appropriate diagnosis, prognosis, and treatment planning.

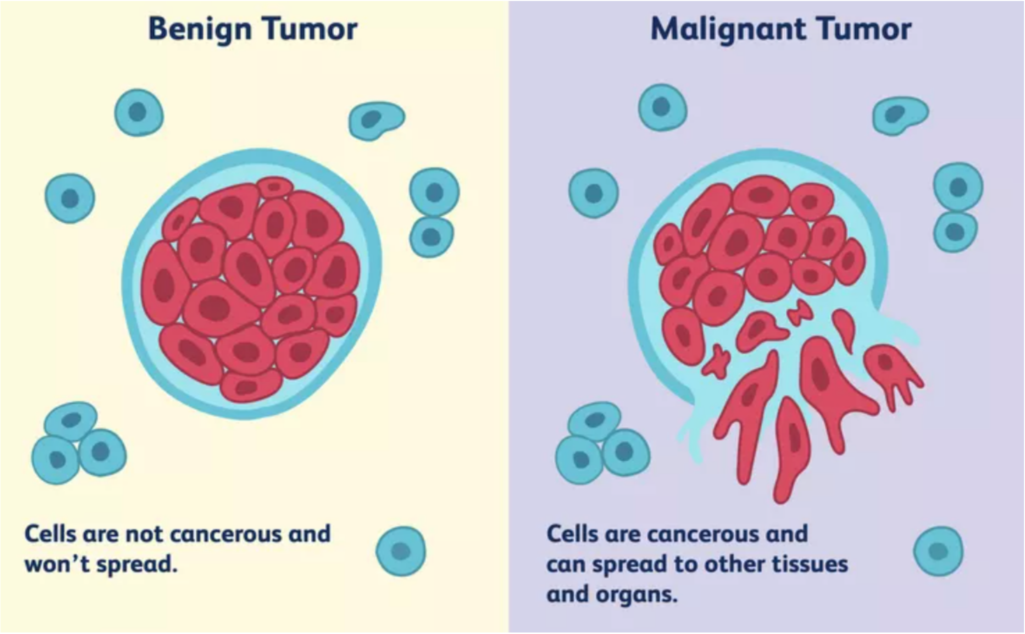

Benign Tumor
Malignant Tumor
Cells are not cancerous and won't spread.
Cells are cancerous and can spread to other tissues and organs.

Fig 1. Benign and Malignant Tumors

Discriminating the type of Tumor based on its Features.

Dataset and Feature Selection using Genetic Algorithm:

To predict the type of tumor, we need data of various features of a tumor. Fortunately, we found a dataset containing 569 tumor samples with 30 different attributes. These attributes are the features of each tumor sample. The dataset has the following features:

id	concave_points_mean	compactness_se	area_worst
radius_mean	symmetry_mean	concavity_se	smoothness_worst
texture_mean	fractal_dimension_mean	concave_points_se	compactness_worst
perimeter_mean	radius_se	symmetry_se	concavity_worst
area_mean	texture_se	fractal_dimension_se	concave_points_worst
smoothness_mean	perimeter_se	radius_worst	symmetry_worst
compactness_mean	area_se	texture_worst	fractal_dimension_worst
concavity_mean	smoothness_se	perimeter_worst	diagnosis

Most of the features shown in the above table are either redundant or useless in determining the type of tumor. Only few of these features might be useful in precisely predicting the tumor type. But we do not know which features to choose. However, we can find the best features using **Genetic Algorithm**.

Genetic algorithms (GAs) are search and optimization techniques inspired by the principles of natural selection and evolution. They are a type of metaheuristic algorithm that can efficiently solve complex problems by mimicking the process of natural evolution.

The basic idea behind a genetic algorithm is to create a population of potential solutions, often represented as strings of binary or real-valued values called "chromosomes" or "genomes." Each chromosome represents a candidate solution to the problem at hand. The population undergoes a series of iterative steps, including selection, crossover, and mutation, to generate new generations of individuals.

Discriminating the type of Tumor based on its Features.

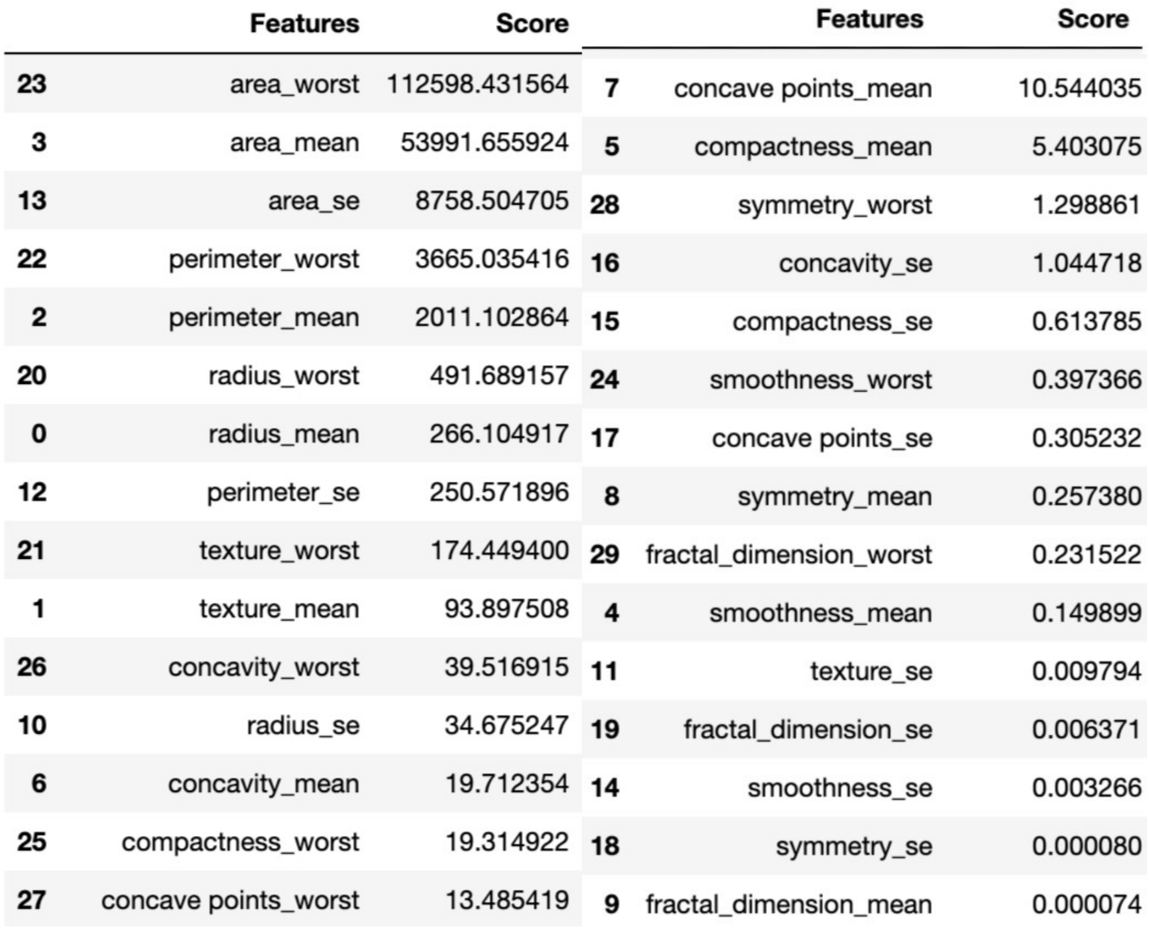
Detailed overview of steps involved in a genetic algorithm:

1. Initialization: A population of randomly generated individuals is created as the initial generation.
2. Fitness Evaluation: Each individual's fitness or objective function is evaluated, which quantifies how well the solution performs with respect to the problem's criteria. The fitness function is problem-specific and can be based on various factors or constraints.
3. Selection: Individuals with higher fitness scores have a higher chance of being selected as parents for the next generation. This process, often referred to as "survival of the fittest," aims to bias the search towards better solutions.
4. Crossover: Selected individuals undergo crossover or recombination, where parts of their genetic material are exchanged or combined to produce offspring. This step mimics the biological process of genetic recombination and introduces new variations into the population.
5. Mutation: A small random change is applied to some individuals' genetic material to introduce additional diversity and prevent premature convergence to suboptimal solutions.
6. Replacement: The offspring, along with a portion of the previous generation, form the new population for the next iteration.
7. Termination: The algorithm continues iterating through selection, crossover, and mutation steps until a termination condition is met. This condition can be a maximum number of iterations, a specific fitness threshold, or the absence of significant improvements in successive generations.

Genetic algorithms excel in solving complex optimization problems with large solution spaces, non-linear relationships, and multiple criteria. They are particularly useful when the problem lacks a clear analytical solution or when the search space is too vast for exhaustive exploration. Genetic algorithms have been successfully applied in various domains, including engineering design, scheduling, data mining, and machine learning.

Discriminating the type of Tumor based on its Features.

Feature Selection: We found a python package called 'sklearn-genetic' which is based on Genetic Algorithm. It has a module called 'GeneticSelectionCV' which is what we used to find the 5 best features by calculating scores for each feature in python and choosing the top 5 features. The scores for each feature are shown below:



	Features	Score
23	area_worst	112598.431564
3	area_mean	53991.655924
13	area_se	8758.504705
22	perimeter_worst	3665.035416
2	perimeter_mean	2011.102864
20	radius_worst	491.689157
0	radius_mean	266.104917
12	perimeter_se	250.571896
21	texture_worst	174.449400
1	texture_mean	93.897508
26	concavity_worst	39.516915
10	radius_se	34.675247
6	concavity_mean	19.712354
25	compactness_worst	19.314922
27	concave points_worst	13.485419
7	concave points_mean	10.544035
5	compactness_mean	5.403075
28	symmetry_worst	1.298861
16	concavity_se	1.044718
15	compactness_se	0.613785
24	smoothness_worst	0.397366
17	concave points_se	0.305232
8	symmetry_mean	0.257380
29	fractal_dimension_worst	0.231522
4	smoothness_mean	0.149899
11	texture_se	0.009794
19	fractal_dimension_se	0.006371
14	smoothness_se	0.003266
18	symmetry_se	0.000080
9	fractal_dimension_mean	0.000074

Fig.2 Scores after running genetic algorithm to select best features.

We decided to use 5 best features (features with the highest scores). The features that we are going to use in our Discriminant Analysis are:

- area_worst
- area_mean
- area_se
- perimeter_worst
- perimeter_mean

Optimization Method - Discriminant Analysis:

Our job is to predict the type of tumor using the 5 best features that we selected previously. These 5 features are the explanatory variables, and the type of tumor is the response variable. The optimization method that is most appropriate to do this is **Discriminant Analysis**.

Discriminant analysis, also known as Linear Discriminant Analysis (LDA) is a statistical method used to categorize groups or populations. It is a multivariate technique that aims to determine which variables are most effective in differentiating between groups based on their characteristics. The primary objective of discriminant analysis is to find a discriminant function that maximizes the separation between groups.

In discriminant analysis, the groups being studied are referred to as the dependent variable or the categorical variable. The independent variables, also known as predictor variables, features or discriminators, are used to predict or classify the groups. The discriminant function combines these independent variables to create a linear combination that best discriminates between the groups.

The discriminant function is calculated using mathematical techniques, such as linear algebra and optimization algorithms. It assigns weights or coefficients to each independent variable based on its contribution to group separation. These coefficients indicate the relative importance of each variable in discriminating between groups.

Once the discriminant function is determined, it can be used to predict the group membership of new observations or individuals. By comparing the calculated discriminant scores with predefined thresholds or cutoffs, individuals can be classified into the appropriate groups.

Discriminant analysis assumes that the independent variables are normally distributed within each group and that the variance-covariance matrices are equal across the groups. It also assumes linearity between the independent variables and the discriminant function.

Discriminant analysis has numerous applications in various fields. It is commonly used in market research to segment customers, in medical research to predict disease outcomes (which is our major concern in this project), in social sciences to identify group differences in attitudes or behaviors, and in many other areas where group classification or prediction is required.

Discriminating the type of Tumor based on its Features.

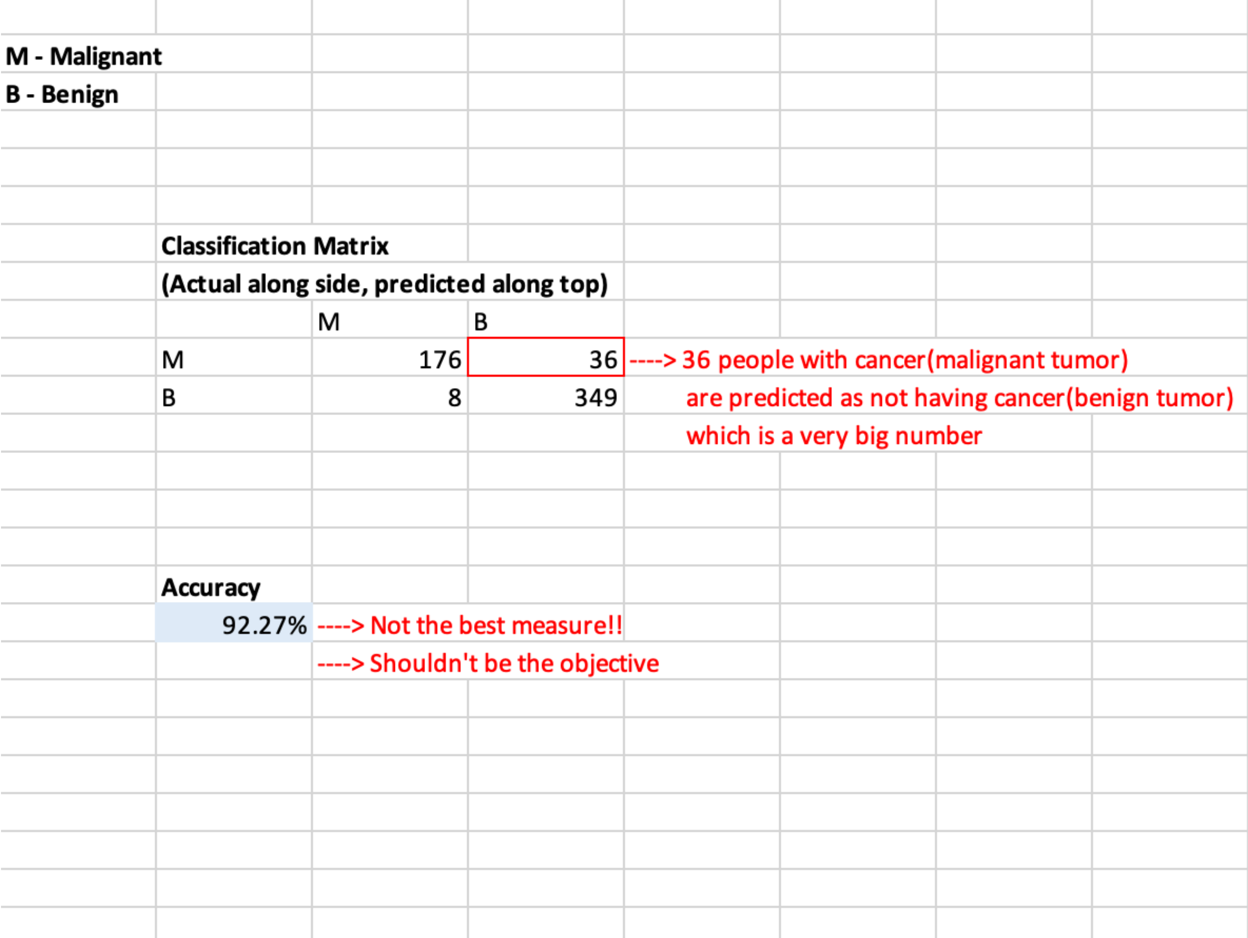
Optimization Model:

We built an Optimization Model in an Excel spreadsheet to perform discriminant analysis on the selected features to predict the type of tumor using the following steps:

- *Tumor Features:* We entered the data for the selected features in the spreadsheet (B15:F583). These features are the inputs, and they decide the type of tumor for each sample.
- *Decision Variables:* 'Weights' (B8:F8, Coefficients of the 5 features which are used to form discriminant scores) and 'Cutoff' (B11, Cutoff value for classification) are the decision variables.
- *Discriminant Scores:* Each discriminant score is a weighted combination of area_worst, area_mean, area_se, perimeter_worst and perimeter_mean. To calculate these in column H, we entered the formula: '=SUMPRODUCT(Weights,$B15:$F15)' in H15 and copied it down.
- *Classifications:* A tumor is classified as Benign if the tumor's discriminant score is below the Cutoff value; otherwise, the tumor is Malignant. So, we entered '=IF(H15<Cutoff,"B","M")' in I15 and copied it down.
- *Tallies:* It is important to tally the classifications in a classification matrix, as we did in range K15:M17. The best way to calculate the tallies is to use the COUNTIFS excel function. We entered the formula: '=COUNTIFS($G$15:$G$583,$K16,$I$15:$I$583,L$15)' in L16 and copied it to the range L16:M17. We then calculated the percentage of tumors correctly classified in cell K23 with the formula '=(L16+M17)/SUM(L16:M17)'. This is the accuracy of the discriminant analysis.
- *Punishment on type 1 Error (Malignant predicted as Benign):* To calculate the punishment of type1 error, we entered the formula: '=(L16+M17-5*M16)/SUM(L16:M17)' in cell L23. This the **Objective** to **Maximize** for the model.

Discriminating the type of Tumor based on its Features.

Reason for choosing Punishment on Error over Accuracy as the objective to maximize: Cancer is a life-threatening disease. Therefore, it is very important that people with Malignant Tumor shouldn't be diagnosed/predicted as Benign. If this happens, patients will not get the required diagnosis and treatment which may lead to death. If we choose Accuracy as the objective to maximize, we get the following results after performing discriminant Analysis:



M - Malignant
B - Benign

Classification Matrix
(Actual along side, predicted along top)

	M	B
M	176	36
B	8	349

----> 36 people with cancer(malignant tumor) are predicted as not having cancer(benign tumor) which is a very big number

Accuracy
92.27% ----> Not the best measure!!
----> Shouldn't be the objective

Fig 3. Accuracy shouldn't be the objective function.

As, you can see from Fig. 3; if we choose accuracy as the objective function, we may get a highly accurate (92%) discriminant analysis but we mis predicted 36 people with malignant Tumor as having a Benign Tumor. These 36 people might not get required treatment and 36 is a very big number considering the context and the sample size.

Discriminating the type of Tumor based on its Features.

So, we need to minimize this error of predicting "M" as "B" (Maximize the Punishment). This is the reason why we took Punishment on this error as the objective function to maximize.

Calculating the upper and lower bounds for the Constraints:

'Weights' are the fractional coefficients of the features. So, **-1 <= Weights <= 1**.

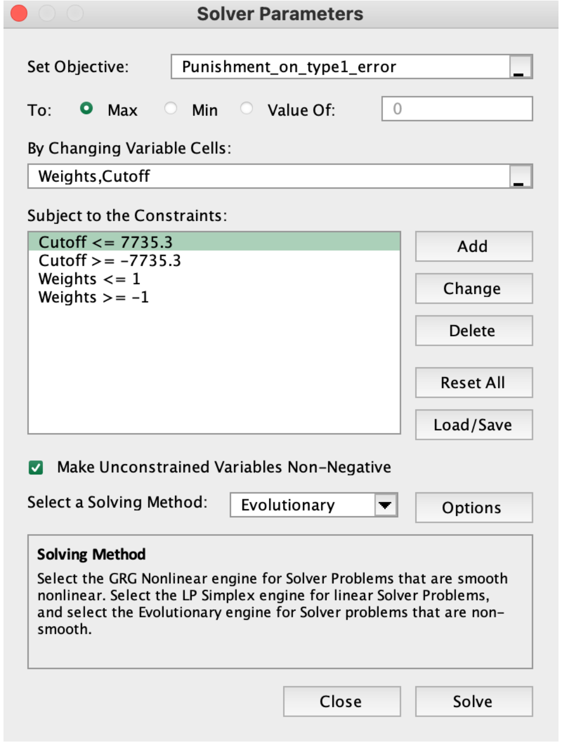
To find the upper and lower bounds for the 'Cutoff' value, we must find the maximum of all the summation values of the 5 features, which is 7735.3.

Hence, **-7735.3 <= Cutoff <= 7735.3**

The constraints are:

- Weights >= -1
- Weights <= 1
- Cutoff >= -7735.3
- Cutoff <= 7735.3

Evolutionary Method must be selected in the solver table. Ultimately, the solver table will look like this:



Solver Parameters

Set Objective: Punishment_on_type1_error

To: ● Max ○ Min ○ Value Of: 0

By Changing Variable Cells:
Weights,Cutoff

Subject to the Constraints:
Cutoff <= 7735.3
Cutoff >= -7735.3
Weights <= 1
Weights >= -1

☑ Make Unconstrained Variables Non-Negative

Select a Solving Method: Evolutionary

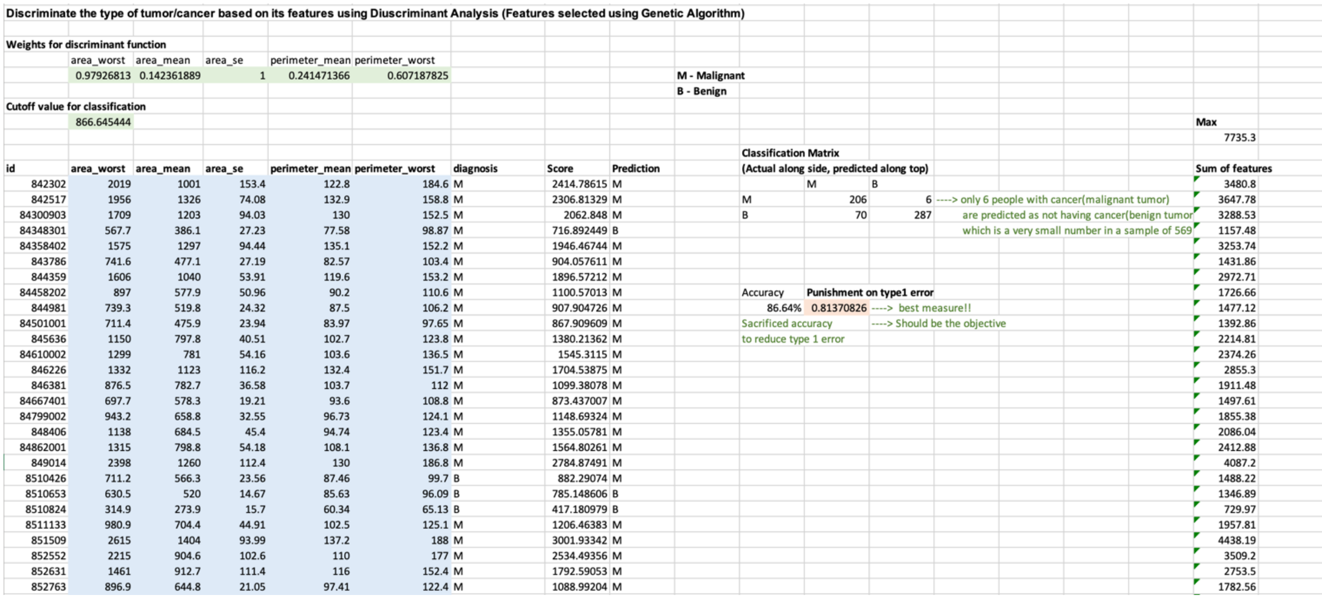
Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.

Fig 4. Solver

Discriminating the type of Tumor based on its Features.

Results & Discussion:

After maximizing the punishment on error by running the evolutionary method in solver, we got the following results. Our spreadsheet is shown in Fig 5.



Discriminate the type of tumor/cancer based on its features using Diuscriminant Analysis (Features selected using Genetic Algorithm)

Weights for discriminant function

area_worst	area_mean	area_se	perimeter_mean	perimeter_worst
0.97926813	0.142361889	1	0.241471366	0.607187825

M - Malignant
B - Benign

Cutoff value for classification
866.645444

Max
7735.3

Classification Matrix
(Actual along side, predicted along top)

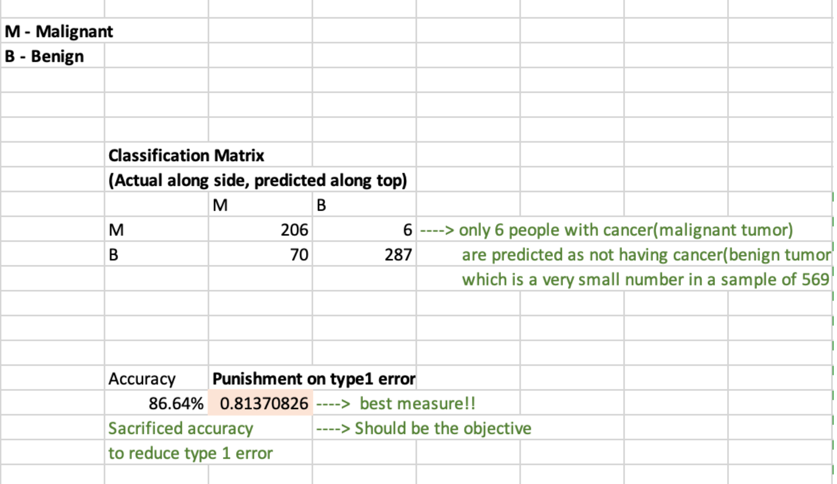
	M	B	
M	206	6	----> only 6 people with cancer(malignant tumor)
B	70	287	are predicted as not having cancer(benign tumor
			which is a very small number in a sample of 569

Accuracy	**Punishment on type1 error**	
86.64%	0.81370826	----> best measure!!
Sacrificed accuracy		----> Should be the objective
to reduce type 1 error		

id	area_worst	area_mean	area_se	perimeter_mean	perimeter_worst	diagnosis	Score	Prediction	Sum of features
842302	2019	1001	153.4	122.8	184.6	M	2414.78615	M	3480.8
842517	1956	1326	74.08	132.9	158.8	M	2306.81329	M	3647.78
84300903	1709	1203	94.03	130	152.5	M	2062.848	M	3288.53
84348301	567.7	386.1	27.23	77.58	98.87	M	716.892449	B	1157.48
84358402	1575	1297	94.44	135.1	152.2	M	1946.46744	M	3253.74
843786	741.6	477.1	27.19	82.57	103.4	M	904.057611	M	1431.86
844359	1606	1040	53.91	119.6	153.2	M	1896.57212	M	2972.71
84458202	897	577.9	50.96	90.2	110.6	M	1100.57013	M	1726.66
844981	739.3	519.8	24.32	87.5	106.2	M	907.904726	M	1477.12
84501001	711.4	475.9	23.94	83.97	97.65	M	867.909609	M	1392.86
845636	1150	797.8	40.51	102.7	123.8	M	1380.21362	M	2214.81
84610002	1299	781	54.16	103.6	136.5	M	1545.3115	M	2374.26
846226	1332	1123	116.2	132.4	151.7	M	1704.53875	M	2855.3
846381	876.5	782.7	36.58	103.7	112	M	1099.38078	M	1911.48
84667401	697.7	578.3	19.21	93.6	108.8	M	873.437007	M	1497.61
84799002	943.2	658.8	32.55	96.73	124.1	M	1148.69324	M	1855.38
848406	1138	684.5	45.4	94.74	123.4	M	1355.05781	M	2086.04
84862001	1315	798.8	54.18	108.1	136.8	M	1564.80261	M	2412.88
849014	2398	1260	112.4	130	186.8	M	2784.87491	M	4087.2
8510426	711.2	566.3	23.56	87.46	99.7	B	882.29074	M	1488.22
8510653	630.5	520	14.67	85.63	96.09	B	785.148606	B	1346.89
8510824	314.9	273.9	15.7	60.34	65.13	B	417.180979	B	729.97
8511133	980.9	704.4	44.91	102.5	125.1	M	1206.46383	M	1957.81
851509	2615	1404	93.99	137.2	188	M	3001.93342	M	4438.19
852552	2215	904.6	102.6	110	177	M	2534.49356	M	3509.2
852631	1461	912.7	111.4	116	152.4	M	1792.59053	M	2753.5
852763	896.9	644.8	21.05	97.41	122.4	M	1088.99204	M	1782.56

Fig 5. Optimization Model

We had to make a sacrifice on the accuracy of the analysis (86.6% now, which is 92% if accuracy is maximized). But the most important thing in this scenario (cancer diagnosis) is that the number of people with malignant tumor but predicted as benign should be as low as possible. (6 in the sample of 569, which is achieved by maximizing the punishment on this error).



M - Malignant
B - Benign

Classification Matrix
(Actual along side, predicted along top)

	M	B	
M	206	6	----> only 6 people with cancer(malignant tumor)
B	70	287	are predicted as not having cancer(benign tumor
			which is a very small number in a sample of 569

Accuracy	**Punishment on type1 error**	
86.64%	0.81370826	----> best measure!!
Sacrificed accuracy		----> Should be the objective
to reduce type 1 error		

Fig 6. Punishment on error is the Objective to Maximize.

The detailed classification is as follows:

Total number of tumor samples	569
Number of samples correctly predicted as Malignant	206
Number of samples correctly predicted as Benign	287
Number of Benign tumors wrongly predicted as Malignant	70
Number of Malignant tumors wrongly predicted as Benign	6
Accuracy (Percentage of tumor samples that are correctly predicted)	86.64%
Punishment on type 1 error (predicting M as B)	0.8137

We may think that 70 Benign samples are wrongly predicted as Malignant, but it is not as serious as predicting Malignant tumor as Benign, which can lead to death. If Benign samples are wrongly predicted as Malignant, they'll still get to know that they don't have cancer when they go for further advanced diagnosis or treatment.

Conclusion:

Determining the type of tumor is a key component of cancer diagnosis and treatment. Precise tumor identification is now possible with the help of improved diagnostic procedures such as imaging, biopsy, and genetic testing. But one of the most important benefits of determining the kind of tumor using a simple Discriminant Analysis is to decide if further diagnosis and treatment is required AND it is Very Cheap!!

The application of Discriminant Analysis and Feature Selection techniques has proven to be effective in discriminating different types of tumors based on their features. The study showcased the importance of selecting relevant and discriminative features to enhance the accuracy and efficiency of tumor classification.

By employing Discriminant Analysis, we were able to identify the key features that contribute significantly to the classification of tumors. This statistical technique allowed for the creation of a discriminant function that maximizes the separation between different tumor types, enabling accurate predictions based on the selected features.

Feature selection played a crucial role in enhancing the performance of the classification model. By eliminating irrelevant or redundant features, the complexity of the model was reduced, resulting in improved efficiency and interpretability. Since we found a python module for feature selection which is based on Genetic Algorithm, we didn’t need to worry about the internal mathematics involved in the Feature Selection process.

The combined use of discriminant analysis and feature selection provided a robust framework for discriminating tumor types based on their features. This approach has the potential to aid medical professionals in diagnosing and predicting tumor types more accurately, leading to better treatment strategies and improved patient outcomes.

It is important to note that the success of tumor discrimination using discriminant analysis and feature selection relies on the quality and availability of data. Therefore, ongoing efforts to collect comprehensive and well-annotated tumor datasets will further enhance the effectiveness of this approach in clinical practice.

In conclusion, the integration of discriminant analysis and feature selection techniques holds great promise in the field of tumor classification. The findings from this study provide valuable insights into the potential applications of these methods, highlighting their importance in advancing medical research and improving patient care.

References:

1. https://www.kaggle.com/datasets/erdemtaha/cancer-data
2. https://www.mathworks.com/help/gads/what-is-the-genetic-algorithm.html
3. https://www.geeksforgeeks.org/genetic-algorithms/
4. https://towardsdatascience.com/feature-selection-with-genetic-algorithms-7dd7e02dd237
5. https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240
6. https://my.clevelandclinic.org/health/diseases/21881-tumor
7. https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis
8. https://chat.openai.com/
9. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/discriminant-analysis - :~:text=Discriminant analysis (DA) is a,variable in separating the groups.
10. https://help.xlstat.com/6691-discriminant-analysis-excel-tutorial
11. Practical Management Science, Winston & Albright, Edition 4.
12. https://hastie.su.domains/Papers/pda.pdf
13. https://arxiv.org/pdf/1904.03469.pdf