MACHINE LEARNING-BASED CLASSIFICATION OF EPILEPTIC PATIENTS INTO MILD AND MILD SEVERE CATEGORIES USING EEG SIGNALS

SIGNAL AND IMAGEGING ACQUISITION AND MODELING IN HEATHCARE

GROUP: 11

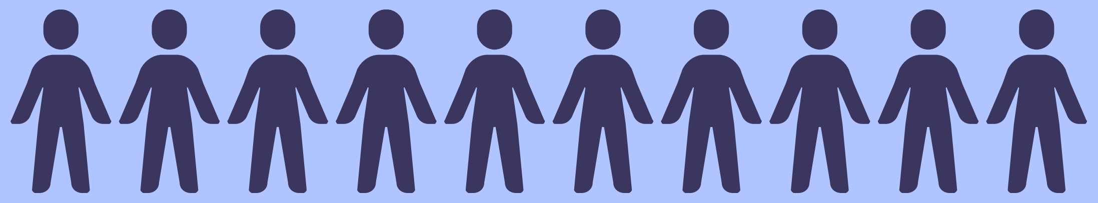
Yesim Nur Tortop
Mustafa Soydan
Kerem Erciyes

Supervised By: Isabella Castiglioni

EPILEPSY AND DIAGNOSIS CHALLANGES

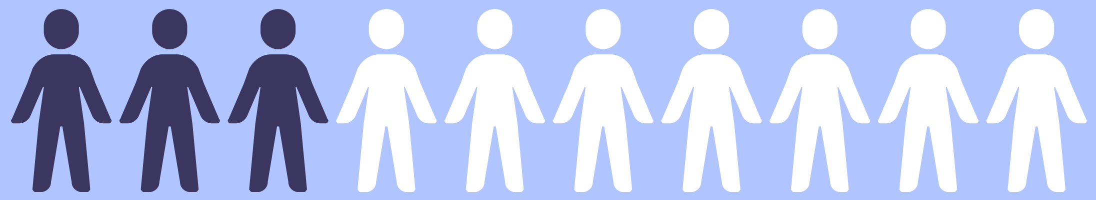
50M

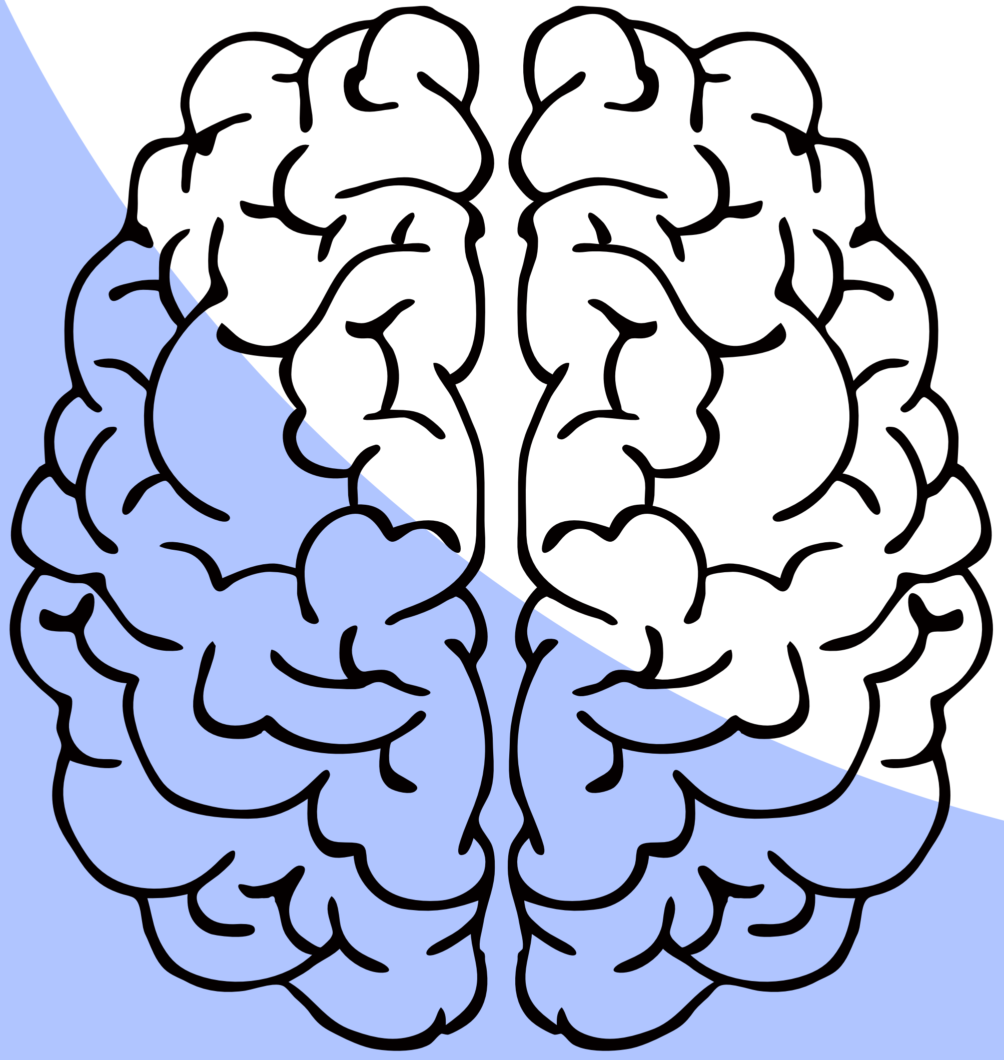
People affected by epilepsy [1]



30%

Patients experiencing diagnostic challanges [2]



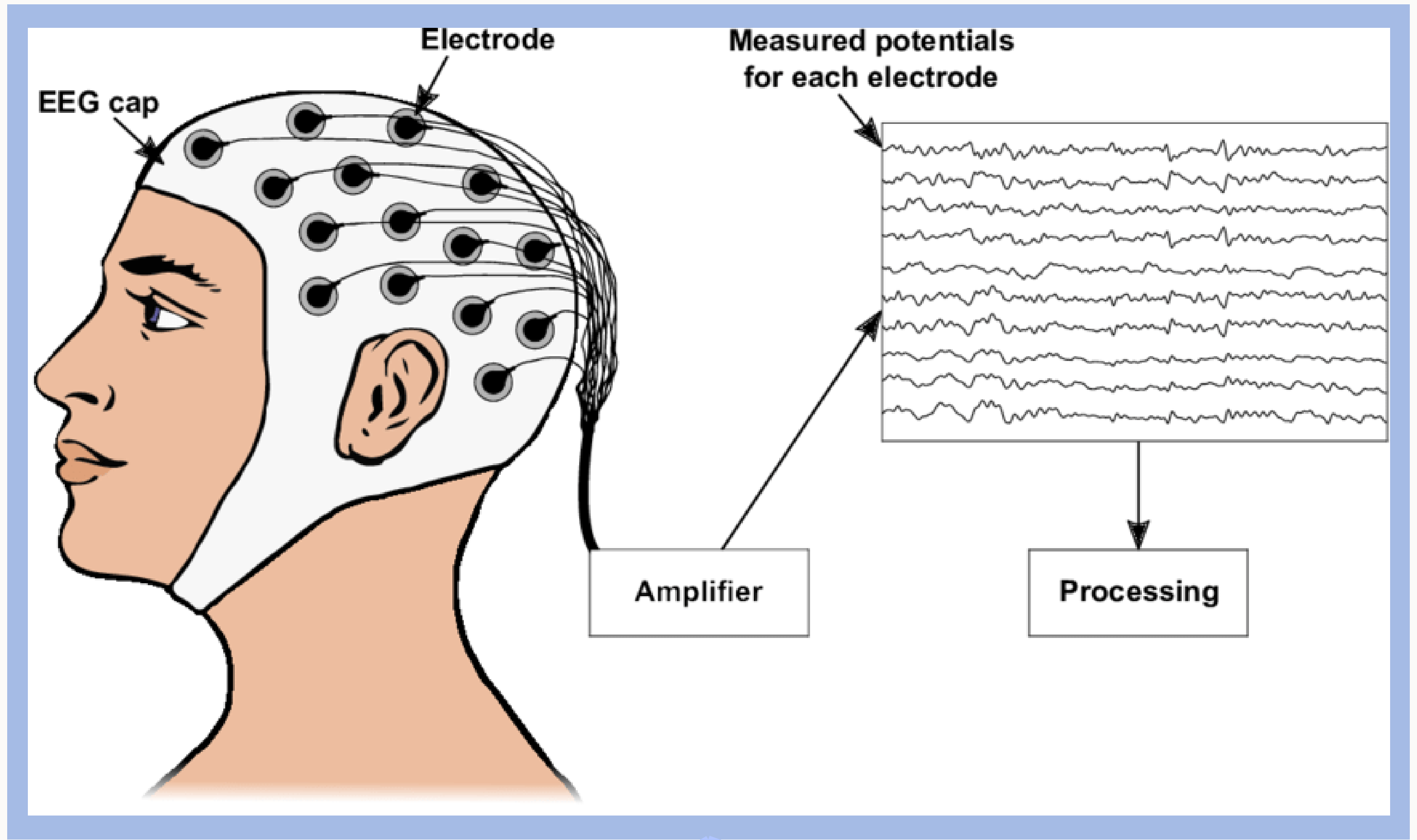


- [1] WHO. (2020). Epilepsy: a public health imperative. [Online]. Available: _
- [2] Krumholz et al. (2015). Management of an unprovoked first seizure in adults. Neurology, 85(17), 1526-1537.

AIM OF THE PROJECT

The principal objective of this project is to establish a machine learning methodology for the classification of epilepsy severity based on EEG data analysis.

EEG DATASET



An electroencephalogram (EEG) is a test that measures electrical activity in the brain using small, metal discs (electrodes) attached to the scalp

EEG data provides insights as;

- Aiding in diagnosis
- Classification
- Localization
- Treatment monitoring

DATA FROM 500 INDIVIDUALS

4094 FEATURE

MODEL PREPARATION PROCEDURE

1. Data Preparation
2. Defining Feature Selection Techniques
3. Classification with SVM
4. Classification with Random Forest
5. Model Comparision
6. Conclusion

DATA PREPARATION

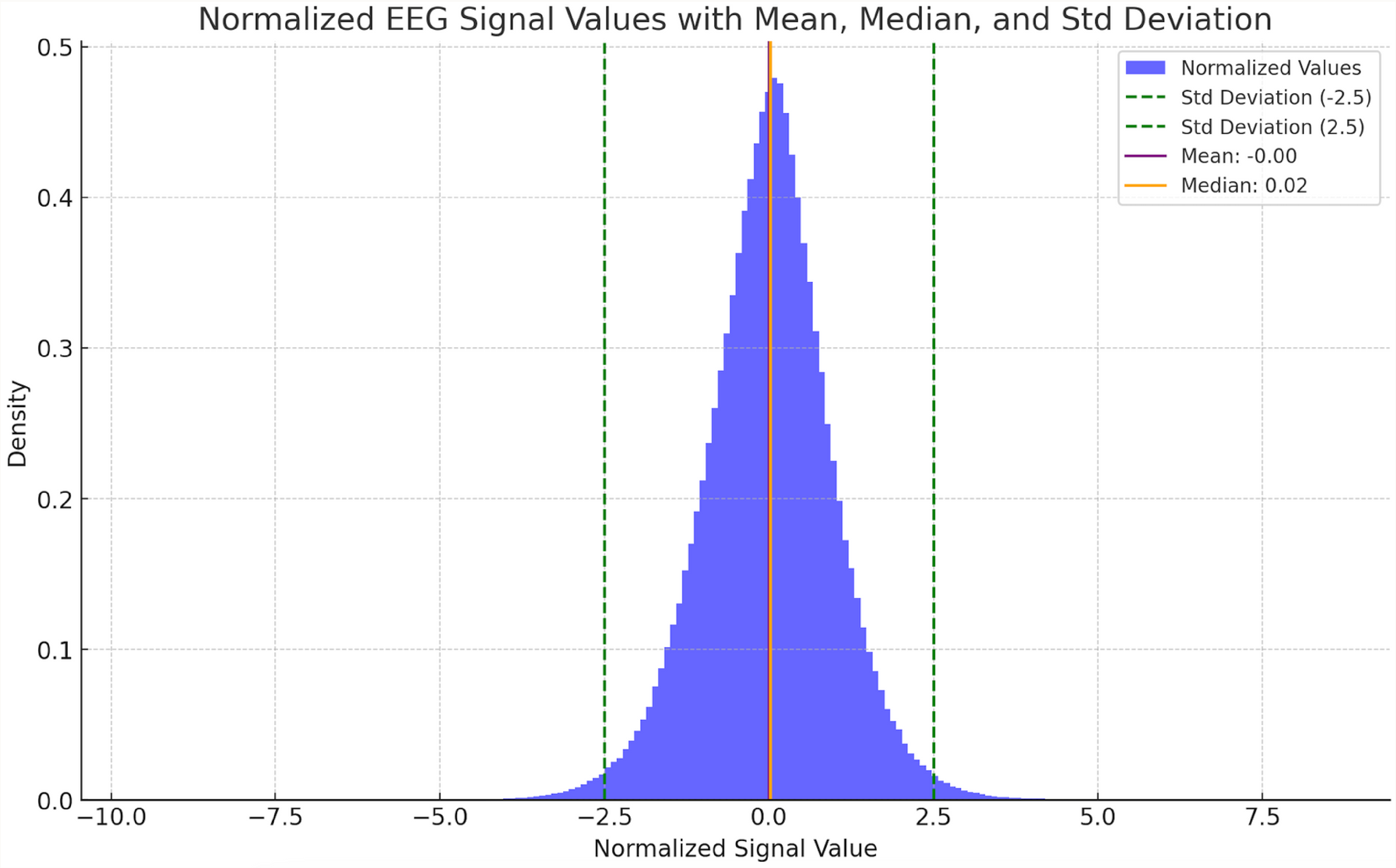
01 SELECTING CLASS 3 AND CLASS 4 PATIENTS

02 NORMALIZATION

03 TRAIN AND TEST SPLIT

04 OUTLIER DETECTION

NORMALIZATION



$$X_{normalized} = \frac{X - X_{mean}}{X_{stddev}}$$

OUTLIER DETECTION

Z SCORE THRESHOLD = 2.5

IF THERE ARE MORE THAN 250 COLUMN IS ABOVE THE THRESHOLD FOR ONE PATIENT

REMOVE THAT ROW

4 ROW DROPPED FROM CLASS 3

5 ROW DROPPED FROM CLASS 4

FEATURE SELECTION

PCA

1. DIMENSIONALITY REDUCTION
2. FEATURE COMPRESSION
3. IMPROVED MODEL PERFORMANCE
4. REMOVAL OF REDUNDANCY
5. SPEEDING UP LEARNING ALGORITHMS

SELECT K BEST

1. DIMENSIONALITY REDUCTION
2. REDUCED OVERFITTING
3. IMPROVED MODEL PERFORMANCE
4. INTERPRETABILITY
5. SPEEDING UP LEARNING ALGORITHMS

SELECTED FEATURES FOR KBEST

K = 1

Indices of selected features: [3934]

K = 5

IIIndices of selected features: [1350 2271 2309 3934 4077]

K = 50

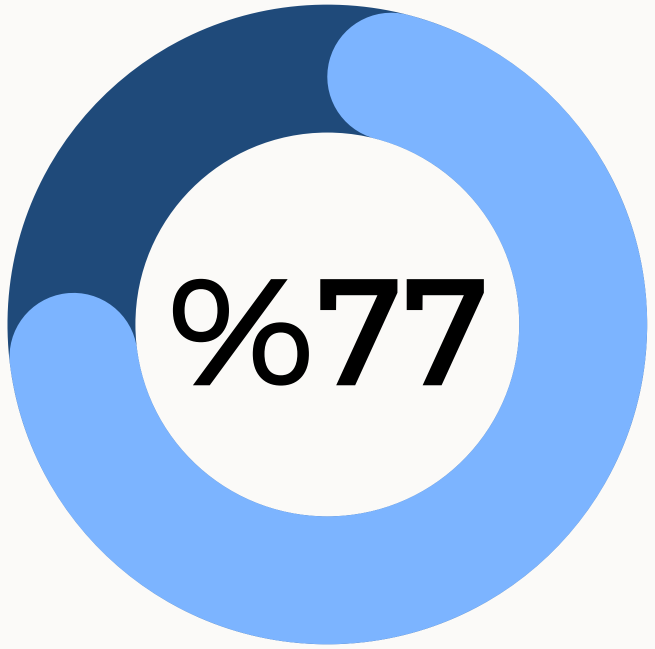
Indices of selected features: [382 383 384 513 514 543 650 1282 1283 1347 1349 1350 1468 1469 1470 2148 2181 2182 2183 2184 2232 2233 2271 2272 2273 2291 2308 2309 2310 2330 2331 3101 3105 3118 3120 3121 3122 3166 3167 3168 3320 3321 3934 3935 3936 3976 4075 4076 4077 4078]

K = 100

Indices of selected features: [14 143 363 382 383 384 429 448 513 514 517 543 583 612 650 658 858 1222 1228 1276 1281 1282 1283 1347 1348 1349 1350 1351 1393 1468 1469 1470 1471 1510 1543 1562 1564 2139 2141 2148 2181 2182 2183 2184 2207 2232 2233 2270 2271 2272 2273 2286 2291 2307 2308 2309 2310 2330 2331 2332 2539 2541 2644 2906 3068 3098 3101 3105 3118 3120 3121 3122 3123 3166 3167 3168 3169 3289 3295 3320 3321 3363 3366 3917 3933 3934 3935 3936 3975 3976 3977 3978 3979 4006 4030 4031 4075 4076 4077 4078]

THE MOST IMPORTANT FEATURE ACCORDING TO KBEST

COLUMN 3934



ACCURACY

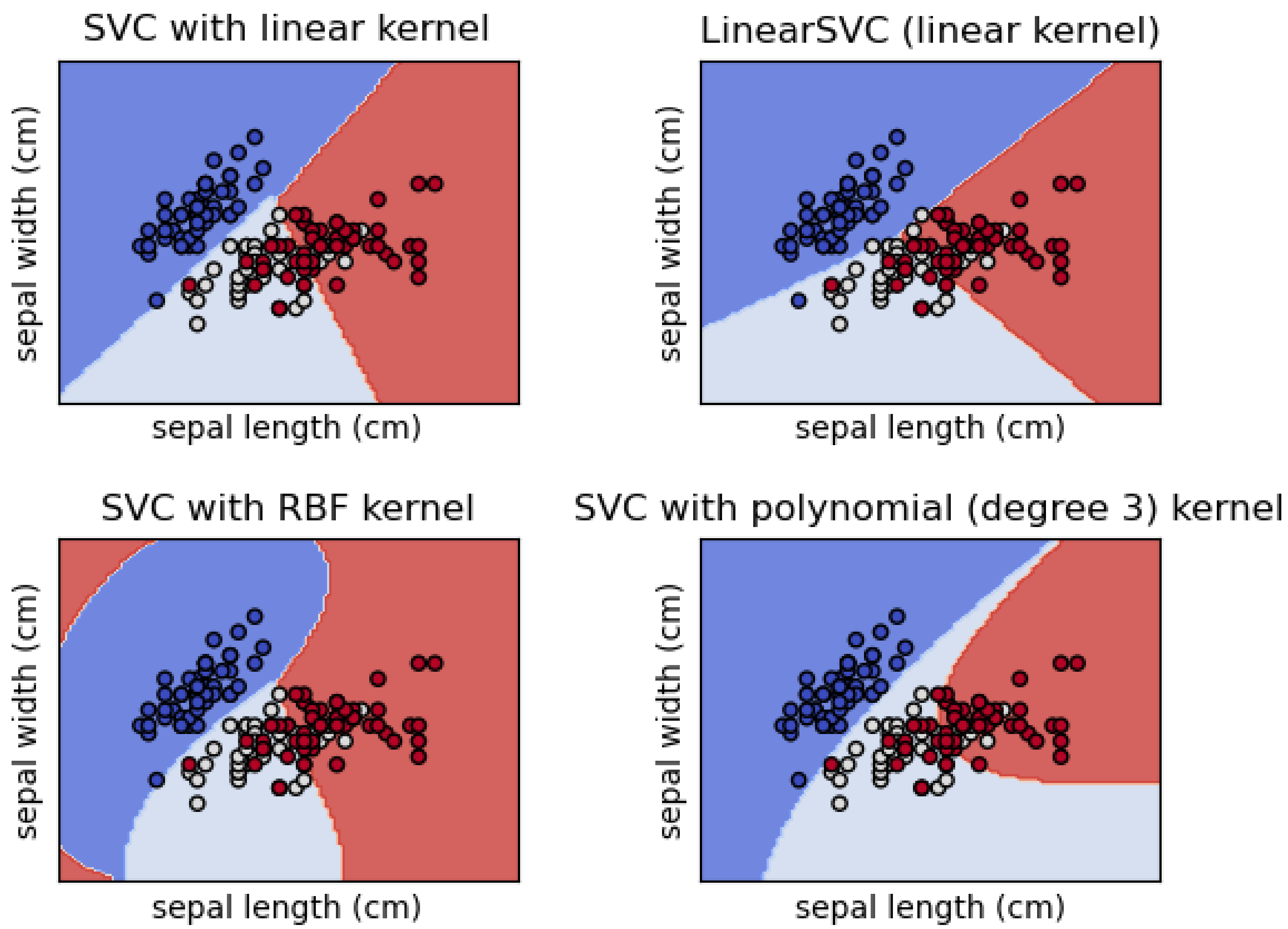
EUI
X3934
-34
28
131
11
17
131
29
-45
23
22
13
21

SVM

SVM is a supervised learning algorithm. It constructs a hyperplane in a high or infinite dimensional space.

HYPERPARAMETERS

- C: Controls the trade-off between smooth decision boundaries and classifying training points correctly.
- Kernel: The function used to map the dataset into a higher dimensional space where it is easier to classify the data linearly.
- Gamma (γ): Determines the distance of influence of a single training example, with low values indicating far and high values indicating close.



5-FOLD CROSS-VALIDATION

WITH OUTLIERS

FOLD 1 ACCURACY
0.59375

FOLD 2 ACCURACY
0.625

FOLD 3 ACCURACY
0.71875

FOLD 4 ACCURACY
0.84375

FOLD 5 ACCURACY
0.6875

MEAN ACCURACY
0.69375

STANDARD DEVIATION
0.087

WITHOUT OUTLIERS

FOLD 1 ACCURACY
0.6875

FOLD 2 ACCURACY
0.6875

FOLD 3 ACCURACY
0.65625

FOLD 4 ACCURACY
0.78125

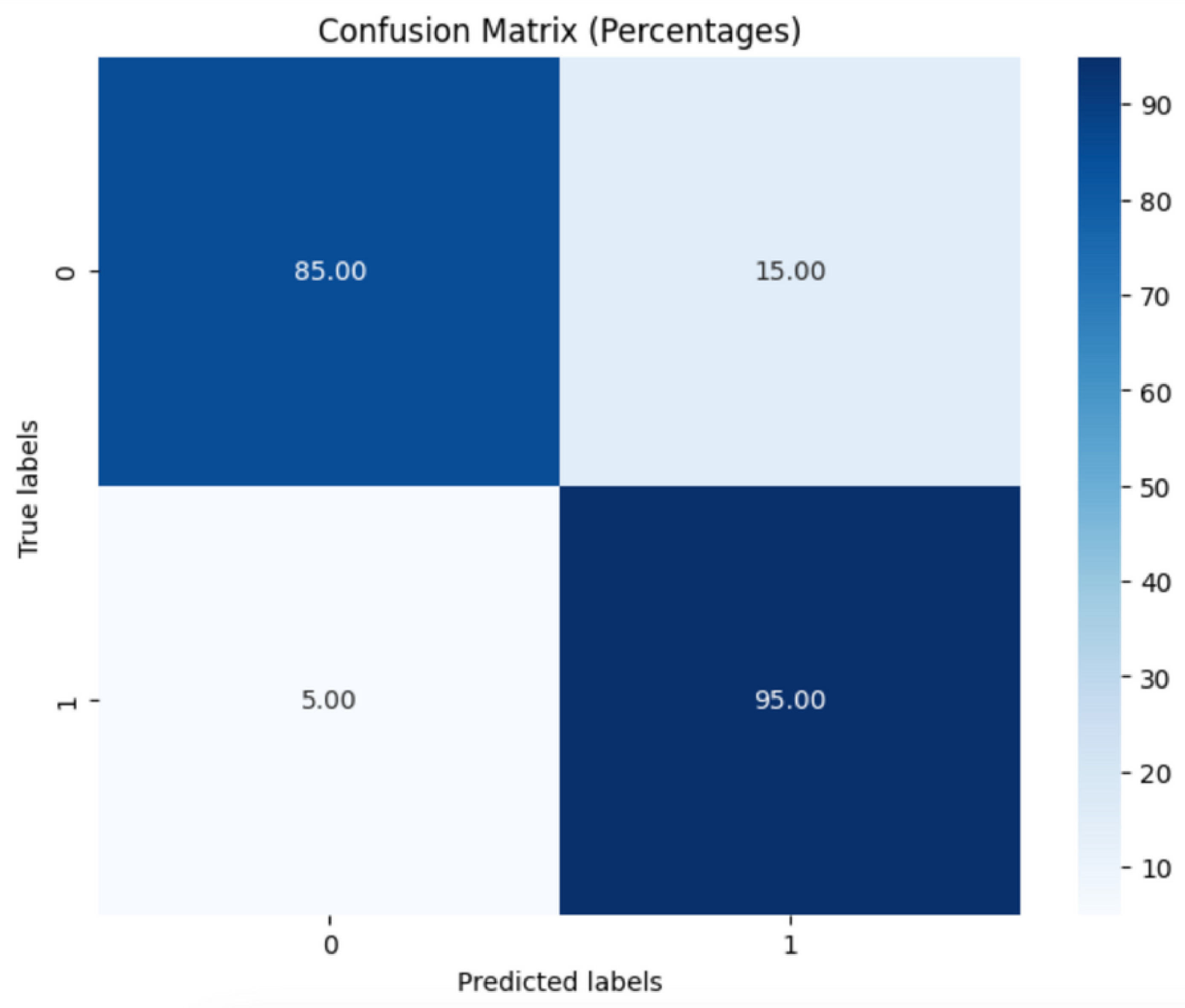
FOLD 5 ACCURACY
0.5625

MEAN ACCURACY
0.675

STANDARD DEVIATION
0.0701

SVM TEST

CONFUSION MATRIX



SELECTED HYPERPARAMETERS

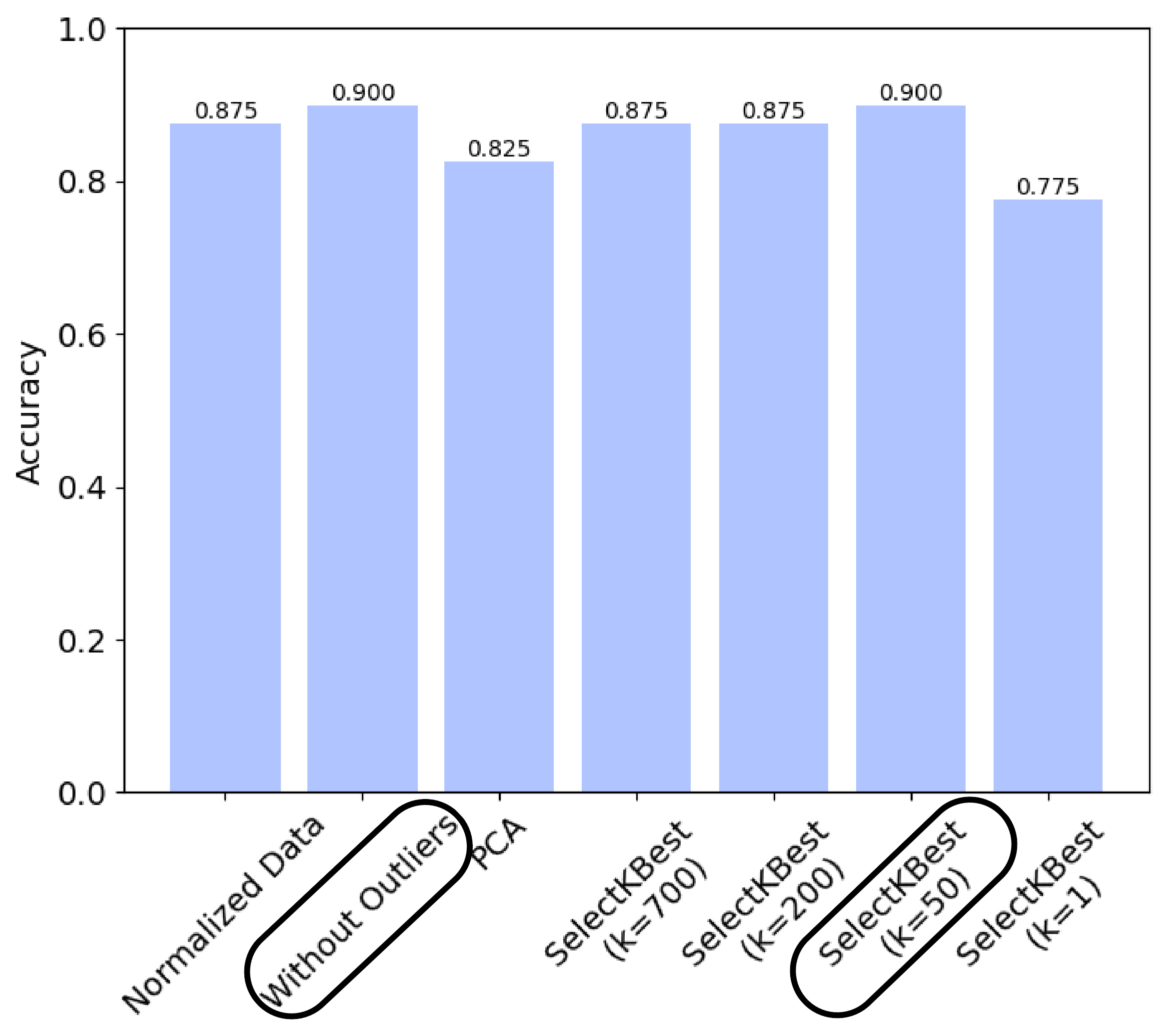
Kernel = RBF

C= 10

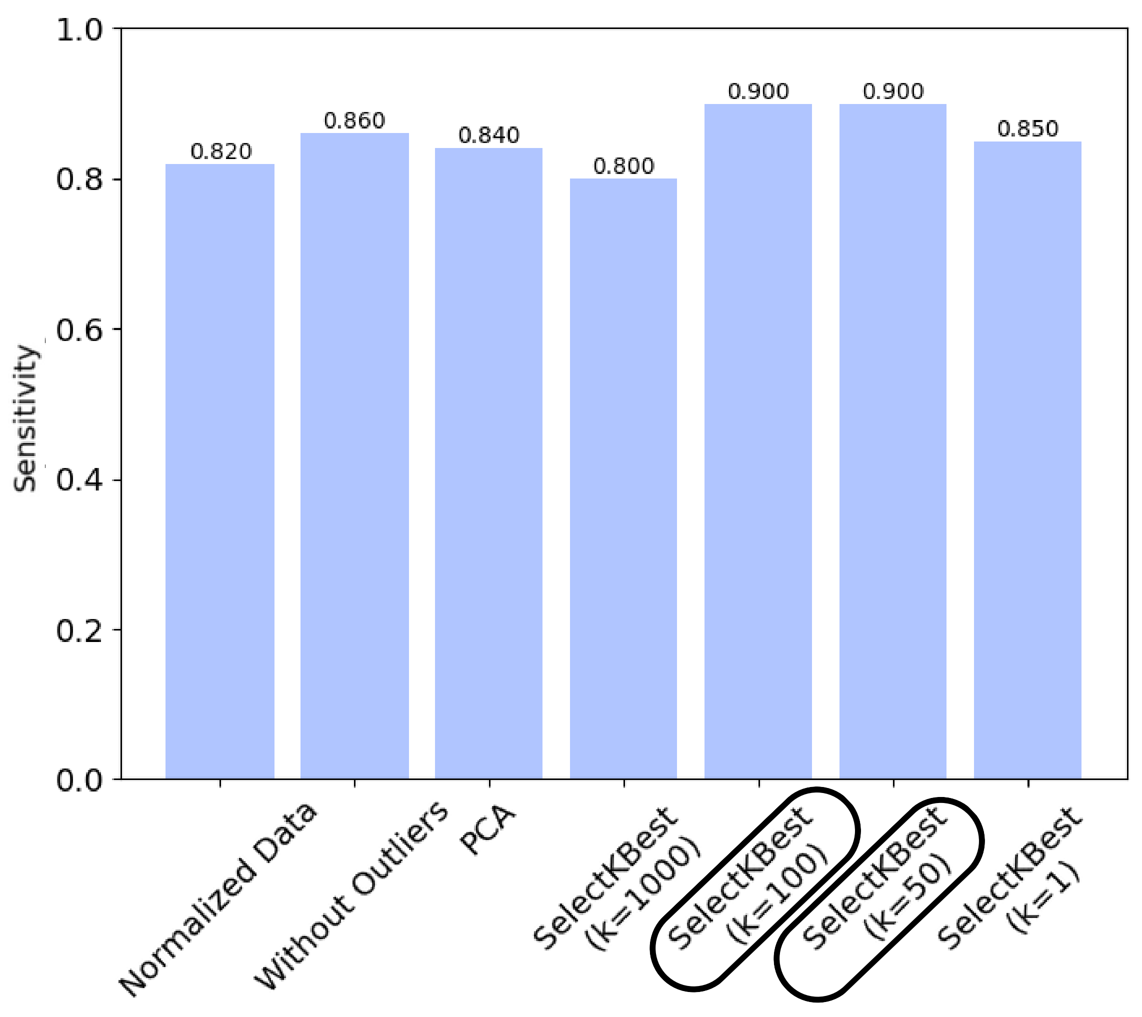
Gamma = 0.0005

RESULT COMPARISON FOR SVM

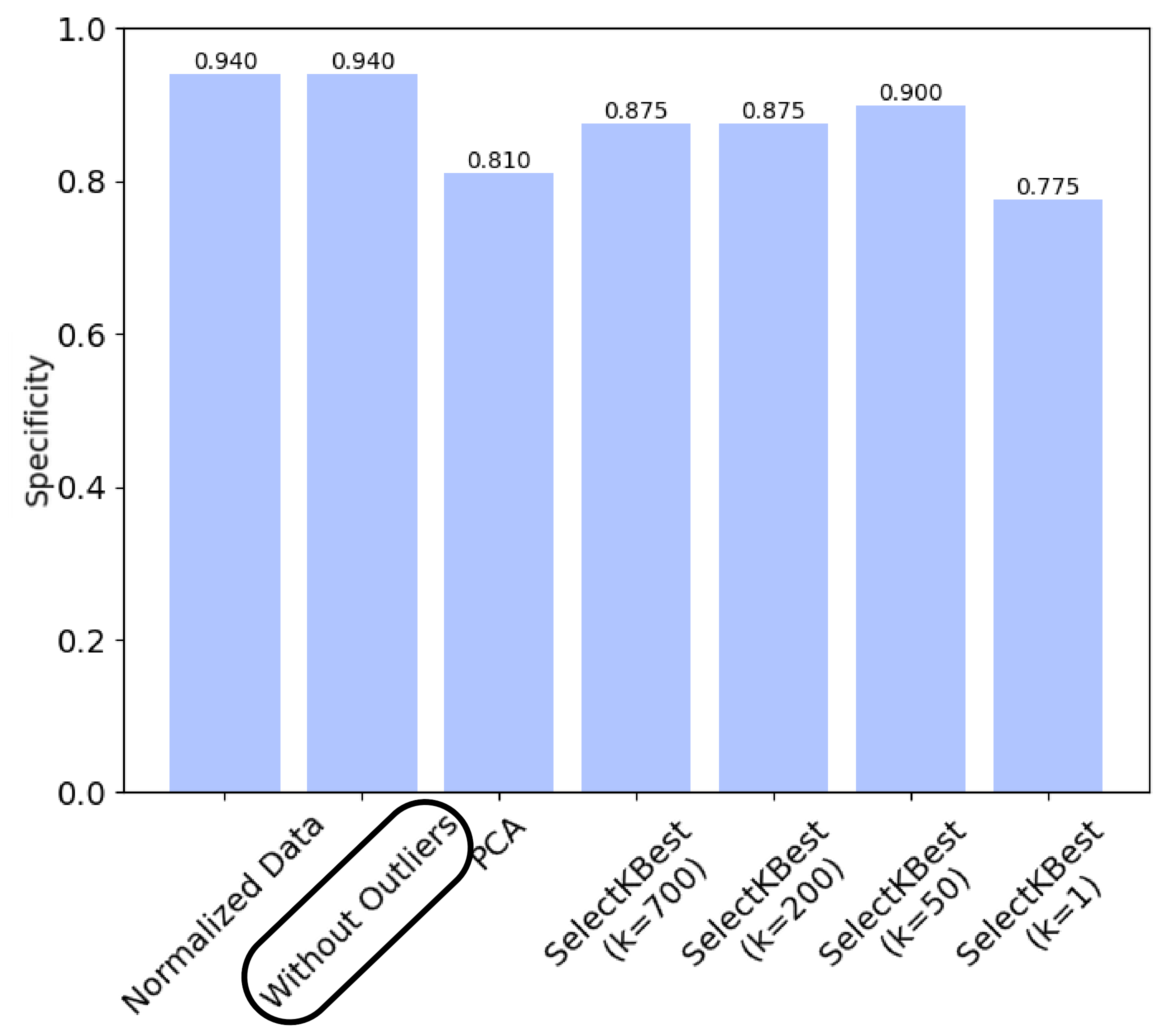
ACCURACY



SENSITIVITY (TRUE POSITIVE RATE)

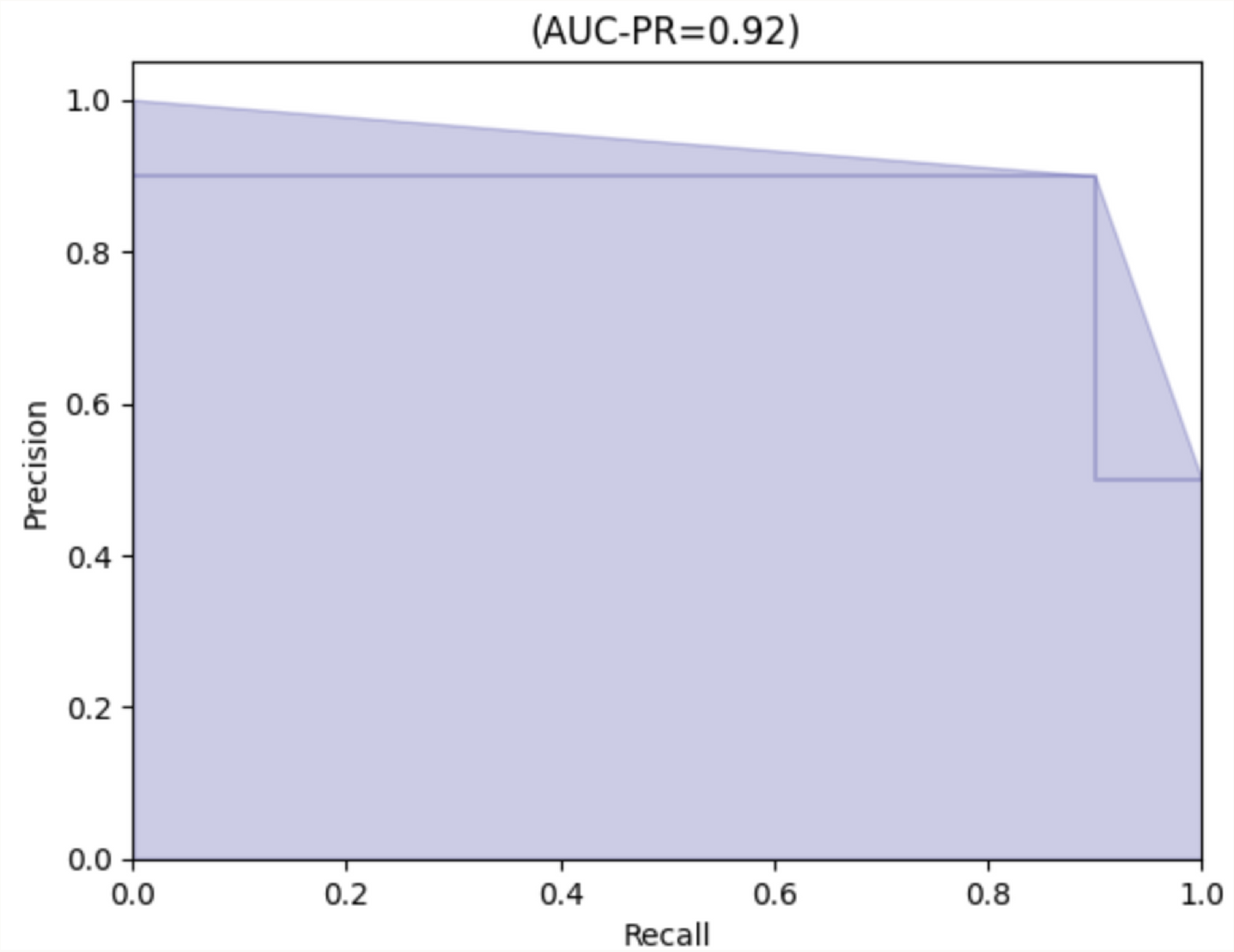


SPECIFICITY (TRUE NEGATIVE RATE)

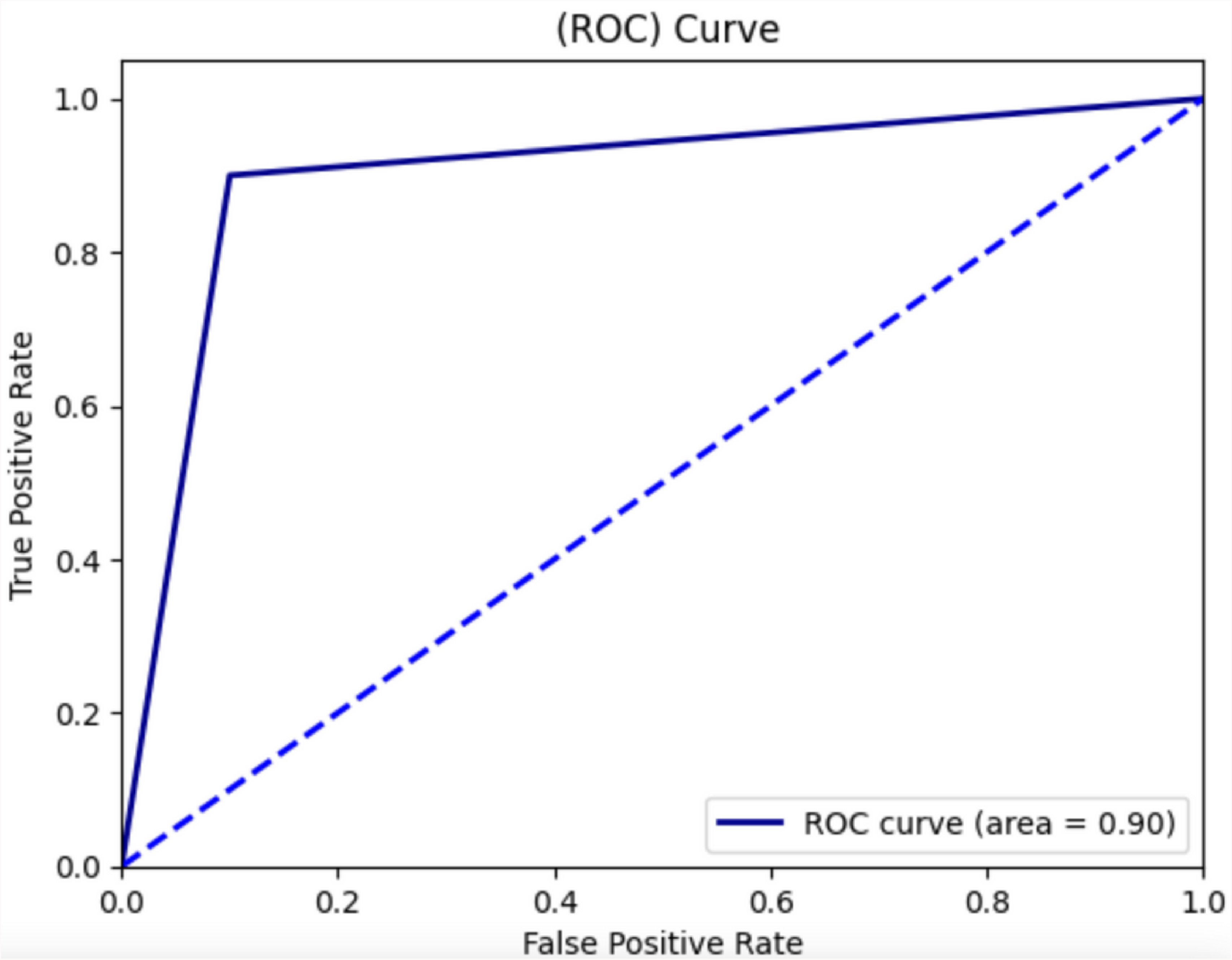


METRIC RESULTS FOR SVM

PRECISION RECALL CURVE



ROC CURVE



RANDOM FOREST

Random Forest is an ensemble learning technique. It builds multiple decision trees and merges them together to get a more accurate and stable prediction.

HYPERPARAMETERS

- N-Estimators: This parameter specifies the number of trees in the forest.
- Max Depth: It's the maximum length of the paths from the root to any leaf.
- Criterion: These measures affect how the decision trees decide to split data at a node.
- Max_features: The number of features to consider when looking for the best split.

5-FOLD CROSS-VALIDATION

WITH OUTLIERS

FOLD 1 ACCURACY
0.75

FOLD 2 ACCURACY
0.78125

FOLD 3 ACCURACY
0.78125

FOLD 4 ACCURACY
0.6875

FOLD 5 ACCURACY
0.6875

MEAN ACCURACY
0.7375

STANDARD DEVIATION
0.0424

WITHOUT OUTLIERS

FOLD 1 ACCURACY
0.90322

FOLD 2 ACCURACY
0.866

FOLD 3 ACCURACY
0.8

FOLD 4 ACCURACY
0.866

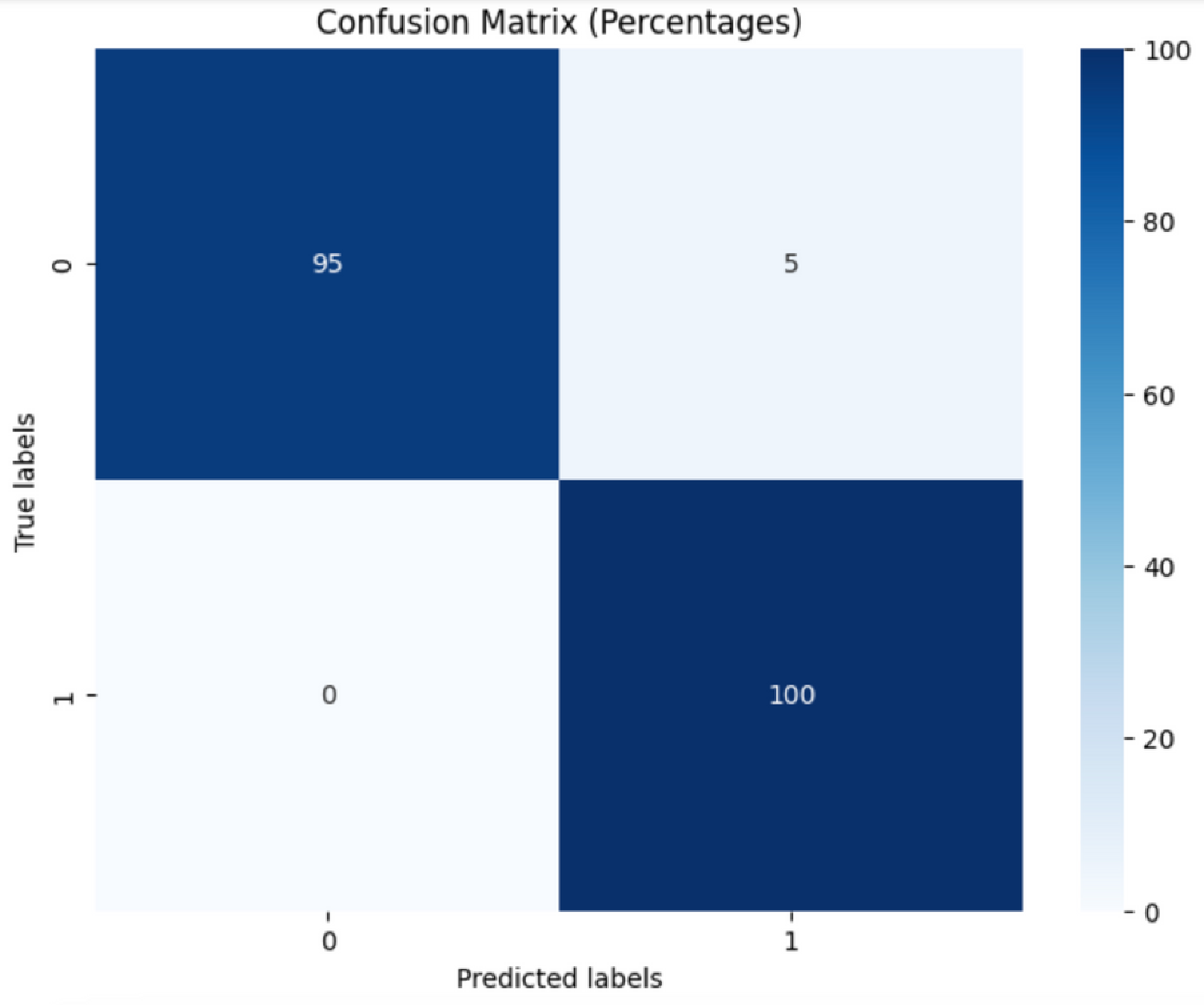
FOLD 5 ACCURACY
0.833

MEAN ACCURACY
0.854

STANDARD DEVIATION
0.0348

RANDOM FOREST TEST

CONFUSION MATRIX



SELECTED HYPERPARAMETERS

criterion = entropy

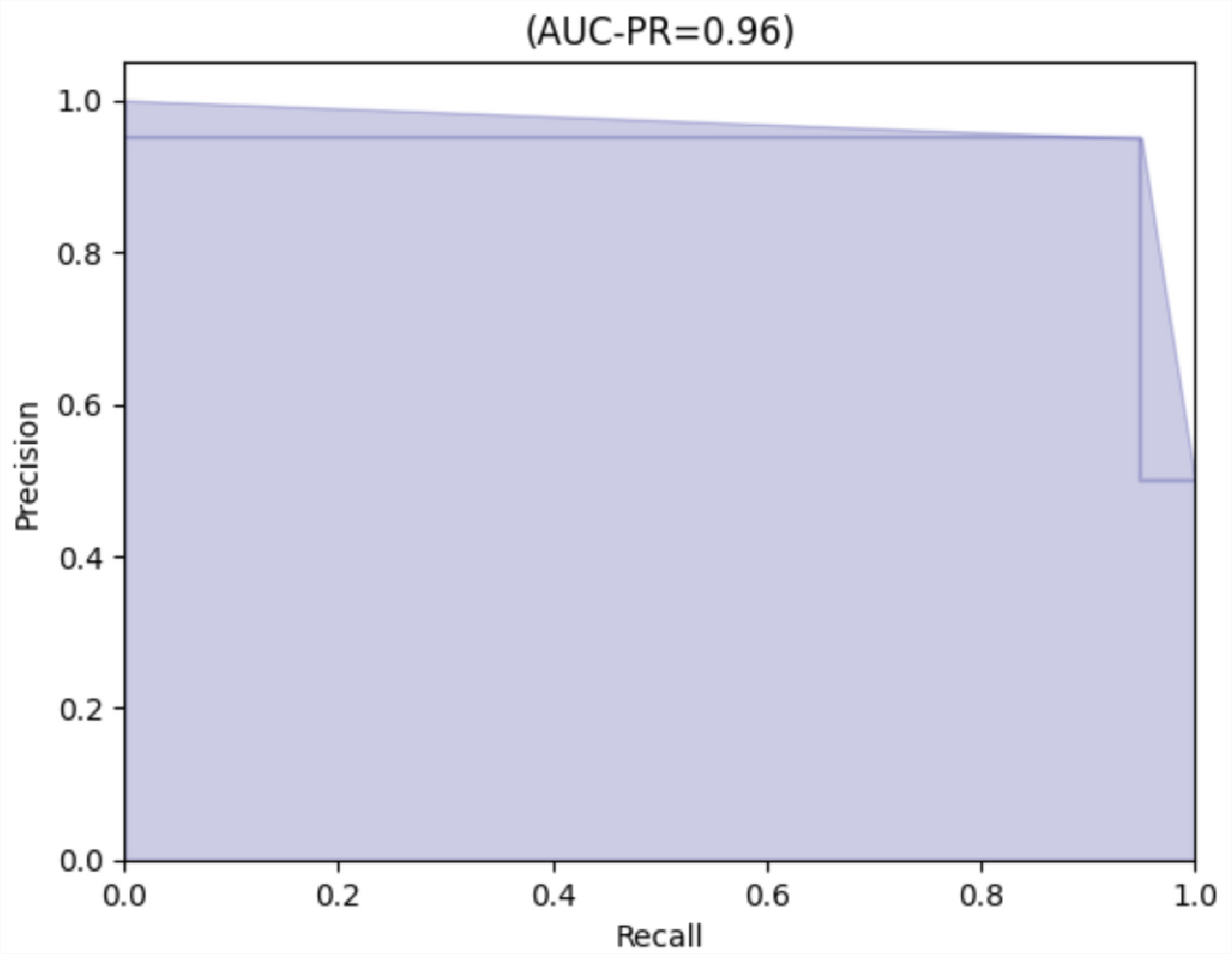
max_depth = 10

max_features = log2

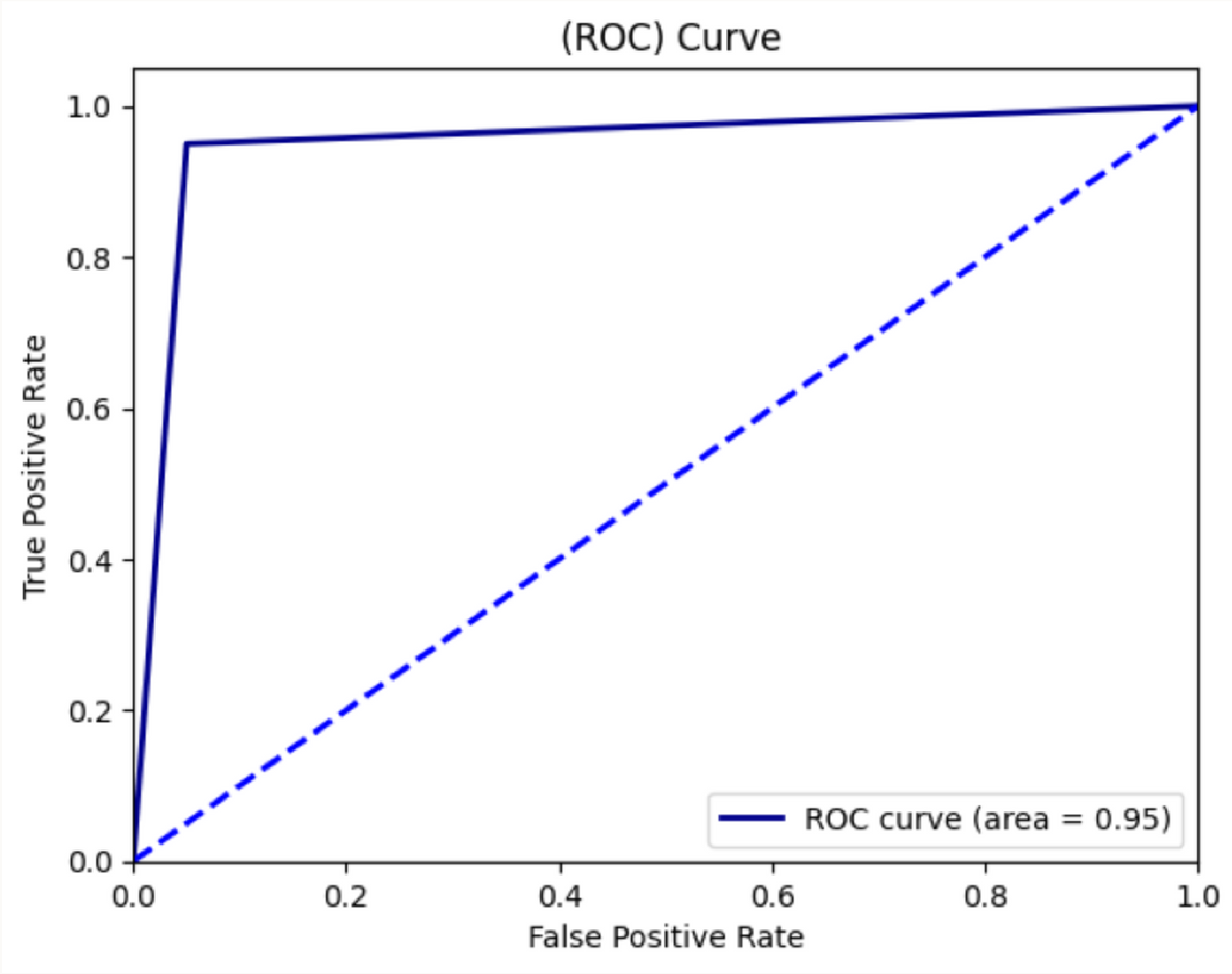
n_estimators = 200

METRIC RESULTS FOR RANDOM FOREST

PRECISION RECALL CURVE

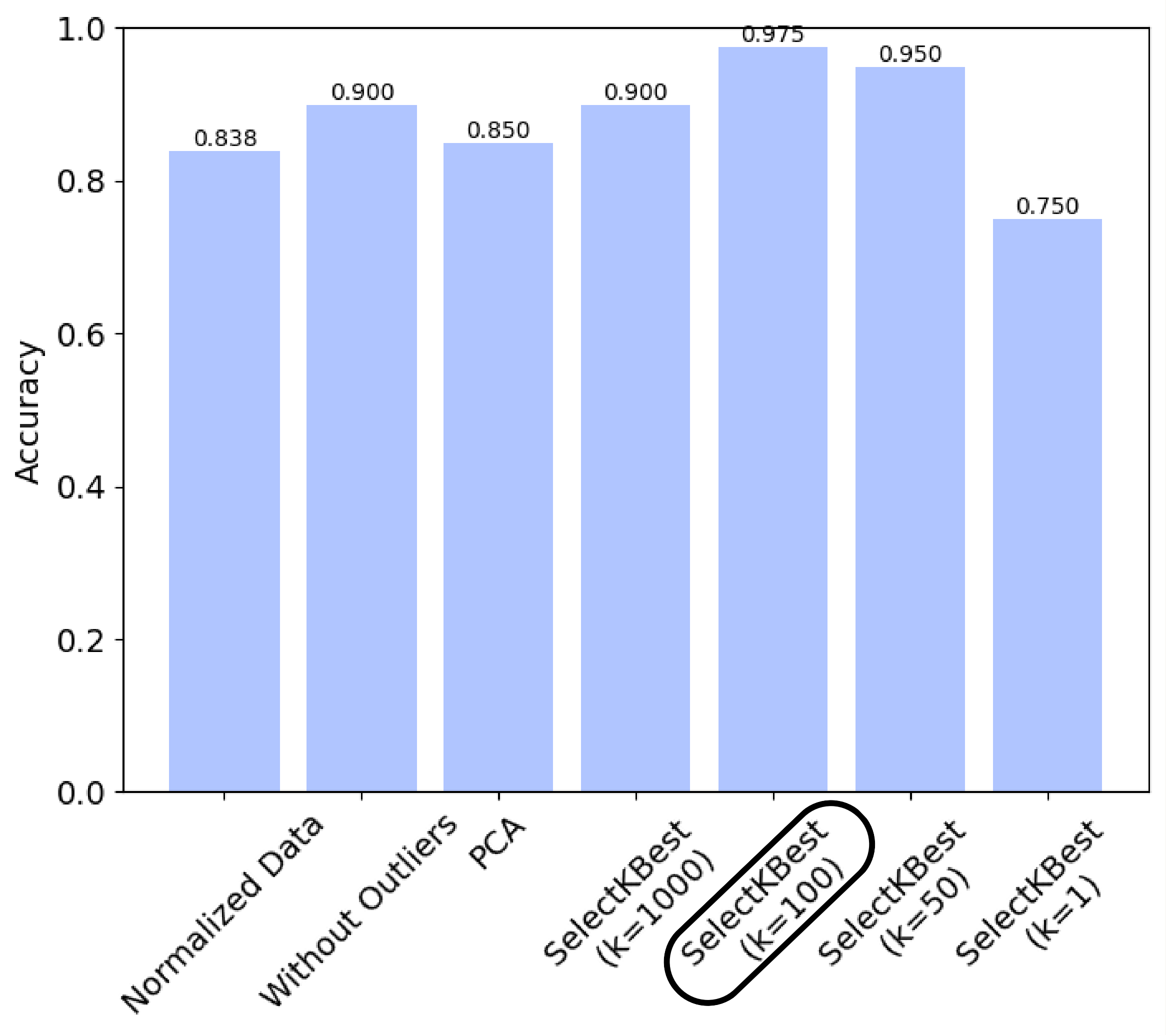


ROC CURVE

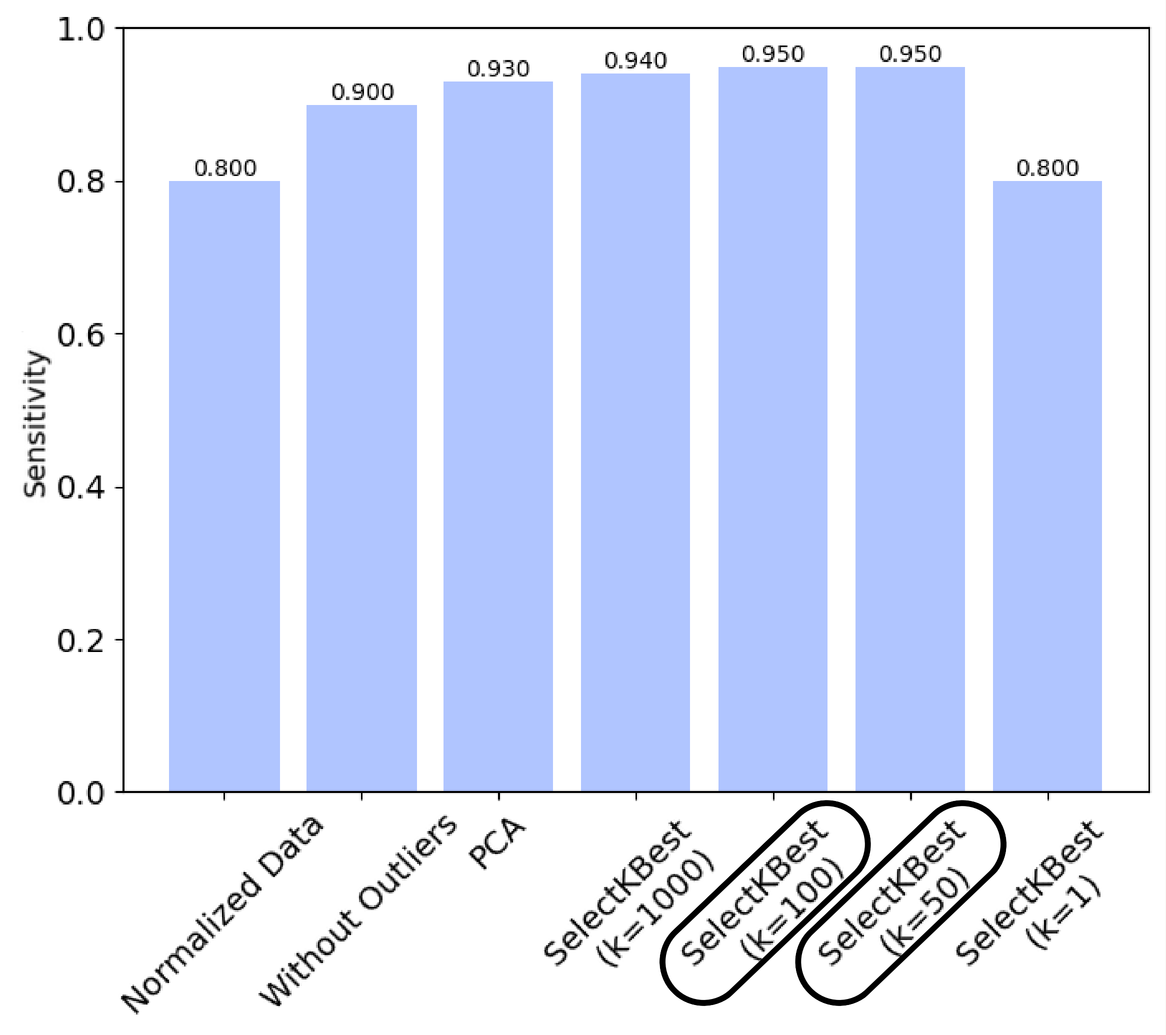


RESULT COMPARISON FOR RANDOM FOREST

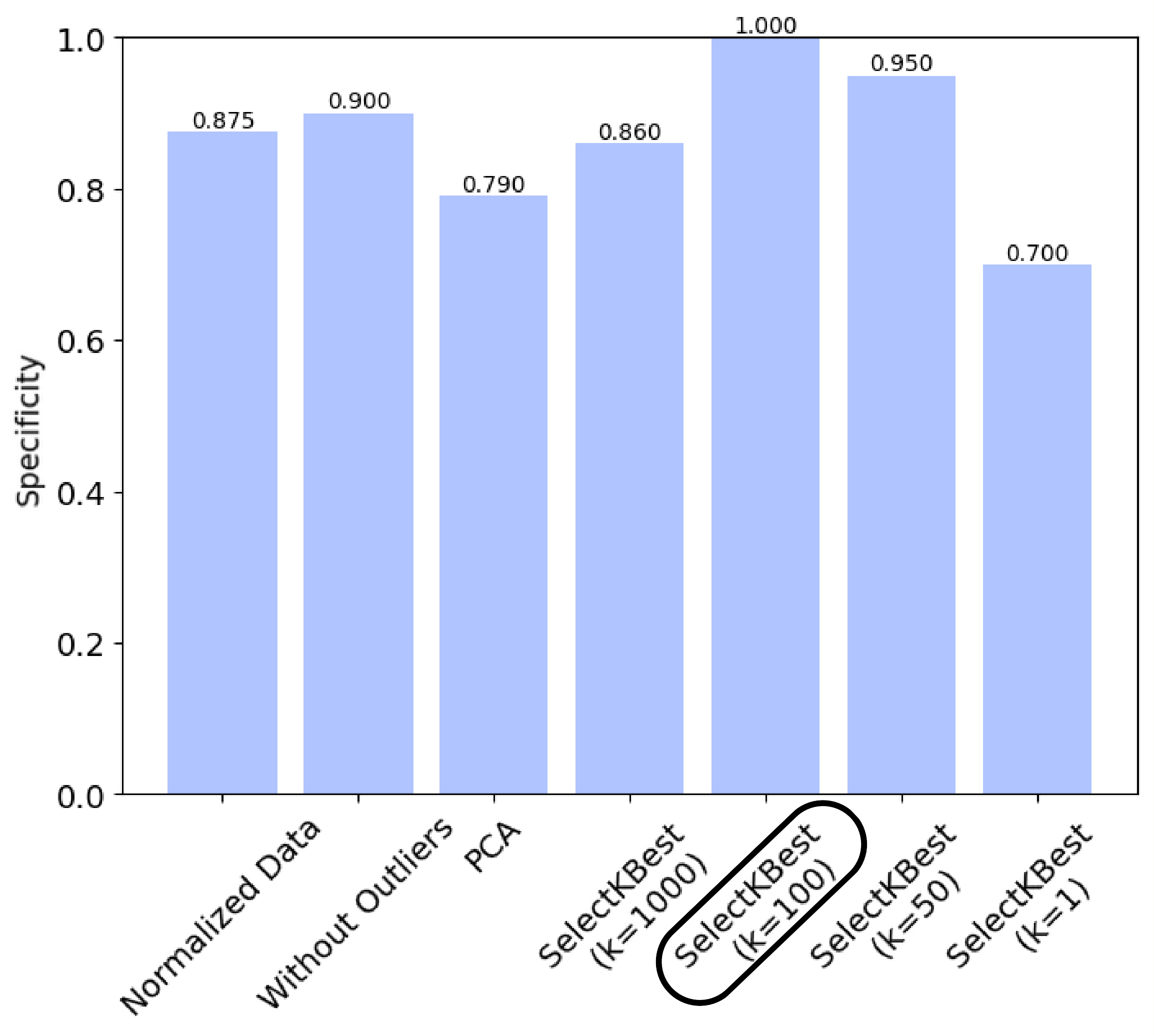
ACCURACY



SENSITIVITY TRUE POSITIVE RATE



SPECIFICITY TRUE NEGATIVE RATE



THE BEST MODELS

1 RANDOM FOREST

- OUTLIERS REMOVED
- FEATURE SELECTION SELECTKBEST(K=100)

ACCURACY

2 SVM

- OUTLIERS REMOVED
- FEATURE SELECTION SELECTKBEST(K=50)

ACCURACY

CONCLUSION

In our comparative study of machine learning models applied to a high-dimensional dataset with limited samples, the Random Forest algorithm paired with K-Best feature selection observed as the superior method, particularly when the top 100 features were retained. This model's success can be attributed to its ensemble nature. By leveraging multiple decision trees and focusing on the most statistically significant features, the Random Forest model demonstrated robust generalization capabilities, making it well-suited for scenarios where the feature space is large relative to the number of available data points.