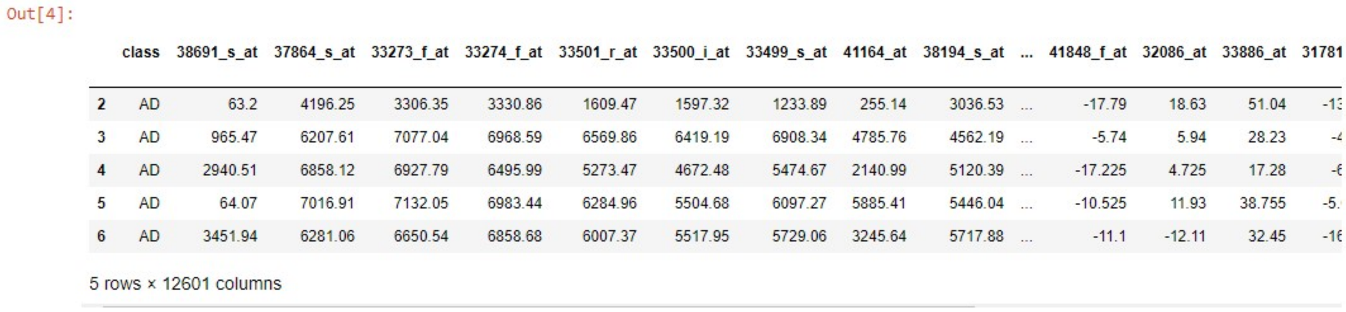
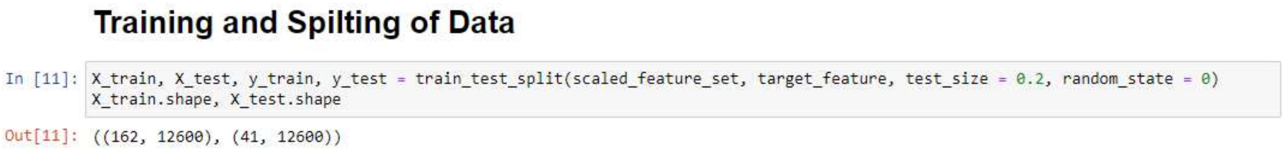
Lung Sample Dataset



Out[4]:

	class	38691_s_at	37864_s_at	33273_f_at	33274_f_at	33501_r_at	33500_i_at	33499_s_at	41164_at	38194_s_at	...	41848_f_at	32086_at	33886_at	31781
2	AD	63.2	4196.25	3306.35	3330.86	1609.47	1597.32	1233.89	255.14	3036.53	...	-17.79	18.63	51.04	-13
3	AD	965.47	6207.61	7077.04	6968.59	6569.86	6419.19	6908.34	4785.76	4562.19	...	-5.74	5.94	28.23	-4
4	AD	2940.51	6858.12	6927.79	6495.99	5273.47	4672.48	5474.67	2140.99	5120.39	...	-17.225	4.725	17.28	-6
5	AD	64.07	7016.91	7132.05	6983.44	6284.96	5504.68	6097.27	5885.41	5446.04	...	-10.525	11.93	38.755	-5.
6	AD	3451.94	6281.06	6650.54	6858.68	6007.37	5517.95	5729.06	3245.64	5717.88	...	-11.1	-12.11	32.45	-16

5 rows × 12601 columns



Training and Spilting of Data

```
In [11]: X_train, X_test, y_train, y_test = train_test_split(scaled_feature_set, target_feature, test_size = 0.2, random_state = 0)
         X_train.shape, X_test.shape

Out[11]: ((162, 12600), (41, 12600))
```

1. Total number of features present in the dataset = 12601 (including target feature) and 12600(excluding target feature)
2. Number of features extracted by each filter method = 2394(approx. 19%)
3. Average accuracy of KNN Classifier is obtained by calculating the accuracy of 19 neighbors from 1to 19
4. Filter Methods Used:

   4.1 Mutual Information

   4.2 F Classification

   4.3 T Test
5. Wrapper Methods Used:

   5.1 Sequential Forward Search

   5.2 Sequential Backward Search

Assumptions:

1. F1 = Mutual Information
2. F2 = F Classification
3. F3 = T Test
4. SFS = Sequential Forward Search
5. SBS = Sequential Backward Search
6. KNN = K Nearest Neighbors
7. SVM = Support Vector Machine
8. S1 = F1(N features) → F2(2N/3 features out of selected features from F1) → F3(N/3 features out of selected features from F2)
9. S2 = F2(N features) → F3(2N/3 features out of selected features from F2) → F1(N/3 features out of selected features from F3)
10. S3 = F3(N features) → F1(2N/3 features out of selected features from F3) → F2(N/3 features out of selected features from F1)
11. Union = F1 U F2 U F3
12. TP = True Positive
13. FP = False Positive
14. FN = False Negative
15. TN = True Negative

Here,

N = 12600

2N/3 = 8400

N/3 = 4200

U = Union of set

Filter method used by each wrapper method = F Classification(because Wrapper methods takes a lot of time to extract features in comparison to filter methods)

Classification method used by wrapper methods to extract features=Support Vector Machine

Number of features extracted by F Classification filter method to give the wrapper method for further feature extraction = 500

Number of features used by each wrapper method = 500

Number of features extracted by Wrapper Methods:

1. Sequential Forward Search = 100
2. Sequential Backward Search = 400(because it is takes a lot of time to remove 1 feature from set)

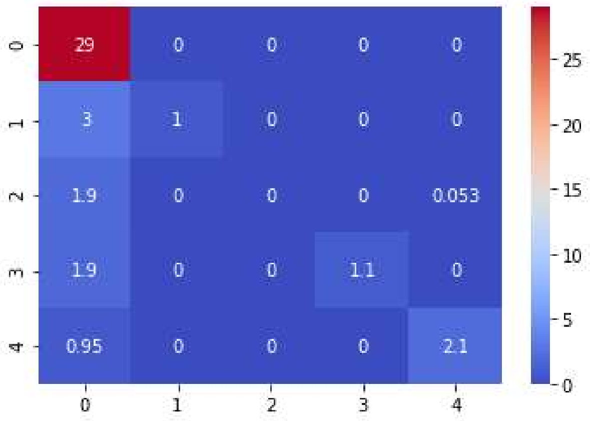
Here, confusion matrix displayed in case of KNN neighbors will correspond to the maximum accuracy achieved by the KNN neighbors.

Results:

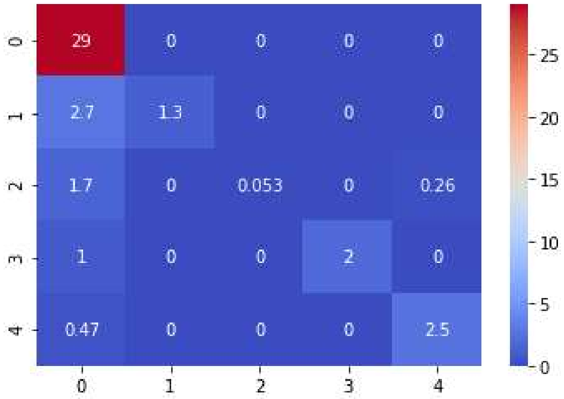
Comparison Table of KNN Classifier

Parameters	Time Taken For Execution (seconds)	Average		Best/Maximum	
Method		Accuracy		Accuracy	F-Score
F1	2109.1994	84.9807	0.8498	95.122	0.9512
F2	2158.3983	80.8729	0.8087	85.3659	0.8536
F3	147.6579	86.6496	0.8664	97.5610	0.9756
Union	4415.2556	83.0552	0.8305	85.3659	0.8536
S1	2027.5659	91.2709	0.9127	97.5610	0.9756
S2	3277.7884	92.4262	0.9242	97.5610	0.9756
S3	1019.6803	85.4942	0.8549	95.1220	0.9512
SFS	501	81.0013	0.81	82.9268	0.8292
SBS	4217	79.0757	0.7907	85.3659	0.8536

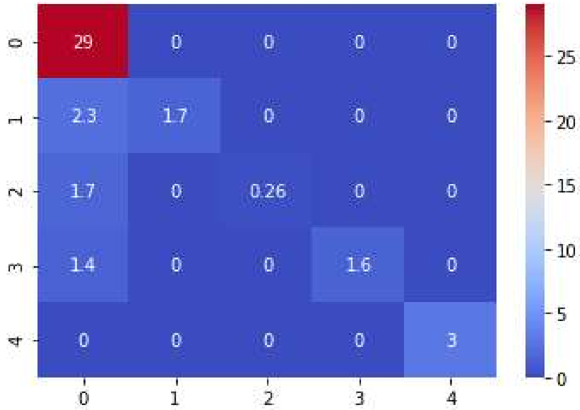
1. F1 Confusion Matrix:



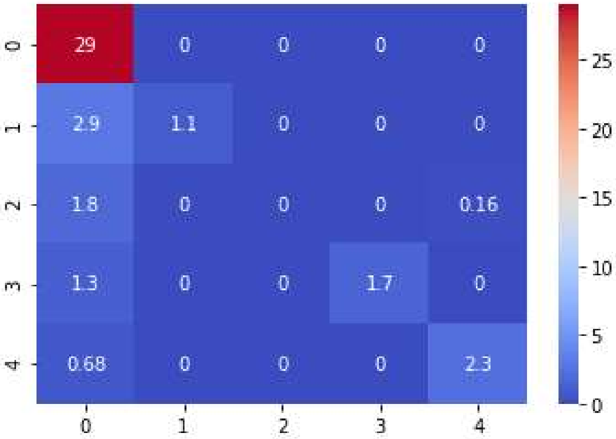
2. F2 Confusion Matrix:



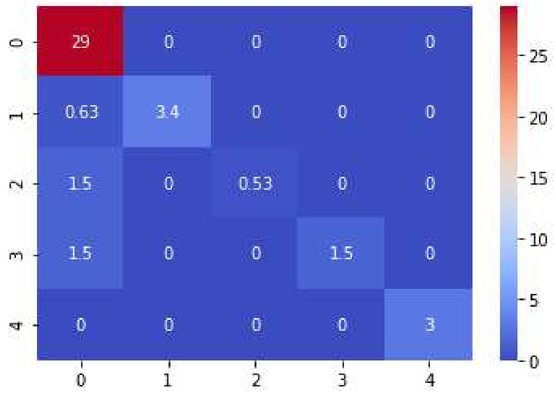
3. F3 Confusion Matrix:



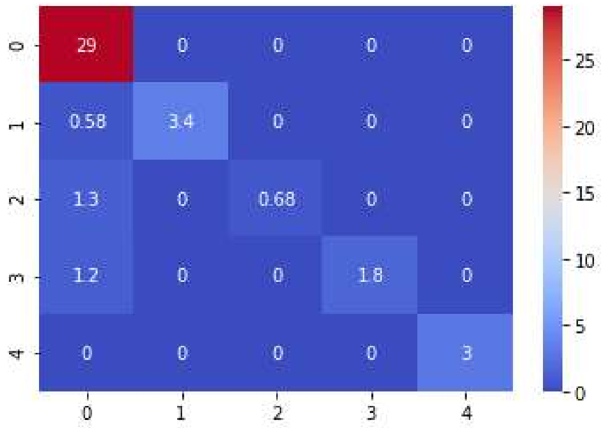
4. Union Confusion Matrix:



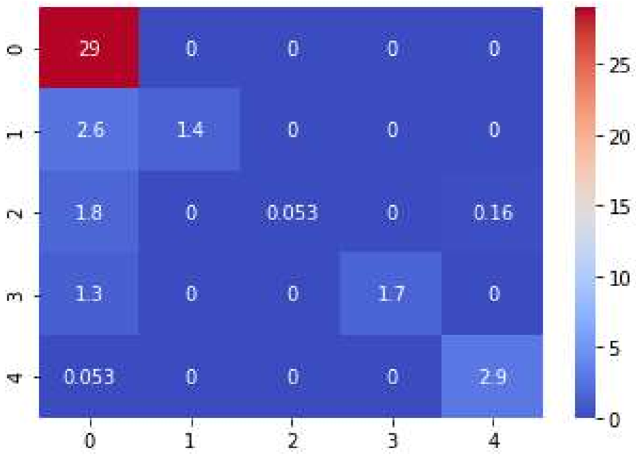
5. S1 Confusion Matrix:



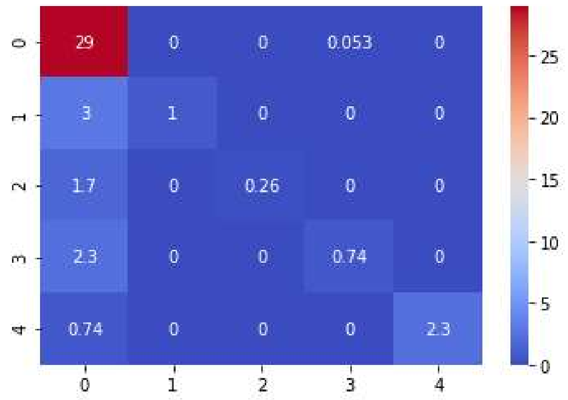
6. S2 Confusion Matrix:



7. S3 Confusion Matrix:

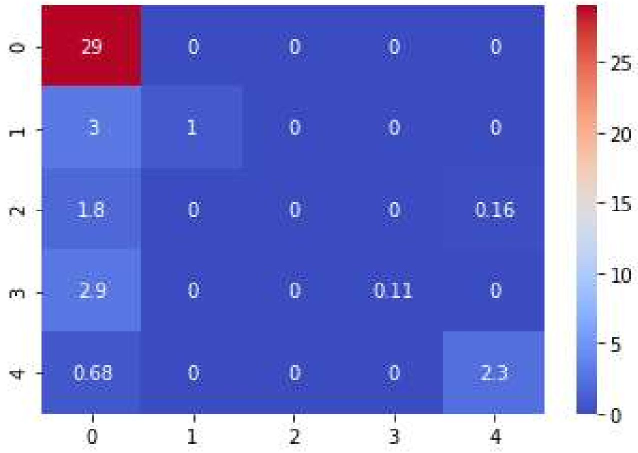


8. SFS Confusion Matrix:



	0	1	2	3	4
0	29	0	0	0.053	0
1	3	1	0	0	0
2	1.7	0	0.26	0	0
3	2.3	0	0	0.74	0
4	0.74	0	0	0	2.3

9. SBS Confusion Matrix:

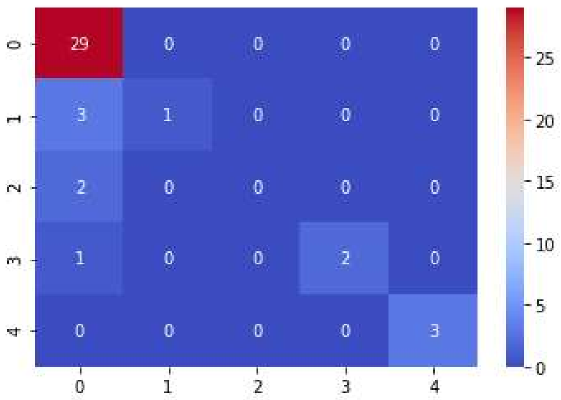


	0	1	2	3	4
0	29	0	0	0	0
1	3	1	0	0	0
2	1.8	0	0	0	0.16
3	2.9	0	0	0.11	0
4	0.68	0	0	0	2.3

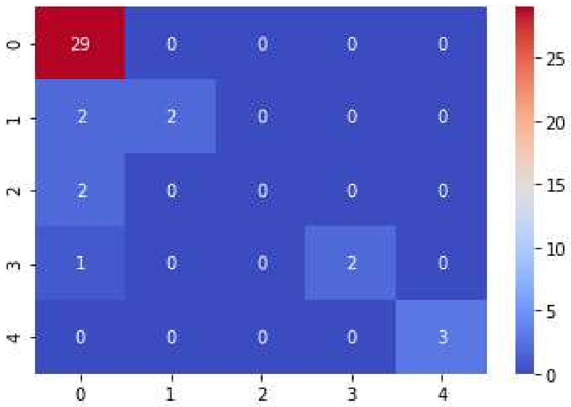
Comparison Table of SVM Classifier

Parameters → ↓ Method	Time Taken For Execution (seconds)	Accuracy	F-Score
F1	2109.1994	87.8049	0.878
F2	2158.3983	85.3659	0.8536
F3	147.6579	87.8049	0.878
Union	4415.2556	85.3659	0.8536
S1	2027.5659	90.2439	0.9024
S2	3277.7884	90.2439	0.9024
S3	1019.6803	87.8049	0.878
SFS	501	82.9268	0.8292
SBS	4217	80.4878	0.8048

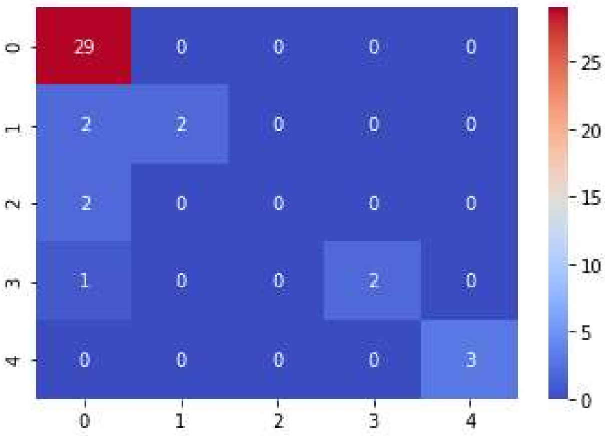
1. F1 Confusion Matrix:



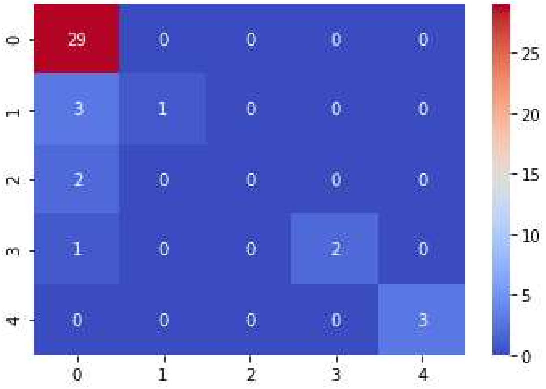
2. F2 Confusion Matrix:



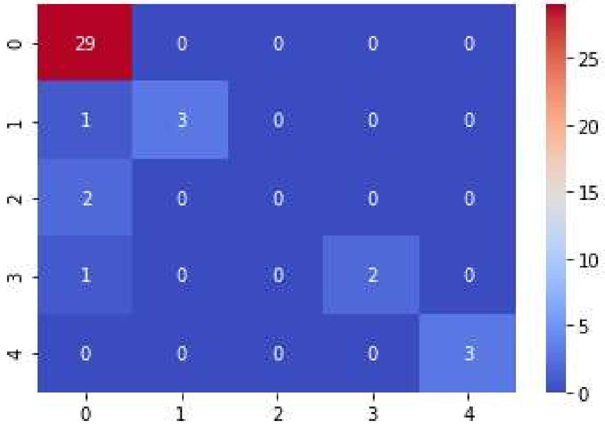
3. F3 Confusion Matrix:



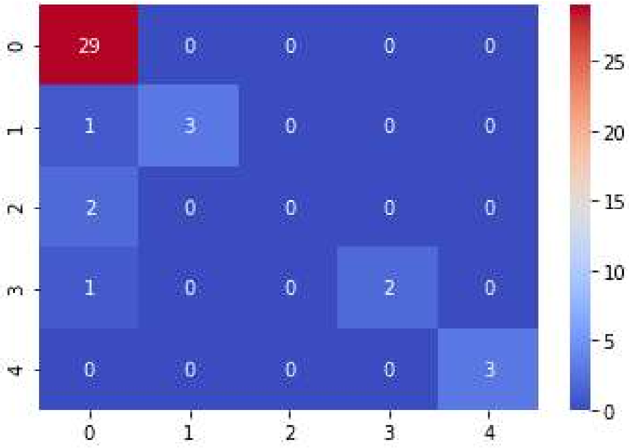
4. Union Confusion Matrix:



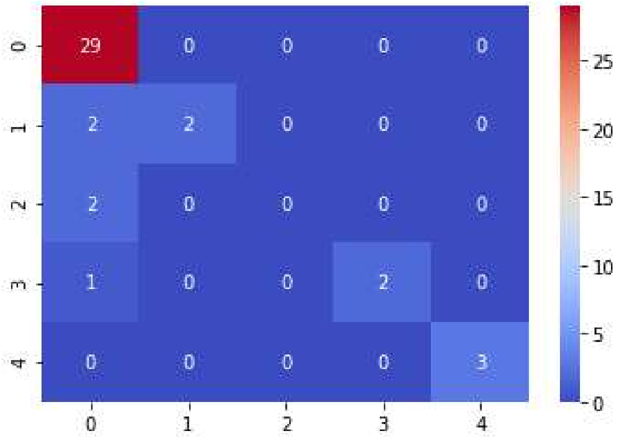
5. S1 Confusion Matrix:



6. S2 Confusion Matrix:

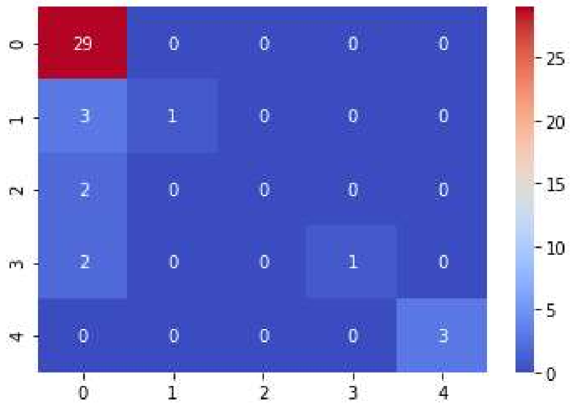


7. S3 Confusion Matrix:



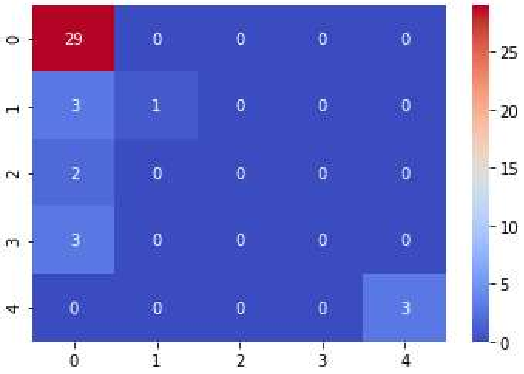
	0	1	2	3	4
0	29	0	0	0	0
1	2	2	0	0	0
2	2	0	0	0	0
3	1	0	0	2	0
4	0	0	0	0	3

8. SFS Confusion Matrix:



	0	1	2	3	4
0	29	0	0	0	0
1	3	1	0	0	0
2	2	0	0	0	0
3	2	0	0	1	0
4	0	0	0	0	3

9. SBS Confusion Matrix:



	0	1	2	3	4
0	29	0	0	0	0
1	3	1	0	0	0
2	2	0	0	0	0
3	3	0	0	0	0
4	0	0	0	0	3

Observations and Conclusions:

1. F3(T-Test) is providing the best accuracy, confusion matrix and f-score among all filter methods for KNN Classifier but it is also taking the least time among all the other filter method.
2. S2(F2->F3->F1) cascading method is providing the best accuracy, confusion matrix and f-score among all the features extraction methods for KNN Classification.
3. T-test is better than Mutual Information as it has higher average accuracy in KNN classification than mutual information and it also takes less time for execution than mutual information. Both are giving same accuracy in SVM Classification but T-Test is taking less time than Mutual Information for feature extraction.
4. T-test is taking least time among all feature extraction methods.
5. S1(F1->F2->F3) and S2(F2->F3->F1) is providing the better solution than F3(T-test) in terms of average accuracy and but time taken for the execution of S1 and S2 is more in case of KNN classification.
6. Combination wrapper and filter methods is not providing better solution in comparison of using single filter method or cascading S1, S2, S3 for feature extraction.
7. Highest average accuracy achieved by KNN is in case of S2(F2->F3->F1) which is 92.4262%.
8. Highest maximum accuracy achieved by KNN is in case of F3(T-test), S1(F1->F2->F 3) and S2(F2->F3->F1) which is 97.5610%.
9. Highest accuracy achieved by SVM is in case of S1(F1->F2->F3) and S2(F2->F3->F1) which is 93.3333%.
10. KNN is providing better average and maximum accuracy than SVM classification for this dataset.
11. Using T-test for feature extraction is also a good solution as it is providing 86.6496.% average accuracy in KNN and 87.8049% accuracy in SVM classification with least ti me.

12. Cascading of filter methods S1, S2, S3 is a very solution for this dataset in terms of accuracy, f score.
13. Maximum accuracy and maximum f-score achieved in case of SVM (Support Vector Machine) is achieved in case of S1(F1->F2->F3) and S2(F2->F3->F1).But S2(F2->F3->F1) is better than S1(F2->F3->F1) here because S2(F2->F3->F1) is taking less time than S1(F1->F2->F3)