

FINAL YEAR PROJECT REPORT

PROJECT - SUPERVISED MACHINE LEARNING APPROACHES FOR CANCER CLASSIFICATION USING GENE EXPRESSION DATA.

DATASET 1 :- Cancer gene expression datasets presented in the paper [leukemia \(Golub et al.\)](#)

FEATURES :- **Number of genes:** 5147
Number of samples: 72

DIAGNOSTIC CLASSES:

Acute lymphoblastic leukemia (**ALL**): 47 examples (65.3%)
Acute myeloid leukemia (**AML**): 25 examples (34.7%)

FEATURE SELECTION: Top 20 gene features having maximum cancerous cells based on PEARSON'S CORRELATION.

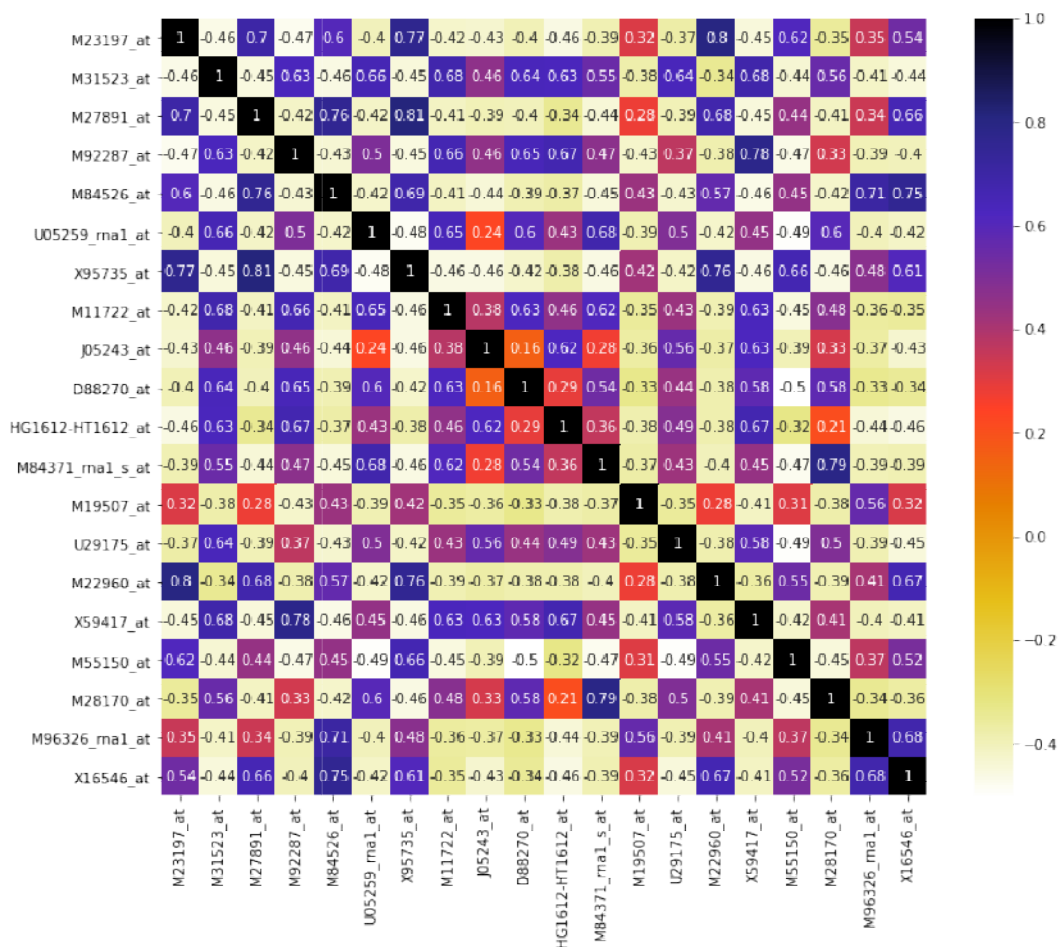


Diagram 1: Correlation diagram of 20 radviz visualizations with 8 attributes.

TRAINING DATA - TESTING DATA SPLIT:

1. Training Data : Testing Data = 1 : 1 i.e. 50-50 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	100%	100%	100%	100%
Support Vector Machine (SVM)	97.22%	100%	100%	100%	100%
Naive Bayes Classifier	100%	100%	100%	100%	100%
eXtreme Gradient Boosting	100%	94.44%	90.91%	90.91%	90.91%
Random Forest Classifier	100%	100%	100%	100%	100%

2. Training Data : Testing Data = 7 : 3 i.e. 70-30 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	100%	100%	100%	100%
Support Vector Machine (SVM)	96%	100%	100%	100%	100%
Naive Bayes Classifier	100%	100%	100%	100%	100%
eXtreme Gradient Boosting	100%	100%	100%	100%	100%
Random Forest Classifier	100%	95.45%	85.71%	100%	92.31%

3. Training Data : Testing Data = (8:2) 4 : 1 i.e. 80-20 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	100%	100%	100%	100%
Support Vector Machine (SVM)	96.49%	100%	100%	100%	100%
Naive Bayes Classifier	100%	100%	100%	100%	100%
eXtreme Gradient Boosting	100%	100%	100%	100%	100%
Random Forest Classifier	100%	93.33%	80%	100%	88.89%

DATASET 2 :- Cancer gene expression datasets presented in the paper
[SRBCT \(Khan et al.\)](#)

FEATURES :- Number of genes: 2308
Number of samples: 83

DIAGNOSTIC CLASSES:

Ewing's sarcoma (EWS): 29 examples (34.9%)
Burkitt's lymphoma (BL): 11 examples (13.3%)

Neuroblastoma (NB): 18 examples (21.7%)
Rhabdomyosarcoma (RMS): 25 examples (30.1%)

FEATURE SELECTION: Top 20 gene features having maximum cancerous cells based on PEARSON'S CORRELATION.

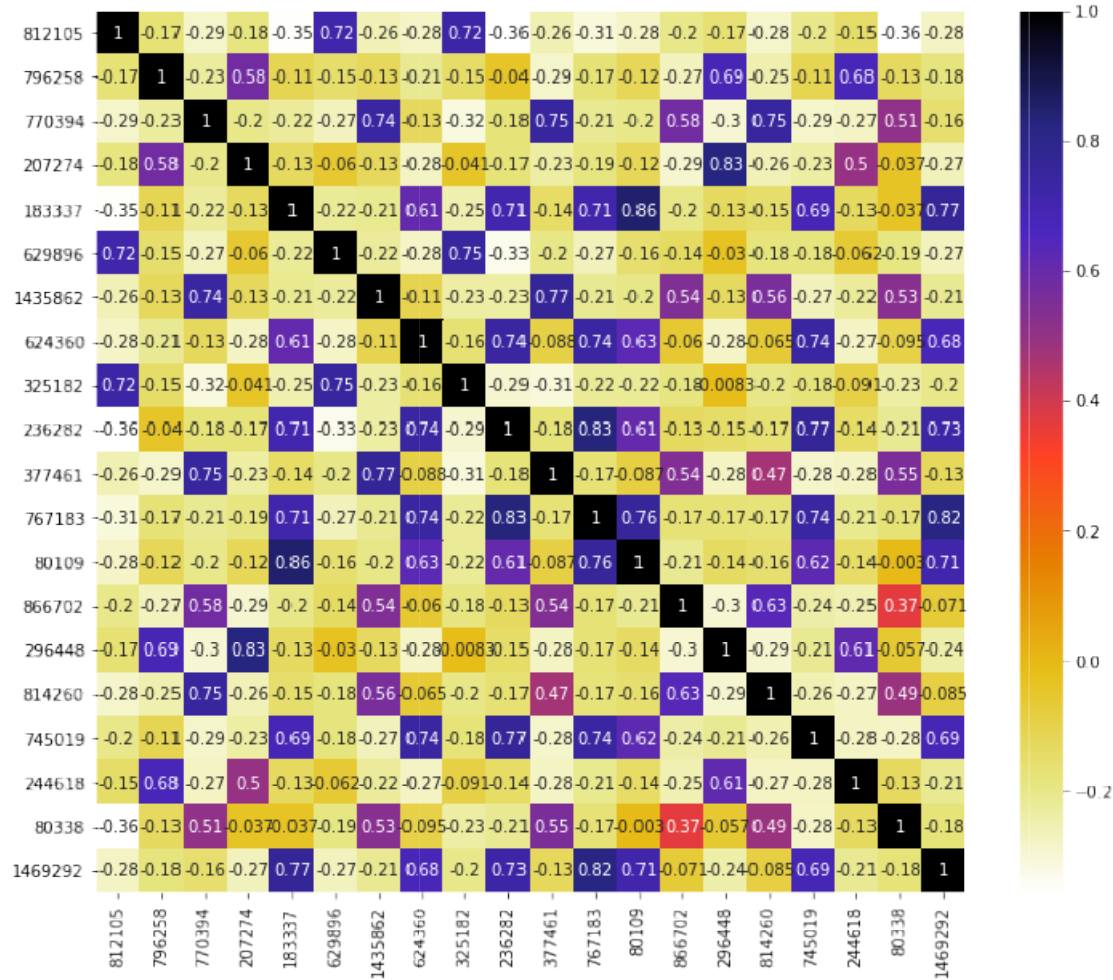


Diagram 2 : Correlation diagram of 20 radviz visualizations with 8 attributes.

TRAINING DATA - TESTING DATA SPLIT:

1. Training Data : Testing Data = 1 : 1 i.e. 50-50 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	97.56%	97.56%	97.56%	97.56%
Support Vector Machine (SVM)	100%	92.68%	94%	93%	93%
Naive Bayes Classifier	100%	100%	100%	100%	100%
eXtreme Gradient Boosting	100%	87.8%	91.01%	87.8%	87.56%
Random Forest Classifier	100%	100%	100%	100%	100%

2. Training Data : Testing Data = 7 : 3 i.e. 70-30 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	96%	96%	96%	96%
Support Vector Machine (SVM)	100%	96%	96.44%	96%	95.9%
Naive Bayes Classifier	100%	100%	100%	100%	100%
eXtreme Gradient Boosting	100%	100%	100%	100%	100%
Random Forest Classifier	100%	100%	100%	100%	100%

3. Training Data : Testing Data = (8:2) 4 : 1 i.e. 80-20 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	94.12%	95.29%	94.12%	93.86%
Support Vector Machine (SVM)	100%	94.12%	95.29%	94.12%	93.86%
Naive Bayes Classifier	100%	94.12%	95.29%	94.12%	93.86%
eXtreme Gradient Boosting	100%	100%	100%	100%	100%
Random Forest Classifier	100%	100%	100%	100%	100%

DATASET 3 :- Cancer gene expression datasets presented in the paper
[MLL \(Armstrong et al.\)](#)

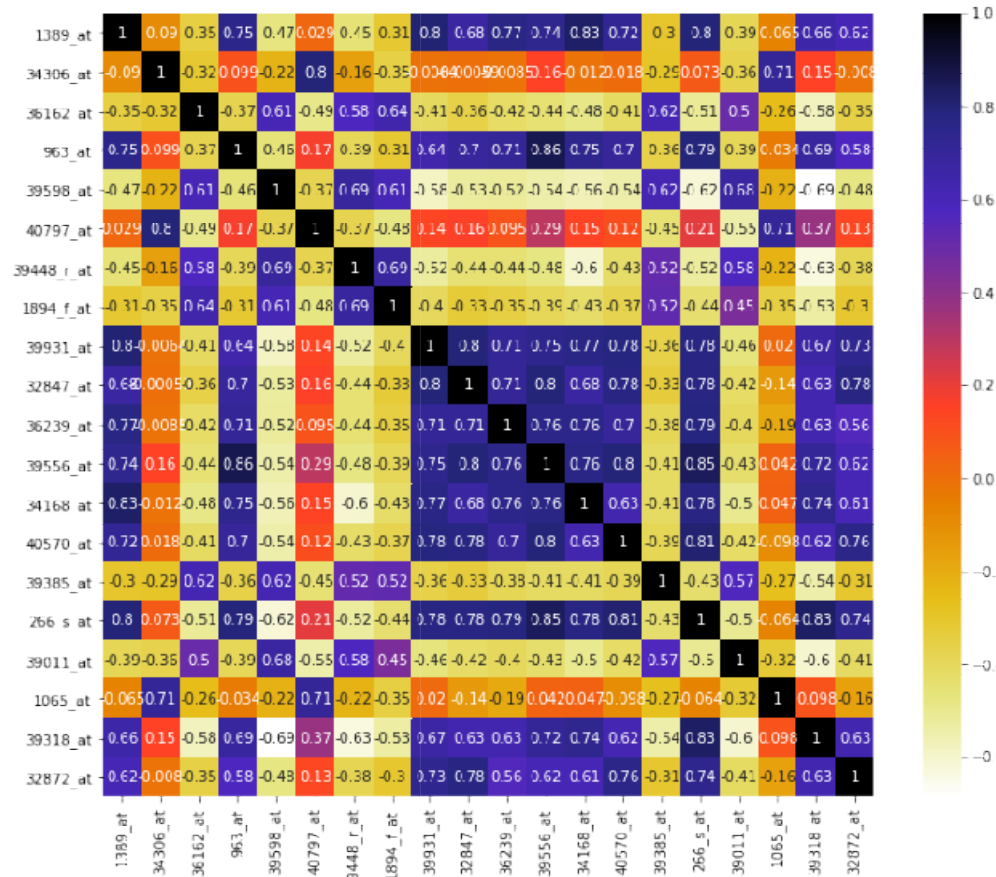
FEATURES :- Number of genes: 12533
Number of samples: 72

DIAGNOSTIC CLASSES:

Acute lymphoblastic leukemia (**ALL**): **24** examples (33.3%)
Mixed-lineage leukemia (**MLL**): **20** examples (27.8%)
Acute myeloid leukemia (**AML**): **28** examples (38.9%)

FEATURE SELECTION: Top 20 gene features having maximum cancerous cells based on
PEARSON'S CORRELATION.

Diagram 3 :- Correlation diagram of 20 radviz visualizations with 8 attributes.
(Contd....)



TRAINING DATA - TESTING DATA SPLIT:

1. Training Data : Testing Data = 1 : 1 i.e. 50-50 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	97.22%	97.47%	97.22%	97.23%
Support Vector Machine (SVM)	100%	94.44%	95%	94.44%	94.44%
Naive Bayes Classifier	100%	86.11%	86.11%	86.11%	86.11%
eXtreme Gradient Boosting	100%	88.89%	89.24%	88.89%	88.64%
Random Forest Classifier	100%	91.67%	91.67%	91.67%	91.67%

2. Training Data : Testing Data = 7 : 3 i.e. 70-30 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	95.45%	96%	95.45%	96%
Support Vector Machine (SVM)	98%	95.45%	96.21%	95.45%	95.53%
Naive Bayes Classifier	100%	95.45%	96.21%	95.45%	95.51%
eXtreme Gradient Boosting	100%	95.45%	95.96%	95.45%	95.34%
Random Forest Classifier	100%	95.45%	96.21%	95.45%	95.51%

3. Training Data : Testing Data = (8:2) 4 : 1 i.e. 80-20 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	93.33%	95%	93.33%	93.55%
Support Vector Machine (SVM)	98.25%	93.33%	95%	93.33%	93.55%
Naive Bayes Classifier	100%	93.33%	95%	93.33%	93.44%
eXtreme Gradient Boosting	100%	100%	100%	100%	100%
Random Forest Classifier	100%	93.33%	95%	93.33%	93.44%

DATASET 4 :- Cancer gene expression datasets presented in the paper
[DLBCL \(Shipp et al.\)](#)

FEATURES :- Number of genes: 7070
Number of samples: 77

DIAGNOSTIC CLASSES:

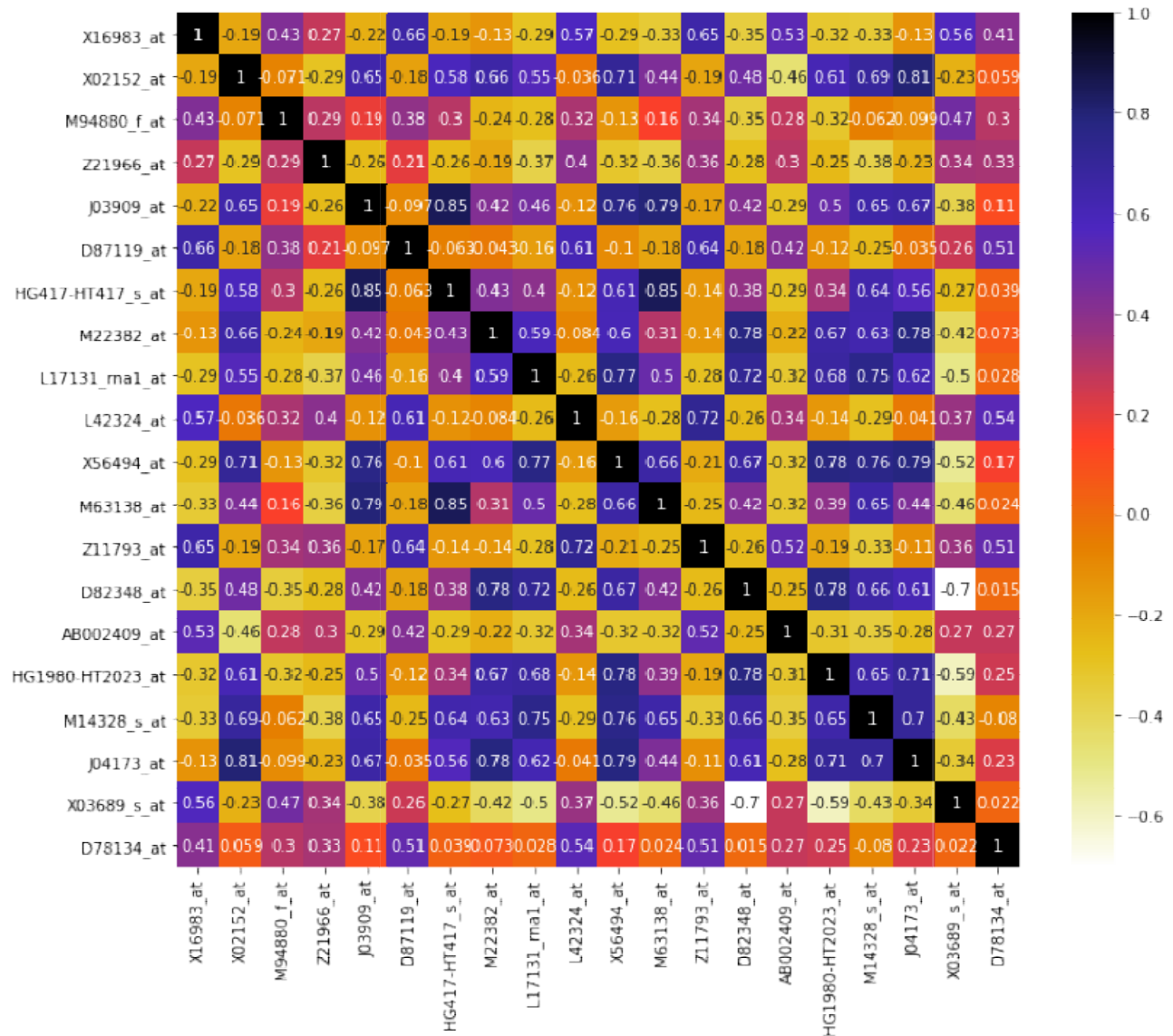
Diffuse large B-cell lymphoma (**DLBCL**): **58** examples (75.3%)
Follicular lymphoma (**FL**): **19** examples (24.7%)

FEATURE SELECTION: Top 20 gene features having maximum cancerous cells based on
PEARSON'S CORRELATION.

List of top 20 genes :

['X16983_at', 'X02152_at', 'M94880_f_at', 'Z21966_at', 'J03909_at', 'D87119_at', 'HG417-
HT417_s_at', 'M22382_at', 'L17131_rna1_at', 'L42324_at', 'X56494_at', 'M63138_at',
'Z11793_at', 'D82348_at', 'AB002409_at', 'HG1980-HT2023_at', 'M14328_s_at', 'J04173_at',
'X03689_s_at', 'D78134_at']

Diagram 4 :- Correlation diagram of 20 radviz visualizations with 8 attributes.
(Contd....)



TRAINING DATA - TESTING DATA SPLIT:

1. Training Data : Testing Data = 1 : 1 i.e. 50-50 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	92.11%	75%	100%	85.71%
Support Vector Machine (SVM)	94.87%	94.74%	81.82%	100%	90%
Naive Bayes Classifier	97.44%	94.74%	81.82%	100%	90%
eXtreme Gradient Boosting	100%	78.95%	57.14%	44.44%	50%
Random Forest Classifier	100%	92.11%	80%	88.89%	84.21%

2. Training Data : Testing Data = 7 : 3 i.e. 70-30 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	95.83%	80%	100%	88.89%
Support Vector Machine (SVM)	94.34%	95.83%	80%	100%	88.89%
Naive Bayes Classifier	96.23%	95.83%	80%	100%	88.89%
eXtreme Gradient Boosting	100%	83.33%	50%	25%	33.33%
Random Forest Classifier	100%	87.5%	66.67%	50%	57.14%

3. Training Data : Testing Data = (8:2) 4 : 1 i.e. 80-20 split

Machine Learning Algorithms	Model Score	Accuracy	Precision	Recall	F1 Score
Logistic Regression	100%	93.75%	80%	100%	88.89%
Support Vector Machine (SVM)	98.36%	93.75%	80%	100%	88.89%
Naive Bayes Classifier	96.72%	93.75%	80%	100%	88.89%
eXtreme Gradient Boosting	100%	81.25%	66.67%	50%	57.14%
Random Forest Classifier	100%	81.25%	66.67%	50%	57.14%

References:-

Datasets for Cancerous Gene Classification(Orange Dataset):

<https://file.biolab.si/biolab/supp/bi-cancer/projections/info/SRBCT.html>