PAPER REPORT:-

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

A) DATASET: - Cancer gene expression datasets presented in the paper **leukemia (Golub et al.)** Famously known as LEUKEMIA DATASET.

B) FEATURES :- Number of genes: 5147
Number of samples: 72

C) DIAGNOSTIC CLASSES:

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

D) FEATURE SELECTION TECHNIQUES:

I have used **FOUR** feature selection techniques to find out the **best 30 features** that are selected from each feature selection.

Target No. of Features:- 30

The feature selection techniques are as follows:-

1. **Chi-Squared test with SelectKBest algorithm :-**
   I have selected 30 best features using Chi-Squared test.
   NOTE: - As the actual leukemia gene dataset contained negative values, and given that Chi-Squared test doesn't work for negative inputs, I have used **Min-Max Scaler** to scale the whole dataset in the **range (0-1)**, for better efficiency.

1. **Pearson's Correlation:-**
   I have selected 30 best features using Pearson's Correlation method.

2. **Feature Importance method:-**
   Feature Importance is a process used to select features in the dataset that contributes the most in predicting the target variable.
   I have selected top 30 features having the most contributions using Feature Importance method.

3. **Information Gain:-** Information Gain can be used for feature selection, by evaluating the gain of each variable in the context of the target variable.
   I have selected 30 best features using Information Gain.

F) UNION & INTERSECTION OF DATA:

After feature selection, I had 30 features from four Feature Selection Processes. So altogether I had 120 numbers of features, with many features common as output from each of the processes.
So my next job was to find out the distinct features. So I moved into **union** and **intersection** of data.

1. **UNION of 120 features:-**
   After performing union among these 120 features, I got distinct 63 features.
   So, my final processed dataset looks like
   No. of rows: - **72**
   No. of columns: - 63 features + 1 label = **64**
   Shape of the dataset: (72, 64)

2. **INTERSECTION of 120 features:-**
   After performing intersection among these 120 features, I got distinct 7 features.
   So, my final processed dataset looks like
   No. of rows:- **72**
   No. of columns: - 7 features + 1 label = **8**
   Shape of the dataset: (72, 8)

G) MACHINE LEARNING CLASSIFIERS WITH OPTIMIZATION METHODS:

For both UNION dataset and INTERSECTION dataset:
Training Data : Testing Data Split = 7:3 i.e. 70-30 Split.

1. STANDARD VECTOR MACHINE (SVM):

The first machine learning classifier is **Standard Vector Machine** or **SVM**. We know it as Standard Vector Classifier (SVC). Because the dataset is small with less number of features, I have chosen **Linear Standard Vector Classifier** or **Linear SVC**. It will be applicable for both UNION and INTERSECTION dataset.
For optimization of hyperparameters in Linear SVC, I have used two optimization algorithms namely **RANDOMISED SEARCH CV** and **GRID SEARCH CV**. Using these, I have found the best hyperparameters for Linear SVC which can give the highest accuracy, both in UNION and INTERSECTION dataset.

LINEAR STANDARD VECTOR CLASSIFIER (LINEAR SVC) :

Datasets	Score Linear SVC	Accuracy Linear SVC	Linear SVC Accuracy with (RANDOMISED SEARCH CV)	Linear SVC Accuracy with (GRID SEARCH CV)
UNION	100%	86.3636%	90.90909091%	90.90909091%
INTERSECTION	100%	95.454546%	95.454546%	95.454546%

2. MULTI LAYERED PERCEPTRON (MLP):

The second machine learning classifier is **Multi Layered Perceptron** or **MLP**. It will be applicable for both UNION and INTERSECTION dataset.

Also, for optimization of hyperparameters in MLP, I have used two optimization algorithms namely **RANDOMISED SEARCH CV** and **GRID SEARCH CV**. Using these, I have found the best hyperparameters for MLP which can give the highest accuracy, both in UNION and INTERSECTION dataset.

MULTI LAYERED PERCEPTRON (MLP) :

Datasets	Training Accuracy	Testing Accuracy	Testing Accuracy with (RANDOMISED SEARCH CV)	Testing Accuracy with (GRID SEARCH CV)
UNION	100%	90.909091%	95.454546%	90.90909091%
INTERSECTION	100%	72.727273%	90.909091%	95.454546%

3. ARTIFICIAL NEURAL NETWORK (ANN) :-

The third algorithm is **Artificial Neural Network** or **ANN**. It will be applicable for both UNION and INTERSECTION dataset.

For hyperparameter optimization in ANN, I have used a library from Deep Learning framework **Keras** known as **KERAS TUNER**. It gives me the best hyperparameters in the neural metwork as well as finds out the best accuracy of the neural network model trained using those hyperparameters. For weight optimization of the neurons, I have used **ADAM** optimizer.

ARTIFICIAL NEURAL NETWORK (ANN) :

UNION DATASET:

Keras Tuner optimizer (BEST HYPERPARAMETERS and BEST ACCURACY):

No. of layers = 3(Hidden) + 2
Input Layer = No. of features(X)
First Hidden Layer = 32 Activation : Relu
Second Hidden Layer = 160 Activation : Relu
Third Hidden Layer = 32 Activation : Relu
Output Layer = Category of labels (2)
Optimizer = Adam
Learning Rate = 0.01
Number of epochs = 5 (Total Trials = 10*3 = 30)
***Accuracy* = 0.9696969588597616**

INTERSECTION DATASET:

Keras Tuner optimizer (BEST HYPERPARAMETERS and BEST ACCURACY):

No. of layers = 9(hidden) + 2
Input layer = No. of features(X)
First Hidden Layer = 208 Activation: Sigmoid
Second Hidden Layer = 16 Activation: Sigmoid
Third Hidden Layer = 80 Activation: Sigmoid
Fourth Hidden layer = 144 Activation: Relu
Fifth Hidden layer = 144 Activation: Relu
Sixth Hidden layer = 208 Activation: Sigmoid
Seventh Hidden layer = 16 Activation: Sigmoid
Eighth Hidden layer = 16 Activation: Relu
Ninth Hidden layer = 208 Activation: Sigmoid
Output Layer = Category of labels(2)

Weight Optimizer = Adam
Learning Rate = 0.001
Number of epochs = 5 (Total trials = 5*3 = 15)
Accuracy = **0.939393937587738**

Therefore, in Artificial Neural Network, using the best optimizers, if we train both the UNION dataset and the INTERSECTION dataset with their training data, we will get validation accuracies of **96.96969588597616 %** and **93.9393937587738 %** on their testing data respectively.

*****SOME NECESSARY UPDATES*****

Best Hyperparameters for Linear SVC using Randomised Search CV and Grid Search CV: (UNION DATASET)

Best Hyperparameters	Randomised Search CV	Grid Search CV
Linear Support Vector Classifier	'C': 1.0, 'loss': 'squared_hinge', 'max_iter': 5000, 'multi_class': 'crammer_singer', 'penalty': 'l1'	'C': 0.001, 'loss': 'squared_hinge', 'max_iter': 1000, 'multi_class': 'crammer_singer', 'penalty': 'l1'
Multi Layered Perceptron	'activation': 'tanh', 'alpha': 0.1, 'hidden_layer_sizes': (32, 64, 128), 'learning_rate': 'adaptive', 'max_iter': 700, 'momentum': 0.2, 'solver': 'lbfgs'	'activation': 'tanh', 'alpha': 1.0, 'hidden_layer_sizes': (32, 64, 128), 'learning_rate': 'adaptive', 'max_iter': 500, 'momentum': 0.01, 'solver': 'lbfgs'

Best Hyperparameters for Linear SVC using Randomised Search CV and Grid Search CV: (INTERSECTION DATASET)

Best Hyperparameters	Randomised Search CV	Grid Search CV
Linear Support Vector Classifier	'C': 10, 'multi_class': 'ovr', 'max_iter': 3000, 'loss': 'squared_hinge', 'penalty': 'l1'	'C': 10.0, 'loss': 'squared_hinge', 'max_iter': 2000, 'multi_class': 'ovr', 'penalty': 'l1'
Multi Layered Perceptron	'activation': 'tanh' 'alpha': 0.01, 'hidden_layer_sizes': (100,), 'learning_rate': 'constant', 'max_iter': 900, 'momentum': 0.5, 'solver': 'lbfgs'	'activation': 'tanh', 'alpha': 0.01, 'hidden_layer_sizes': (16, 32, 64), 'learning_rate': 'constant', 'max_iter': 800, 'momentum': 0.5, 'solver': 'adam'

PAPER REPORT:-

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

A) DATASET: - Cancer gene expression datasets presented in the paper **MLL (Armstrong et al.)** Famously known as MLL DATASET.

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

C) 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%)

D) FEATURE SELECTION TECHNIQUES:

I have used **THREE** feature selection techniques to find out the **best 30 features** that are selected from each feature selection.

Target No. of Features:- 30

The feature selection techniques are as follows:-

1. **Chi-Squared test with SelectKBest algorithm :-**
   I have selected 30 best features using Chi-Squared test.
   NOTE: - As the actual leukemia gene dataset contained negative values, and given that Chi-Squared test doesn't work for negative inputs, I have used **Min-Max Scaler** to scale the whole dataset in the **range (0-1)**, for better efficiency.

2. **Feature Importance method:-**
   Feature Importance is a process used to select features in the dataset that contributes the most in predicting the target variable.
   I have selected top 30 features having the most contributions using Feature Importance method.

3. **Information Gain:-** Information Gain can be used for feature selection, by evaluating the gain of each variable in the context of the target variable.
   I have selected 30 best features using Information Gain.

F) UNION & INTERSECTION OF DATA:

After feature selection, I had 30 features from three Feature Selection Processes. So altogether I had 90 numbers of features, with many features common as output from each of the processes.
So my next job was to find out the distinct features. So I moved into **union** and **intersection** of data.

1. **UNION of 90 features:-**
   After performing union among these 90 features, I got distinct 66 features.
   So, my final processed dataset looks like
   No. of rows: - **72**
   No. of columns: - 66 features + 1 label = **67**
   Shape of the dataset: (72, 67)

2. **INTERSECTION of 90 features:-**
   After performing intersection among these 90 features, I got distinct 4 features.
   So, my final processed dataset looks like
   No. of rows:- **72**
   No. of columns: - 4 features + 1 label = **5**
   Shape of the dataset: (72, 5)

G) MACHINE LEARNING CLASSIFIERS WITH OPTIMIZATION METHODS:

For both UNION dataset and INTERSECTION dataset:
Training Data : Testing Data Split = 7:3 i.e. 70-30 Split.

1. STANDARD VECTOR MACHINE (SVM):

The first machine learning classifier is **Standard Vector Machine** or **SVM**. We know it as Standard Vector Classifier (SVC). Because the dataset is small with less number of features, I have chosen **Linear Standard Vector Classifier** or **Linear SVC**. It will be applicable for both UNION and INTERSECTION dataset.
For optimization of hyperparameters in Linear SVC, I have used two optimization algorithms namely **RANDOMISED SEARCH CV** and **GRID SEARCH CV**. Using these, I have found the best hyperparameters for Linear SVC which can give the highest accuracy, both in UNION and INTERSECTION dataset.

LINEAR STANDARD VECTOR CLASSIFIER (LINEAR SVC) :

Datasets	Score Linear SVC	Accuracy Linear SVC	Linear SVC Accuracy with (RANDOMISED SEARCH CV)	Linear SVC Accuracy with (GRID SEARCH CV)
UNION	100%	95.454546%	95.454546%	95.454546%
INTERSECTION	96%	90.9090901%	90.9090901%	90.9090901%

2. MULTI LAYERED PERCEPTRON (MLP):

The second machine learning classifier is **Multi Layered Perceptron** or **MLP**. It will be applicable for both UNION and INTERSECTION dataset.

Also, for optimization of hyperparameters in MLP, I have used two optimization algorithms namely **RANDOMISED SEARCH CV** and **GRID SEARCH CV**. Using these, I have found the best hyperparameters for MLP which can give the highest accuracy, both in UNION and INTERSECTION dataset.

MULTI LAYERED PERCEPTRON (MLP) :

Datasets	Training Accuracy	Testing Accuracy	Testing Accuracy with (RANDOMISED SEARCH CV)	Testing Accuracy with (GRID SEARCH CV)
UNION	100%	95.454546%	86.363636%	95.454546%
INTERSECTION	100%	86.363636%	95.454546%	81.818182

3. ARTIFICIAL NEURAL NETWORK (ANN) :-

The third algorithm is **Artificial Neural Network** or **ANN**. It will be applicable for both UNION and INTERSECTION dataset.

For hyperparameter optimization in ANN, I have used a library from Deep Learning framework **Keras** known as **KERAS TUNER**. It gives me the best hyperparameters in the neural metwork as well as finds out the best accuracy of the neural network model trained using those hyperparameters. For weight optimization of the neurons, I have used **ADAM** optimizer.

ARTIFICIAL NEURAL NETWORK (ANN) :

UNION DATASET:

Keras Tuner optimizer (BEST HYPERPARAMETERS and BEST ACCURACY):

```
units_0: 464     act_0: tanh
units_1: 432     act_1: relu
units_2: 368     act_2: relu
units_3: 16      act_3: sigmoid
units_4: 48      act_4: relu
units_5: 48      act_5: tanh
units_6: 176     act_6: relu
units_7: 80      act_7: relu
units_8: 144     act_8: sigmoid          learning rate :- 0.001
units_9: 112     act_9: relu             No. of epochs :- 5 (10*3 trials each)
units_10: 400    act_10: relu
units_11: 336    act_11: relu
units_12: 112    act_12: tanh
units_13: 208    act_13: sigmoid
units_14: 112    act_14: relu
units_15: 432    act_15: sigmoid
units_16: 272    act_16: tanh
units_17: 208    act_17: tanh
units_18: 400    act_18: tanh
Score: 0.954545438289642
```

INTERSECTION DATASET:

Keras Tuner optimizer (BEST HYPERPARAMETERS and BEST ACCURACY):

```
num_layers: 6
units_0: 400
act_0: tanh
units_1: 400
act_1: tanh
units_2: 272
act_2: tanh
units_3: 144
act_3: sigmoid                          learning rate:- 0.01
units_4: 144                          No.of epochs:- 10 (5 * 3 trials each)
act_4: tanh
units_5: 144
act_5: relu
units_6: 272
act_6: tanh
units_7: 400
act_7: sigmoid

Score: 0.9090909361839294
```

Therefore, in Artificial Neural Network, using the best optimizers, if we train both the UNION dataset and the INTERSECTION dataset with their training data, we will get validation accuracies of **95.4545438289642 %** and **90.90909361839294 %** on their testing data respectively.

PAPER REPORT:-

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

A) DATASET: - Cancer gene expression datasets presented in the paper **DLBCL (Shipp et al.)** Famously known as DLBCL DATASET.

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

C) DIAGNOSTIC CLASSES:

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

D) FEATURE SELECTION TECHNIQUES:

I have used **THREE** feature selection techniques to find out the **best 30 features** that are selected from each feature selection.

Target No. of Features:- 30

The feature selection techniques are as follows:-

1. **Chi-Squared test with SelectKBest algorithm :-**
   I have selected 30 best features using Chi-Squared test.
   NOTE: - As the actual leukemia gene dataset contained negative values, and given that Chi-Squared test doesn't work for negative inputs, I have used **Min-Max Scaler** to scale the whole dataset in the **range (0-1)**, for better efficiency.

2. **Feature Importance method:-**
   Feature Importance is a process used to select features in the dataset that contributes the most in predicting the target variable.
   I have selected top 30 features having the most contributions using Feature Importance method.

3. **Information Gain:-** Information Gain can be used for feature selection, by evaluating the gain of each variable in the context of the target variable.
   I have selected 30 best features using Information Gain.

F) UNION & INTERSECTION OF DATA:

After feature selection, I had 30 features from three Feature Selection Processes. So altogether I had 90 numbers of features, with many features common as output from each of the processes.
So my next job was to find out the distinct features. So I moved into **union** and **intersection** of data.

1. **UNION of 90 features:-**
   After performing union among these 90 features, I got distinct 68 features.
   So, my final processed dataset looks like
   No. of rows: - **77**
   No. of columns: - 68 features + 1 label = **69**
   Shape of the dataset: (77, 69)

2. **INTERSECTION of 90 features:-**
   After performing intersection among these 90 features, I got distinct 4 features.
   So, my final processed dataset looks like
   No. of rows:- **77**
   No. of columns: - 4 features + 1 label = **5**
   Shape of the dataset: (77, 5)

G) MACHINE LEARNING CLASSIFIERS WITH OPTIMIZATION METHODS:

For both UNION dataset and INTERSECTION dataset:
Training Data : Testing Data Split = 7:3 i.e. 70-30 Split.

1. STANDARD VECTOR MACHINE (SVM):

The first machine learning classifier is **Standard Vector Machine** or **SVM**. We know it as Standard Vector Classifier (SVC). Because the dataset is small with less number of features, I have chosen **Linear Standard Vector Classifier** or **Linear SVC**. It will be applicable for both UNION and INTERSECTION dataset.
For optimization of hyperparameters in Linear SVC, I have used two optimization algorithms namely **RANDOMISED SEARCH CV** and **GRID SEARCH CV**. Using these, I have found the best hyperparameters for Linear SVC which can give the highest accuracy, both in UNION and INTERSECTION dataset.

LINEAR STANDARD VECTOR CLASSIFIER (LINEAR SVC) :

Datasets	Score Linear SVC	Accuracy Linear SVC	Linear SVC Accuracy with (RANDOMISED SEARCH CV)	Linear SVC Accuracy with (GRID SEARCH CV)
UNION	100%	91.66666%	91.66666%	91.66666%
INTERSECTION	73.5849067%	83.333334%	75%	75%

2. MULTI LAYERED PERCEPTRON (MLP):

The second machine learning classifier is **Multi Layered Perceptron** or **MLP**. It will be applicable for both UNION and INTERSECTION dataset.

Also, for optimization of hyperparameters in MLP, I have used two optimization algorithms namely **RANDOMISED SEARCH CV** and **GRID SEARCH CV**. Using these, I have found the best hyperparameters for MLP which can give the highest accuracy, both in UNION and INTERSECTION dataset.

MULTI LAYERED PERCEPTRON (MLP) :

Datasets	Training Accuracy	Testing Accuracy	Testing Accuracy with (RANDOMISED SEARCH CV)	Testing Accuracy with (GRID SEARCH CV)
UNION	96.226415%	87.5%	87.5%	87.5%
INTERSECTION	75.47169%	79.16666%	83.3333334%	75%

3. ARTIFICIAL NEURAL NETWORK (ANN) :-

The third algorithm is **Artificial Neural Network** or **ANN**. It will be applicable for both UNION and INTERSECTION dataset.

For hyperparameter optimization in ANN, I have used a library from Deep Learning framework **Keras** known as **KERAS TUNER**. It gives me the best hyperparameters in the neural metwork as well as finds out the best accuracy of the neural network model trained using those hyperparameters. For weight optimization of the neurons, I have used **ADAM** optimizer.

ARTIFICIAL NEURAL NETWORK (ANN) :

UNION DATASET:

Keras Tuner optimizer (BEST HYPERPARAMETERS and BEST ACCURACY):

```
units_0: 176              act_0: sigmoid
units_1: 336              act_1: relu
units_2: 464              act_2: relu
units_3: 400              act_3: tanh
units_4: 80               act_4: tanh
units_5: 400              act_5: sigmoid
units_6: 176              act_6: sigmoid
units_7: 144              act_7: relu                learning rate:- 0.001
units_8: 208              act_8: tanh      No. of epochs:- 5 (10*3 trials each)
units_9: 272              act_9: tanh
units_10: 368             act_10: sigmoid
units_11: 432             act_11: tanh
units_12: 48              act_12: sigmoid
units_13: 464             act_13: tanh
units_14: 208             act_14: tanh
units_15: 144             act_15: relu
units_16: 368             act_16: tanh
units_17: 336             act_17: tanh

Score: 0.8888888955116272
```

INTERSECTION DATASET:

Keras Tuner optimizer (BEST HYPERPARAMETERS and BEST ACCURACY):

```
num_layers: 9
units_0: 16
act_0: relu
units_1: 144
act_1: tanh
units_2: 16
act_2: relu
units_3: 16
act_3: relu                         learning rate:- 0.0001
units_4: 16                    No. of epochs :- 10 (5 * 3 trials each)
act_4: relu
units_5: 16
act_5: relu
units_6: 16
act_6: relu
units_7: 16
act_7: relu
units_8: 16
act_8: relu

Score: 0.8611111044883728
```

Therefore, in Artificial Neural Network, using the best optimizers, if we train both the UNION dataset and the INTERSECTION dataset with their training data, we will get validation accuracies of **88.88888955116272 %** and **86.11111044883728 %** on their testing data respectively.

CONCLUSION :-

Therefore, from the Testing Accuracies obtained from the implementation of three Machine Learning Models and Optimizing them with various Optimization algorithms, we have come to a definite conclusion that, Artificial Neural Network or ANN performs best with Keras Tuner optimization algorithm for both Union and Intersection datasets..

The accuracies given by ANN are stable and higher than that of Linear SVC and MLP algorithms.