Detection of Invasive Ductal Carcinoma (IDC)

Elif Göksu BİÇMEN
Department of Bioengineering
Marmara University
Istanbul, TURKEY
goksubicmen@gmail.com

İrem YARAR
Department of Bioengineering
Marmara University
Istanbul, TURKEY
Iremyararr6@gmail.com

Ömer Faruk OKUMUŞ
Department of Computer Science
Marmara University
Istanbul, TURKEY
omerokumus3@gmail.com

Harun ÖZDEMİR
Department of Computer Science
Marmara University
Istanbul, TURKEY
Hrn.ozdemir52@gmail.com

Sıla İNANIR
Department of Electrical & Electronics Engineering
Marmara University
Istanbul, TURKEY
edasilainanir@gmail.com

***Abstract*—Invasive ductal carcinoma (IDC) is a type of breast cancer. The intention of this study is to predict and detect IDC and non-IDC tissue slices by analysing the images of them selected from the dataset. To image processing, supervised which are SVM, k-NN, Decision Tree, and ANN algorithms, and unsupervised which is k-means, learning methods have been used. Results show that between four supervised algorithms that were used in this study, the SVM algorithm gives the best result in total (accuracy rate=0.82 and best F1 score) similar to the related works which were studied before for breast cancer. Also, the k-mean clustering model was the most suitable when k=7 (lowest entropy value).**

Keywords—Breast Cancer, Invasive Ductal Carcinoma (IDC), Image Processing, Supervised Learning, Unsupervised Learning.

I. INTRODUCTION

The aim of the project is to determine cancerous and non-cancerous images taken from various tissues. The tissues contain microscope images of IDC or non-IDC cancer patients, and the data set consists of photographs. Breast cancer is a type of cancer that develops from breast tissue. Breast cancer is a tumour that occurs due to the change and uncontrolled growth of the cell group that forms the breast tissue [1]. Symptoms of breast cancer may include a lump, a change in breast shape, and a skin rash. Among the risk factors of breast cancer, there are reasons such as being a woman, alcohol and cigarette use, and hormone treatments. There are many types of breast cancer. One of them is Invasive ductal carcinoma (IDC). Invasive ductal carcinoma (IDC) is a type of breast cancer that begins to grow in the milk duct [2]. According to the literature review, one study was found by type of ductal carcinoma in situ (DCIS) [3]. However, no study suitable for the invasive ductal carcinoma IDC type was found.

Machine learning is the subject of artificial intelligence used to create algorithms that develop automatically through data. algorithms are trained to analyse and predict according to the data. The accuracy of the predictions is directly proportional to the algorithm and data size.

In this study, to predict and detect IDC's tissues, RGB values of images have been studied. Supervised models were tried to develop with Support Vector Machine, k-NN, Decision Tree, and Artificial Neural Network algorithms which classify the breast cancer tissue images and invasive ductal carcinoma tissue images. Also, an unsupervised algorithm, k-means, was used to cluster, IDC and non-IDC instances.

II. RELATED WORK

In fact, large enough medical datasets and adequate learning algorithms have been available for decades. There are also thousands of articles applying machine learning algorithms to medical data. However, few of them can be said to contribute significantly to clinical care [4].

Breast cancer is an extremely common type of cancer among women. Therefore, many researches are carried out on the diagnosis of breast cancer with various image processing and classification methods. Scanning images of cancerous tissue is the easiest way to diagnose the disease. According to the literature search, a study was conducted by taking images from the Digital Database for Screening Mammography, which includes 3000 cases. In this study, images are grouped by ROI as harmless, ordinary, and threatening. Then, cancer detection is done with Principal Component Analysis (PCA) and Back Propagation algorithm (Neural Network) algorithms [5].

Depending on two of the studies in the literature, the SVM algorithm has proven its effectiveness in breast cancer prediction and diagnosis. It is also said to have the best performance in terms of precision and low error rate in the same studies. [6][7] According to another study, ANN has become the most common prediction technique used in the medical field compared to traditional methods such as decision trees [8].

III. APPROACH

The data sets contain samples of breast cancer scanned at 40x. There are a total of 277,574 photographs of data, including 198,738 non-IDCs and 78,786 IDCs. The filename of each patch is in the format: uxXyYclassC.png. Where u is the patient ID (10274idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is IDC.

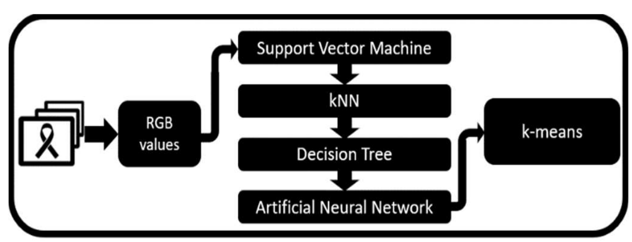


Figure.1. Flow diagram for path of experiments.

A. RGB values

Computers see pictures as array of numbers, namely RGB values for each pixel. Since we were grappling with images, we have directly decided to make use of those RGB values.

B. Classification Models

A classification model is based on estimating the value of many outcomes with one or more inputs. These results can be applied to a dataset.

i. ***Support Vector Machine***

One of the supervised machine learning techniques is the support vector machine abbreviated as SVM, which can be applied to classification and regression problems. It is included in the Scikit-learn collection. It is, however, mostly employed to solve categorization issues. Each data item is plotted as a point in n-dimensional space, where n is the number of features and the value of each feature is the value of a certain coordinate, in this SVM technique. The data is then classified by identifying the hyper-plane that clearly distinguishes the two groups. Gamma, C, and Kernel are the three parameters of SVM.

ii. k-NN

The supervised machine learning method k-nearest neighbours (KNN) may be used to tackle both classification and regression issues. The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data. During the training phase, the KNN algorithm simply stores the dataset, and when it receives new data, it classifies it into a category that is quite similar to the new data.

Choose the Kth neighbour's number. Determine the Euclidean distance between K neighbour's. Take the K closest neighbours based on the Euclidean distance computed. Count the number of data points in each category among these k neighbours. Assign the new data points to the category with the greatest number of neighbours. Our model is complete.

iii. Decision Tree

Starting with the root node, decision tree models iteratively divide all feature values at their middle points; the optimal split is determined by comparing the general entropy changes of all splits applied to the data. The best split is then chosen as the one with the lowest general entropy.

iv. Artificial Neural Network

Artificial neural network (ANN) is a special system that imitates the human brain in terms of processing and analysing data. This concept is fundamental to AI, and a solution to problems which are not possible to be solved by the human brain. Because ANNs are self-learning, they may enhance their performance as additional data becomes available. Neural Networks are analytic approaches sculpted after learning methods inside the psychological feature system and also the neurological processes of the brain, and capable of predicting new observations from previous observations when a learning process has been completed.

IV. EXPERIMENT SETUP

The data we initially collected did not require any additional preparation and was ready to be used. Since we were grappling with images, we have directly decided to make use of those RGB values.

As assessment criteria, we picked accuracy, precision, recall, and f1 scores. Because we have an imbalanced dataset, we have used the selected class as the positive label for each metric, segregating evaluation metrics between classes.

Precision and recall are also extremely helpful as a measure of success comparing models because of this mismatch. The percentage of samples correctly categorized in the positive class is referred to as precision. Recall, on the other hand, refers to the total number of properly identified positive samples, whereas the f1 score is the weighted average of accuracy and recall [9].

V. EXPERIMENTAL RESULTS AND DISCUSSION

Table.1. Results of supervised algorithms.

#	Algorithm	Accuracy	Precision	Recall	F1-macro	F1-micro	Description
1	k-NN	0.8	0.37	0.59	0.52	0.51	80% training set, from dataset-1
2	k-NN	0.8	0.36	0.58	0.53	0.52	90% training set, from dataset-1
3	k-NN	0.76	0.68	0.50	0.52	0.51	80% training set, from dataset-2
4	k-NN	0.76	0.70	0.50	0.53	0.52	90% training set, from dataset-2
5	SVM	0.82	0.82	0.57	0.68	0.68	Dataset-1
6	SVM	0.81	0.78	0.56	0.65	0.65	Dataset-2
7	ANN	0.74	0.32	0.52	0.4	0.4	80% training set
8	ANN	0.72	0.31	0.52	0.39	0.39	90% training set
9	Decision Tree	0.85	0.4	0.61	-	-	10K Images

Table.1 shows results of 4 different classification algorithms that we used for classifying the data. We calculated some evaluation metrics such as Accuracy, Precision, Recall and F1 score. All algorithms gave us nearly close results to each other but some of them may stand out about some metrics. We can say that Decision Tree gave us the best Accuracy rate that is 0.85. SVM follows the Decision Tree about Accuracy with 0.82. SVM looks clearly ahead about the precision metric that is 0.82 and 0.78 with different created models. In addition to this, SVM also has the best F1 score. Depending on all of these metrics, we can easily say that SVM is our best algorithm for this project. If we consider that the Decision Tree algorithm's result row has some missing metrics, we can choose k-NN for our second algorithm that gave good results.

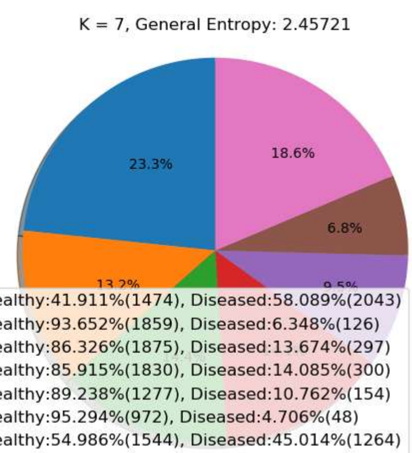


Figure.2. Best clustering (k=7) pie-chart.

In order to create a pie chart, a total of 15063 data were processed with the written codes. Of these data, 10831 show nan-IDC data, while 4232 show IDC data. With the k-means algorithms, clustering was made from 2 to 8 depending on entropy. Accordingly, the lowest entropies were chosen. In the pie chart, the highest rate is the blue cluster with 23.3%. 58.089% of this blue cluster belongs to IDC data. The most suitable for clustering our data set is the one with the lowest entropy. When the entropy values are compared, the lowest entropy value is in the k=7 cluster. In short, k=7 is best.

VI. CONCLUSIONS

Breast cancer detection is an important issue for medical care. This project helps to spot tissues with IDC by using several algorithms which are k-NN, SVM, ANN, and Decision Tree. All algorithms are implemented and tested carefully. The best resulting algorithm was SVM with precision of 0.82.

In this project, we mainly tackled with images, and consequently RGB values were the key to conclude this project. Once we dived deeper into the dataset, we discovered that cancerous tissues and non-cancerous tissues have their own special colours which are magenta and pink, respectively. Focusing on that characteristic property of tissues led us to successive conclusions about spotting IDC. To sum up, it can ultimately be said that cancerous tissues appear as being magenta while non-cancerous tissues do as be pink.

REFERENCES

[1] Gøtzsche, P. C., & Jørgensen, K. J. (2013). Screening for breast cancer with mammography. The Cochrane database of systematic reviews, 2013(6), CD001877. https://doi.org/10.1002/14651858.CD001877.pub5

[2] Sinn, HP; Kreipe, H (May 2013). "A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition". Breast Care (Basel, Switzerland). 8 (2): 149–154.

[3] Mojahed, D., Ha, R. S., Chang, P., Gan, Y., Yao, X., Angelini, B., … Hendon, C. P. (2019). Fully Automated Postlumpectomy Breast Margin Assessment Utilizing Convolutional Neural Network Based Optical Coherence Tomography Image Classification Method. Academic Radiology.

[4] Deo, R. C. (2015). Machine learning in medicine. Circulation, 132(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[5] Prannoy Giri & K Sarava. (2017). Breast Cancer Detection using Image Processing Techniques. ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY Pgs. 391-399

[6] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, 83, 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224

[7] Osareh, A., & Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. 2010 5th International Symposium on Health Informatics and Bioinformatics, HIBIT 2010, 114–120. https://doi.org/10.1109/HIBIT.2010.5478895

[8] S.Kharya, D. Dubey, S. Soni. (2013). Predictive Machine Learning Techniques for Breast Cancer Detection. International Journal of Computer Science and Information Technologies, Vol. 4 (6) , 1023-1028.

[9] Goutte, C., & Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. Lecture Notes In Computer Science, 345-359.

APPENDIX

DELIVERY #1

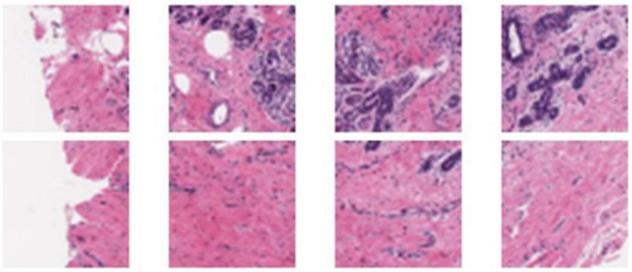


Figure.3. non-IDC images for patient 10274_idx5.

Table.2. non-IDC data for patient 10274_idx5.

Patient ID	x-coordinate	y-coordinate	Class
10274_idx5	101	951	0
10274_idx5	101	1001	0
10274_idx5	101	1051	0
10274_idx5	101	1101	0
10274_idx5	151	951	0
10274_idx5	151	1001	0
10274_idx5	151	1051	0
10274_idx5	151	1101	0

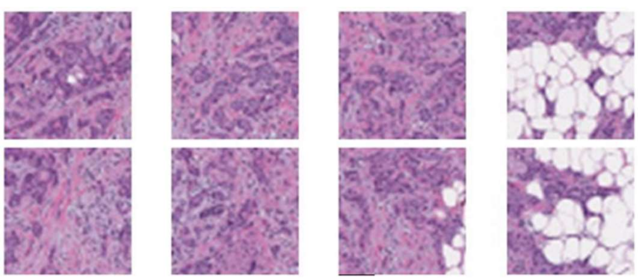


Figure.4. IDC images for patient 10274_idx5.

Table.3. IDC data for patient 10274_idx5.

Patient ID	x-coordinate	y-coordinate	Class
10274_idx5	1351	601	1
10274_idx5	1351	651	1
10274_idx5	1351	701	1
10274_idx5	1351	751	1
10274_idx5	1401	601	1
10274_idx5	1401	651	1
10274_idx5	1401	701	1
10274_idx5	1401	751	1

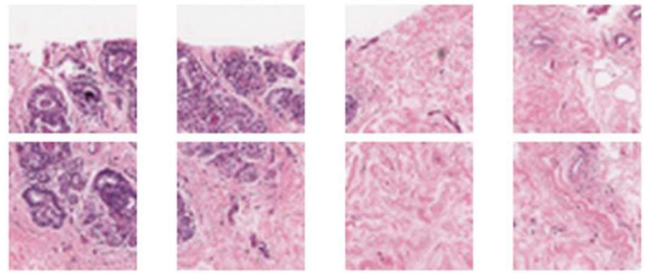


Figure.5. non-IDC images for patient 12751_idx5.

Table.4. non-IDC data for patient 12751_idx5.

Patient ID	x-coordinate	y-coordinate	Class
12751_idx5	251	1251	0
12751_idx5	251	1301	0
12751_idx5	251	1351	0
12751_idx5	251	1401	0
12751_idx5	301	1251	0
12751_idx5	301	1301	0
12751_idx5	301	1351	0
12751_idx5	301	1401	0

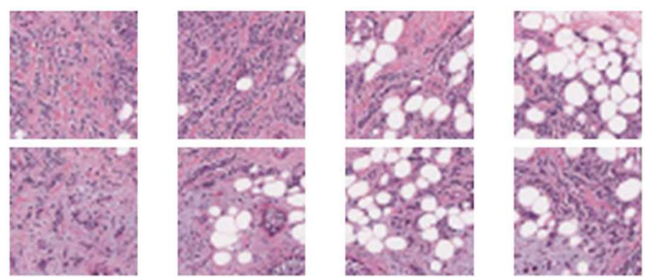


Figure.6. IDC images for patient 12751_idx5.

Table.5. IDC data for patient 12751_idx5.

Patient ID	x-coordinate	y-coordinate	Class
12751_idx5	451	1901	1
12751_idx5	451	1951	1
12751_idx5	451	2001	1
12751_idx5	451	2051	1
12751_idx5	501	1901	1
12751_idx5	501	1951	1
12751_idx5	501	2001	1
12751_idx5	501	2051	1

DELIVERY #2

Table.6. Examples of output values

Class	ID	x Coordinates	y Coordinates	Red Minimum Value	Red Maximum Value	Red Mean	Red Standard Deviation	Green Minimum Value	Green Maximum Value	Green Mean	Green Standard Deviation	Blue Minimum Value	Blue Maximum Value	Blue Mean	Blue Standard Deviation	Enthropy
0	16568	1001	1001	121	246	218.292	22.619	69	243	168.119	32.505	90	244	190.316	23.982	7.962
0	16568	1001	1051	130	243	225.761	11.435	75	236	162.552	19.708	98	240	185.290	15.772	7.227
0	16568	1001	1101	142	245	224.022	14.361	66	244	165.313	41.884	97	244	186.373	31.257	7.759
0	16568	1001	1151	121	247	233.461	20.058	52	244	219.724	46.196	90	244	224.666	33.731	4.667
0	16568	1001	251	44	247	227.897	37.156	10	247	218.964	48.641	10	246	216.720	49.579	6.731
0	16568	1001	801	166	246	230.838	10.102	72	246	185.990	41.894	87	244	201.572	30.733	7.499
0	16568	1001	851	168	241	228.200	8.255	63	230	159.525	21.144	73	234	182.387	20.025	7.020
0	16568	1001	901	172	243	231.362	7.374	106	216	168.839	13.626	130	222	190.767	10.914	6.623
0	16568	1001	951	155	244	229.284	9.553	69	236	162.883	23.899	86	239	185.104	21.980	7.122
0	16568	101	851	116	246	220.263	15.610	57	245	153.874	48.032	49	245	178.264	35.010	7.612
0	16568	101	901	104	247	215.073	16.017	31	246	138.055	43.988	26	244	166.903	32.567	7.734
0	16568	101	951	104	247	230.870	19.646	42	245	203.723	59.377	49	243	212.524	43.920	5.223
0	16568	1051	1001	137	244	221.976	14.677	75	240	167.892	25.283	91	241	190.348	19.321	7.644
0	16568	1051	1051	117	244	221.542	18.104	78	242	184.466	30.115	107	243	202.425	21.865	7.781
0	16568	1051	1101	122	242	217.787	15.692	62	233	146.226	23.256	80	237	171.636	19.447	7.614
0	16568	1051	1151	97	245	214.238	18.527	37	246	149.096	39.522	72	243	174.557	29.394	8.052
0	16568	1051	1201	78	246	218.601	28.827	37	245	184.398	63.894	75	245	199.524	45.931	6.685
0	16568	1051	751	145	247	233.536	9.561	69	246	201.745	40.866	86	245	212.867	30.012	7.285
0	16568	1051	801	171	243	227.736	6.172	71	243	154.526	14.721	89	242	179.215	12.463	6.761
0	16568	1051	851	137	241	229.856	6.669	74	226	163.184	14.098	104	230	186.244	11.424	6.678
0	16568	1051	901	127	241	224.807	16.608	70	226	160.880	17.912	91	231	184.651	13.854	7.122
0	16568	1051	951	109	244	222.265	18.970	51	240	164.698	23.229	12	242	187.892	18.664	7.482
1	16568	1301	551	141	249	227.866	16.962	70	246	193.267	47.066	82	245	206.754	35.145	7.707
1	16568	1301	601	155	246	223.629	9.301	84	242	143.938	21.675	109	242	170.503	17.063	6.998
1	16568	1301	651	128	238	221.473	14.424	78	224	145.624	14.275	121	229	172.821	11.632	6.965
1	16568	1301	701	116	241	208.036	23.279	77	234	142.529	19.653	125	234	171.574	14.650	7.630
1	16568	1351	501	124	247	220.390	25.433	75	246	193.470	45.378	102	245	208.325	31.772	7.880
1	16568	1351	551	98	243	199.839	22.976	51	239	141.580	25.518	90	241	171.551	19.731	7.955
1	16568	1351	601	106	244	199.918	23.309	59	241	145.219	25.370	103	240	175.332	18.327	7.908
1	16568	1351	651	121	240	209.592	20.616	77	237	143.860	18.498	122	239	173.053	14.156	7.477
1	16568	1351	701	109	242	205.276	21.358	76	232	143.758	17.486	124	235	173.504	12.847	7.456
1	16568	1351	751	127	244	218.992	17.648	82	240	154.636	26.529	107	242	179.082	20.012	7.614
1	16568	1401	1251	122	241	203.806	18.466	71	230	136.667	16.129	118	233	167.705	12.086	7.392
1	16568	1401	1301	133	241	203.512	13.503	69	231	128.195	17.905	106	233	160.238	14.366	7.383
1	16568	1401	501	114	245	201.189	23.939	65	244	151.118	36.278	91	243	178.530	26.456	8.204
1	16568	1401	551	114	244	189.771	25.003	65	242	143.269	33.898	86	242	174.412	24.143	8.211
1	16568	1401	601	112	244	184.058	24.618	61	240	139.472	31.623	105	241	172.401	22.472	8.171
1	16568	1401	651	97	242	192.301	24.533	65	239	139.862	24.726	113	239	171.906	17.846	7.923
1	16568	1401	701	127	239	198.228	21.722	82	233	140.028	19.676	125	235	171.360	14.571	7.669
1	16568	1401	751	151	244	222.650	9.412	99	241	149.127	19.538	125	241	174.313	15.607	7.136
1	16568	1451	1251	110	239	204.974	16.921	69	236	145.886	16.390	114	237	174.917	12.194	7.341
1	16568	1451	1301	139	235	208.827	12.844	64	218	138.974	15.082	87	228	168.705	12.540	7.190
1	16568	1451	401	118	248	227.784	20.130	60	247	202.752	43.773	80	245	213.954	31.843	7.583
1	16568	1451	451	103	246	205.822	29.400	70	244	170.368	46.885	108	245	192.624	32.833	8.339
1	16568	1451	501	120	245	194.465	22.751	71	243	141.796	29.077	116	243	173.050	21.400	8.044
1	16568	1451	551	90	244	193.202	25.141	48	241	141.218	32.885	94	243	172.265	24.056	8.210
1	16568	1451	601	107	245	180.629	28.232	55	240	138.221	38.011	104	243	171.229	26.599	8.275
1	16568	1451	651	105	245	182.512	26.152	52	241	136.830	32.078	102	242	169.871	22.969	8.226
1	16568	1451	701	107	243	197.959	25.236	67	238	138.902	24.262	111	240	169.778	18.095	7.945
1	16568	1451	751	127	240	222.078	8.386	100	230	142.591	13.602	136	230	168.788	11.428	6.750
1	16568	1451	801	126	243	217.545	18.724	81	240	148.240	21.269	130	242	174.591	16.047	7.439

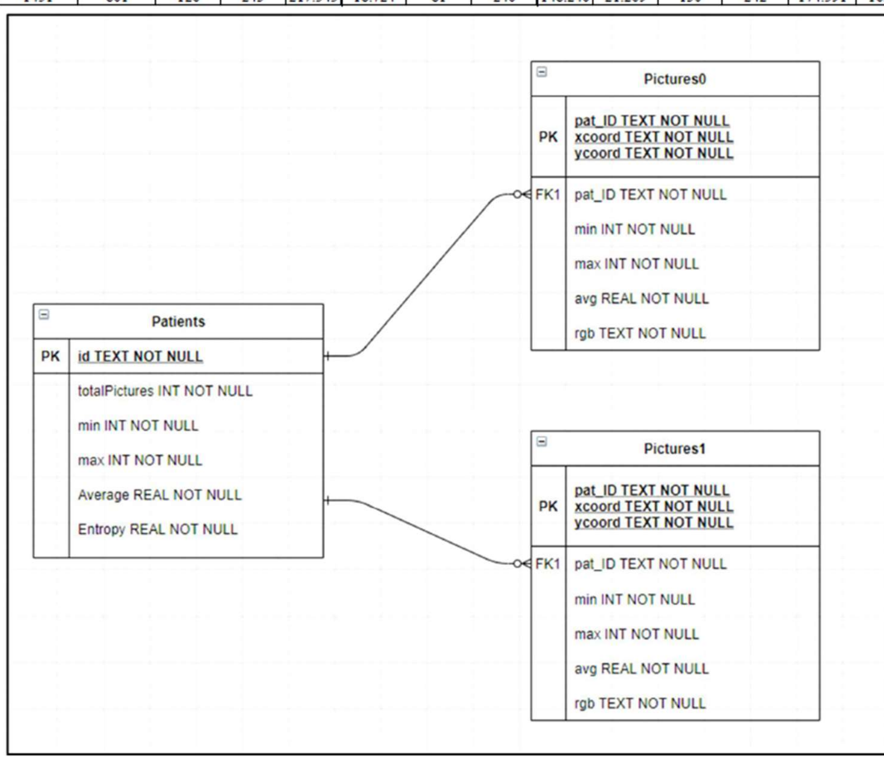


Figure.7. The EER Diagram of Databased named IDC.

DELIVERY #3

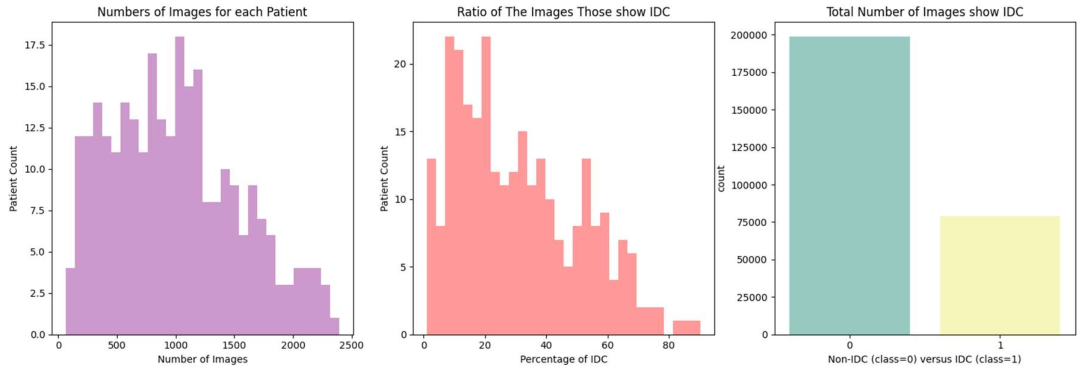


Figure.8. General Aspects of Dataset

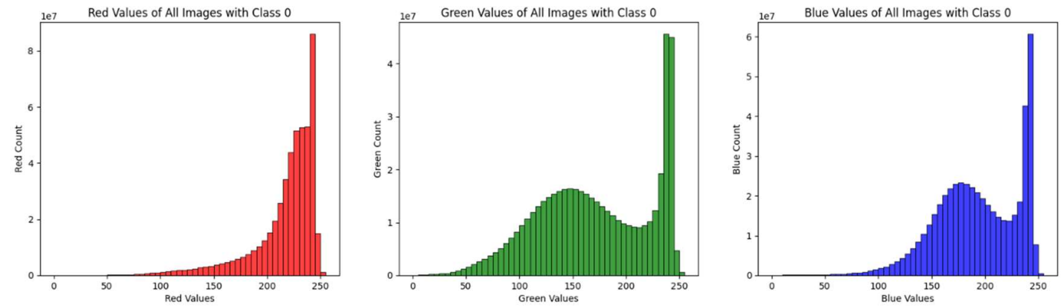


Figure.9. Counts versus values of red, green, and blue graphs for class 0.

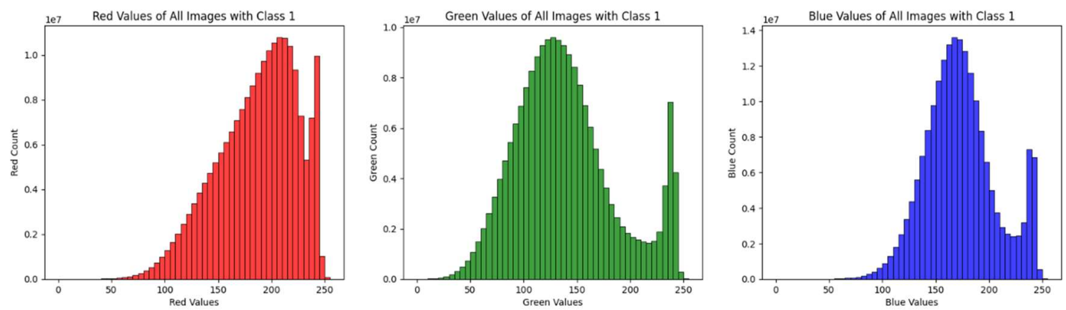


Figure.10. Counts versus values of red, green, and blue graphs for class 1.

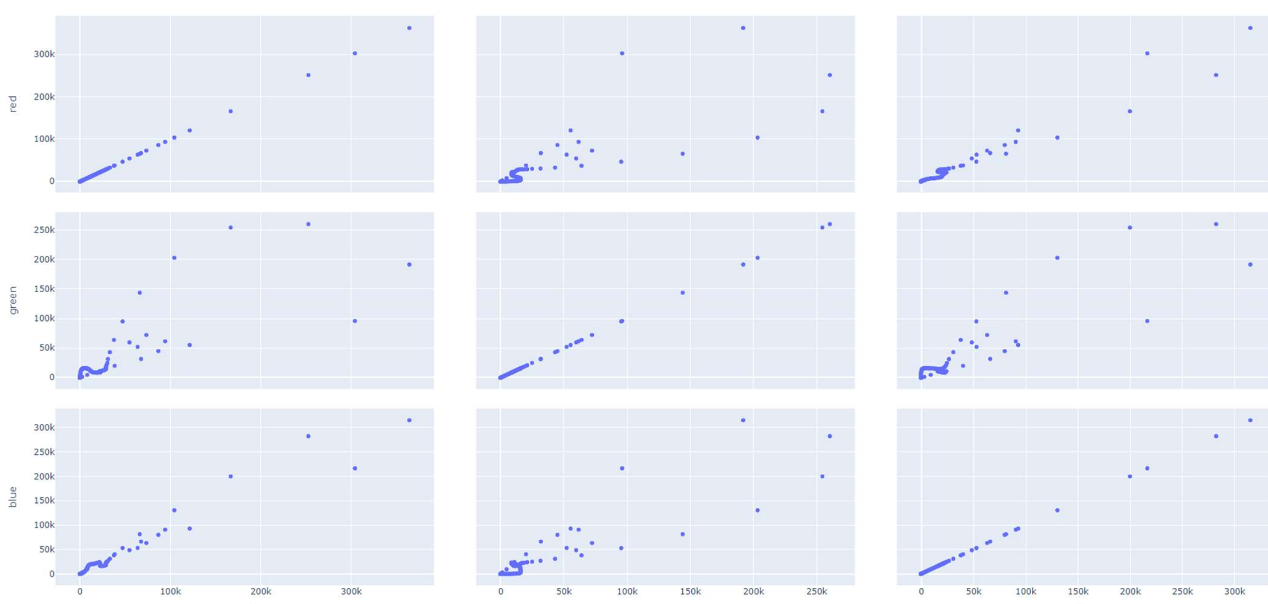


Figure.11. Scatter plot matrix for red, green, and blue attributes.

DELIVERY #4

Table.7. Evaluation Metrics

#	Algorithm	Accuracy	Precision	Recall	F1-macro	F1-micro	Description
1	k-NN	0.8	0.37	0.59	0.52	0.51	80% training set, from dataset-1
2	k-NN	0.8	0.36	0.58	0.53	0.52	90% training set, from dataset-1
3	k-NN	0.76	0.68	0.50	0.52	0.51	80% training set, from dataset-2
4	k-NN	0.76	0.70	0.50	0.53	0.52	90% training set, from dataset-2
5	SVM	0.82	0.82	0.57	0.68	0.68	Dataset-1
6	SVM	0.81	0.78	0.56	0.65	0.65	Dataset-2
7	ANN	0.74	0.32	0.52	0.4	0.4	80% training set
8	ANN	0.72	0.31	0.52	0.39	0.39	90% training set
9	Decision Tree	0.85	0.4	0.61	-	-	10K Images

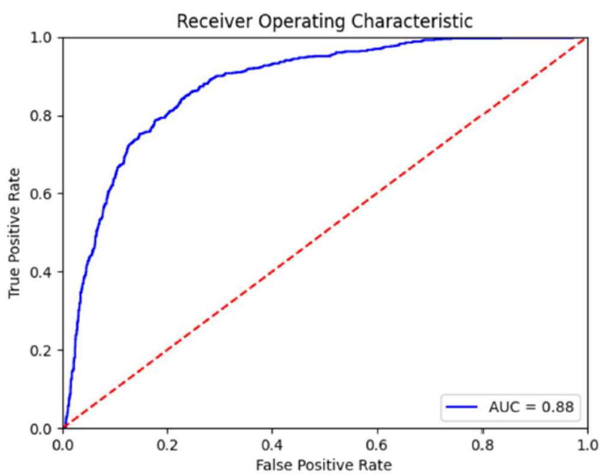


Figure.12. ROC Curve for SVM with Dataset_1.

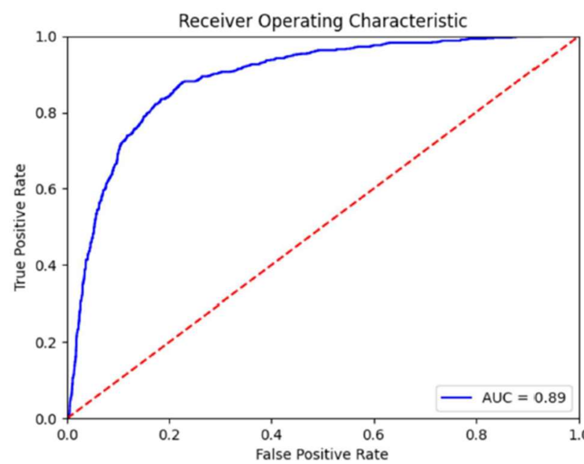
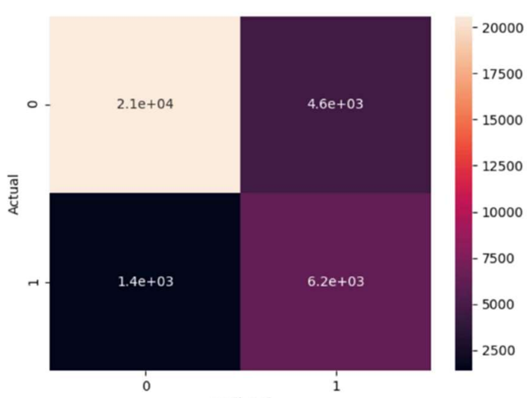


Figure.13. ROC Curve for SVM with Dataset_2



. Figure.14. Confusion Matrix of SVM with Dataset_1

Table.8. Confusion Matrix Parameters for Figure.14.

True Negative	True Positive	False Negative	False Positive
20611	6250	4635	1378

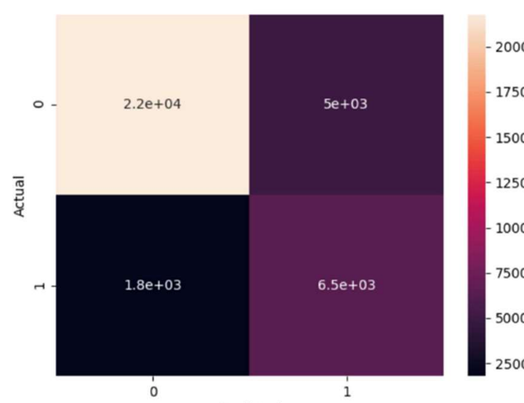


Figure.15. Confusion Matrix of SVM with Dataset_2.

Table.9. Confusion Matrix Parameters for Figure.15.

True Negative	True Positive	False Negative	False Positive
20611	6250	4635	1378

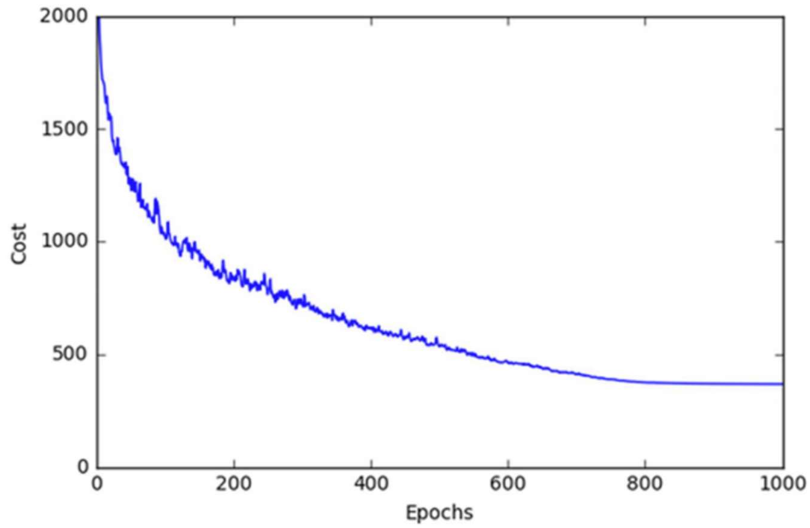


Figure.16. Epoch vs cost plot for ANN.

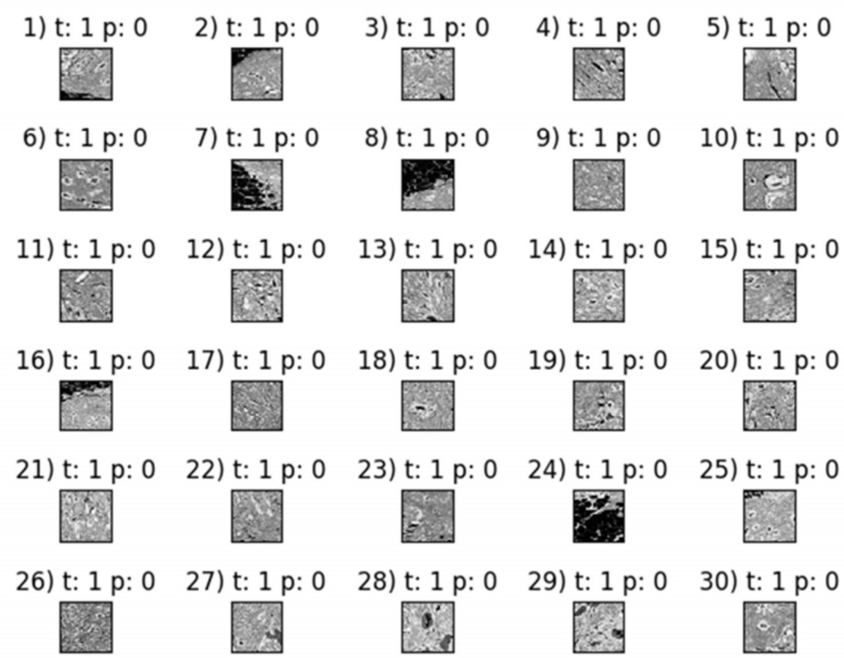


Figure.17. Most wrongly predicted images.

DELIVERY #5

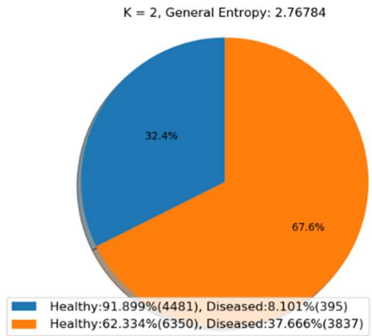


Figure.18. Clusters generated with k-means algorithm with k value 2.

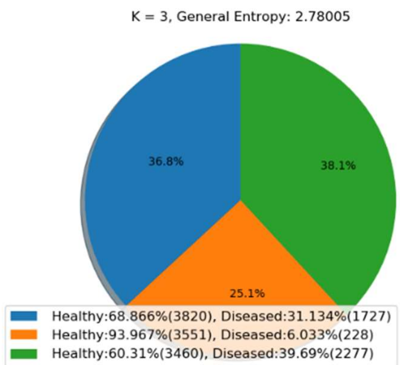


Figure.19. Clusters generated with k-means algorithm with k value 3.

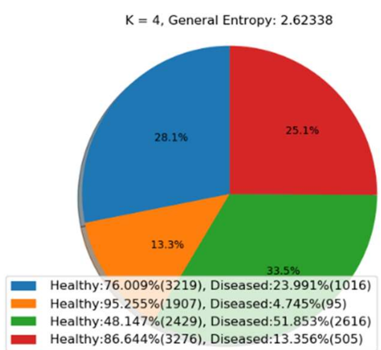


Figure.20. Clusters generated with k-means algorithm with k value 4.

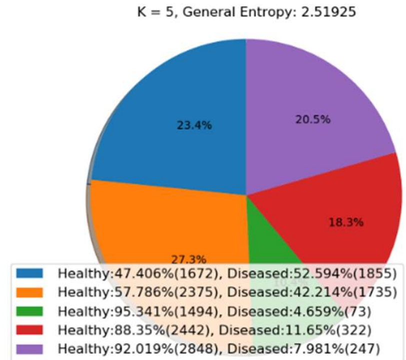


Figure.21. Clusters generated with k-means algorithm with k value 5.

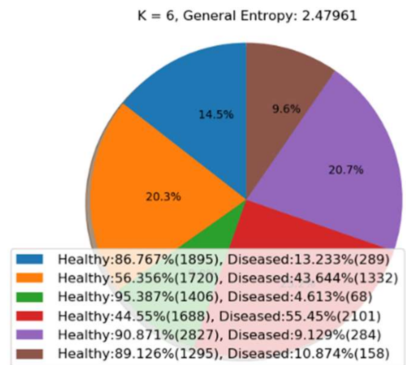


Figure.22. Clusters generated with k-means algorithm with k value 6.

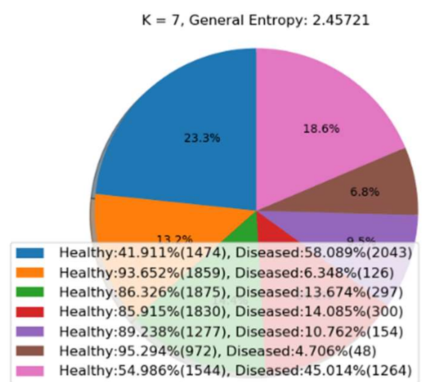


Figure.23. Clusters generated with k-means algorithm with k value 7.

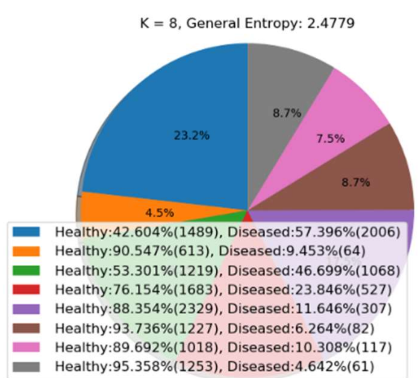


Figure.24. Clusters generated with k-means algorithm with k value 8.

Table.10. Summary of the Pie-charts.

k=2	**General Entropy=2.76784**															
	Blue (%)		**Orange (%)**													
	32.4		67.6													
	non-IDC	**IDC**	**non-IDC**	**IDC**												
	91.899	8.101	62.334	37.666												
k=3	**General Entropy=2.78005**															
	Blue (%)		**Orange (%)**		**Green (%)**											
	36.8		25.1		38.1											
	non-IDC	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**										
	68.866	31.134	93.967	6.033	60.31	39.69										
k=4	**General Entropy=2.62338**															
	Blue (%)		**Orange (%)**		**Green (%)**		**Red (%)**									
	28.1		13.3		33.5		25.1									
	non-IDC	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**								
	76.009	23.991	95.255	4.745	48.147	51.853	86.644	13.356								
k=5	**General Entropy=2.51925**															
	Blue (%)		**Orange (%)**		**Green (%)**		**Red (%)**		**Purple (%)**							
	23.4		27.3		10.4		18.3		20.5							
	non-IDC	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**						
	47.406	52.594	57.786	42.214	95.341	4.659	88.35	11.65	92.019	7.981						
k=6	**General Entropy=2.47961**															
	Blue (%)		**Orange (%)**		**Green (%)**		**Red (%)**		**Purple (%)**		**Brown (%)**					
	14.5		20.3		9.7		25.2		20.7		9.6					
	non-IDC	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**				
	86.767	13.233	56.356	43.644	95.387	4.613	44.55	55.45	90.871	9.129	89.126	10.874				
k=7	**General Entropy=2.45721**															
	Blue (%)		**Orange (%)**		**Green (%)**		**Red (%)**		**Purple (%)**		**Brown (%)**		**Pink (%)**			
	23.2		13.2		14.4		14.2		9.5		6.8		18.6			
	non-IDC	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**		
	41.911	58.089	93.652	6.348	86.326	13.674	85.915	14.085	89.238	10.762	95.294	4.706	54.986	45.014		
k=8	**General Entropy=2.4779**															
	Blue (%)		**Orange (%)**		**Green (%)**		**Red (%)**		**Purple (%)**		**Brown (%)**		**Pink (%)**		**Gray (%)**	
	23.2		4.5		15.2		14.7		17.5		8.7		7.5		8.7	
	non-IDC	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**	**non-IDC**	**IDC**
	42.604	57.396	90.547	9.453	53.301	46.699	76.154	23.846	88.354	11.646	93.736	6.264	89.692	10.308	95.358	4.642