**ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ**

ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

**Διπλωματική εργασία**

**Τίτλος:**

# Επιλογή παραμέτρων και χαρακτηριστικών ταξινομητή με χρήση της καμπύλης λειτουργίας δείκτη (ROC Curve)

**Δημάκου Ανθή**

**Επιτροπή:**

**Καθηγητής Μιχάλης Ζερβάκης (Επιβλέπων)**

**Καθηγητής Καλαϊτζάκης Κωνσταντίνος**

**Καθηγητής Σταυρακάκης Γεώργιος**

# Περιεχόμενα


Περιεχόμενα Εικόνων ..... 4

Περιεχόμενα Πινάκων ..... 8

1. Εισαγωγή ..... 9

1.1. Γενική περιγραφή του προβλήματος ..... 9

1.1.1. Τι είναι Classification ..... 9

1.1.2. Σε ποιους τομείς της επιστήμης χρησιμοποιείται το Classification ..... 10

1.1.3. Classification στους τομείς Ovarian Tumor, Brain Tumor, Breast Tumor 11

1.2. Στόχος Διπλωματικής Εργασίας ..... 12

1.3. Δομή της διπλωματικής εργασίας (Roadmap) ..... 13

2. Θεωρητικό Υπόβαθρο ..... 14

2.1. Classifiers και Classification ..... 14

2.1.1. Περιγραφή τύπων ταξινομητών ..... 14

2.1.2. Ταξινομητές που χρησιμοποιήθηκαν ..... 17

2.2. Confusion Matrix ..... 17

2.3. ROC ..... 18

2.3.1. ROC Curve ..... 19

2.3.2. ROC Space ..... 21

2.4. Confidence Intervals της καμπύλης ROC ..... 22

2.5. Επεξεργασία Δεδομένων ..... 23

2.5.1. Cross Validation ..... 23

2.5.2. K-Folds Cross Validation ..... 24

2.5.3. Stratified K-Folds Cross Validation ..... 25

2.6. Datasets – Τα δεδομένα που χρησιμοποιήθηκαν ..... 26

2.6.1. Ovarian Tumor Dataset ..... 26

2.6.2. Brain Tumor Dataset ..... 27

2.6.3. Breast Tumor Dataset ..... 28

3. Υλοποίηση ..... 30

3.1. Datasets, Training Sets, Testing Sets ..... 30

3.1.1. Μέγεθος και μορφή δεδομένων Testing και Training για Ovarian Tumor Dataset 31

3.1.2. Μέγεθος και μορφή δεδομένων Testing και Training για Brain Tumor Dataset 32

3.1.3. Μέγεθος και μορφή δεδομένων Testing και Training για Brain Tumor Dataset 32

3.2. Classification ..... 33

3.3. Σχεδίαση ROC ..... 36

3.4. Standard Deviation της Καμπύλης ROC ..... 38

4. Αξιολόγηση ..... 40

4.1. Classification ..... 40

4.1.1. Classification Ovarian Tumor Dataset ..... 40

4.1.1.1. Επιλογή μοντέλου ταξινόμησης για Ovarian Tumor Dataset ..... 46

4.1.2. Classification Brain Cancer Dataset ..... 46

4.1.2.1. Παρουσίαση αποτελεσμάτων Brain Cancer Dataset ..... 46

4.1.2.2. Επιλογή μοντέλου ταξινόμησης για Brain Cancer Dataset ..... 51

4.1.3. Classification Breast Cancer Dataset ..... 52

4.1.3.1. Παρουσίαση αποτελεσμάτων Breast Cancer Dataset ..... 52

4.1.3.2. Επιλογή μοντέλου ταξινόμησης για Breast Cancer Dataset ..... 57

4.2. Καμπύλες ROC ..... 57

4.2.1. Καμπύλες ROC 10-Folds ..... 57

4.2.1.1. Καμπύλες ROC 10 –Folds για Ovarian Tumor Dataset ..... 58

4.2.1.2. Καμπύλες ROC 10 –Folds για Brain Cancer Dataset .....60

4.2.1.3. Καμπύλες ROC 10 –Folds για Breast Cancer Dataset .....63

4.3. Καμπύλες Μέσης ROC με Standard Deviation.....65

Μέση καμπύλη ROC για Ovarian Tumor Dataset με std .....66

4.3.1. .....66

4.3.2. Μέση καμπύλη ROC για Brain Tumor Dataset με std .....69

Μέση καμπύλη ROC για Breast Cancer Dataset με std .....71

4.3.3. .....71

4.3.4. Επιλογή του καλύτερου μοντέλου για κάθε dataset .....73

4.3.4.1. Ovarian Tumor Dataset Classifier Selection.....74

4.3.4.2. Brain Tumor Dataset Classifier Selection .....75

4.3.4.3. Breast Tumor Dataset Classifier Selection .....77

5. Συμπεράσματα .....79

6. Μελλοντική Εργασία .....81

7. Αναφορές.....82

8. Παράρτημα .....85

8.1. Καμπύλες ROC Με std.....85

8.1.1. Ovarian Tumor Dataset.....85

8.1.2. Brain Cancer Dataset .....102

8.1.3. Breast Cancer Dataset.....119

## Περιεχόμενα Εικόνων

Εικόνα 1 : Μορφή Καμπύλης ROC.....20

Εικόνα 2 : Χώρος καμπύλης ROC – ROC Space.....22

Εικόνα 3 : Δείγματα ασθενών Καρκίνου Ωοθηκών ανά Ογκολογικό Κέντρο.....26

Εικόνα 4: Τρόπος επεξεργασίας των datasets με τη μέθοδο K-Folds Cross Validation .....31
Εικόνα 5 : Δημιουργία Training Sets , Evaluation Sets, Independent Test Set για τα 10 – Folds από τα Datasets .....31
Εικόνα 6: Διαδικασία και αποτελέσματα Classification για κάθε FOLD .....35
Εικόνα 7: Block Diagram Υπολογισμού των σημείων της καμπύλης ROC .....37
Εικόνα 8 Υπολογισμός του Standard Deviation για κάθε σημείο της καμπύλης ROC 39
Εικόνα 9: Ovarian Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set –Loglc Classifier .....42
Εικόνα 10: Ovarian Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set –Parzenc Classifier .....42
Εικόνα 11: Ovarian Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set –Polyc Classifier.....43
Εικόνα 12: Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers - Ovarian Tumor Dataset – Evaluation Set .....43
Εικόνα 13: Ovarian Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers - Independent Set .....44
Εικόνα 14 : Ovarian Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Evaluation Set .....45
Εικόνα 15: Ovarian Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Independent Set.....46
Εικόνα 16: Brain Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Nmsc Classifier .....48
Εικόνα 17 : Brain Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Klldc Classifier .....49
Εικόνα 18: Brain Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Knnc Classifier ..... **Error! Bookmark not defined.**
Εικόνα 19: Brain Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers – Evaluation Set .....49
Εικόνα 20 : Brain Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers – Independent Set.....50

Εικόνα 21 : Brain Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Evaluation Set .....50
Εικόνα 22: Brain Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Independent Set .....51
Εικόνα 23: Breast Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Klldc Classifier .....53
Εικόνα 24: Breast Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Polyc Classifier.....54
Εικόνα 25: Breast Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Parzenc Classifier .....54
Εικόνα 26: Breast Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers – Evaluation Set .....55
Εικόνα 27: Breast Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers – Independent Set.....55
Εικόνα 28 : Breast Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Evaluation Set .....56
Εικόνα 29 : Breast Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Independent Set .....57
Εικόνα 30 : Ovarian Tumor Dataset – Roc 10 Folds – Fisherc Classifier – Evaluation Set .....58
Εικόνα 31 : Ovarian Tumor Dataset – Roc 10 Folds – Fisherc Classifier – Independent Set .....59
Εικόνα 33 : Ovarian Tumor Dataset – Roc 10 Folds – Klldc Classifier – Independent Set .....60
Εικόνα 34 : Brain Tumor Dataset – Roc 10 Folds – Naïvebc Classifier – Evaluation Set .....61
Εικόνα 35 : Brain Tumor Dataset – Roc 10 Folds – Naïvebc Classifier – Independent Set .....61
Εικόνα 36 : Brain Tumor Dataset – Roc 10 Folds – Udc Classifier – Evaluation Set ....62
Εικόνα 37 : Brain Tumor Dataset – Roc 10 Folds – Udc Classifier – Independent Set 62
Εικόνα 38 : Breast Tumor Dataset – Roc 10 Folds – Polyc Classifier – Evaluation Set 63

Εικόνα 39 : Breast Tumor Dataset – Roc 10 Folds – Polyc Classifier – Independent Set .....64
Εικόνα 41 : Breast Tumor Dataset – Roc 10 Folds – Ldc Classifier – Independent Set65
Εικόνα 42 : Ovarian Cancer Dataset – Roc & Standard Deviation – Udc Classifier – Evaluation Set .....66
Εικόνα 43 : Ovarian Cancer Dataset – Roc & Standard Deviation – Udc Classifier – Independent Set.....67
Εικόνα 45 : Ovarian Cancer Dataset – Roc & Standard Deviation – Naivebc Classifier – Independent Set.....68
Εικόνα 47 : Brain Cancer Dataset – Roc & Standard Deviation – Nmsc Classifier – Independent Set.....69
Εικόνα 48 : Brain Cancer Dataset – Roc & Standard Deviation – Loglc Classifier – Evaluation Set .....70
Εικόνα 49 : Brain Cancer Dataset – Roc & Standard Deviation – Loglc Classifier – Independent Set.....70
Εικόνα 50 : Breast Cancer Dataset – Roc & Standard Deviation – Polyc Classifier – Evaluation Set .....71
Εικόνα 51 : Breast Cancer Dataset – Roc & Standard Deviation – Polyc Classifier – Independent Set.....72
Εικόνα 52 : Breast Cancer Dataset – Roc & Standard Deviation – Ldc Classifier – Evaluation Set .....72
Εικόνα 53 : Breast Cancer Dataset – Roc & Standard Deviation – Ldc Classifier – Independent Set.....73
Εικόνα 54: Ovarian Tumor Dataset – ROC Curve – 17 Classifiers – Evaluation Set.....74
Εικόνα 55 : Ovarian Tumor Dataset – ROC Curve – 17 Classifiers – Independent Set 74
Εικόνα 56 : Brain Tumor Dataset – ROC Curve – 17 Classifiers – Evaluation Set .....75
Εικόνα 57 : Brain Tumor Dataset – ROC Curve – 17 Classifiers – Independent Set ....76
Εικόνα 58 : Breast Tumor Dataset – ROC Curve – 17 Classifiers – Evaluation Set .....77
Εικόνα 59 : Breast Tumor Dataset – ROC Curve – 17 Classifiers – Independent Set ..77

## Περιεχόμενα Πινάκων

Πίνακας 1 : Πίνακας Confusion Matrix ..... 18

Πίνακας 2: Πλήθος Δεδομένων ανά test και train set , Δείγματα ανά κλάση στο Ovarian Tumor Dataset ..... 32

Πίνακας 3: Πλήθος Δεδομένων ανά test και train set , Δείγματα ανά κλάση στο Brain Tumor Dataset ..... 32

Πίνακας 4: Πλήθος Δεδομένων ανά test και train set , Δείγματα ανά κλάση στο Breast Tumor Dataset ..... 33

Πίνακας 5: Ovarian Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Evaluation Set ..... 41

Πίνακας 6: Ovarian Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Independent Set... 41

Πίνακας 7 : Brain Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Evaluation Set ..... 47

Πίνακας 8: Brain Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Independent Set... 47

Πίνακας 9 : Breast Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Evaluation Set ..... 52

Πίνακας 10 : Breast Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Evaluation Set ..... 53

# 1. Εισαγωγή

Η ταξινόμηση είναι μία ευρεία έννοια με εφαρμογή σε διάφορους τομείς της ανθρώπινης δραστηριότητας. Μας απαλλάσσει από την άτακτη και συγκεχυμένη πληροφορία και μας προσφέρει μια ποιο δομημένη άποψη της πραγματικότητας, απαλλαγμένη από περιττές πληροφορίες. Η ταξινόμηση μας δίνει επίσης τη δυνατότητα διάκρισης μεταξύ ομοειδών πραγμάτων και κατά συνέπεια διαφορετική μεταχείριση σε διαφορετικές κατηγορίες και είδη. Θα προσπαθήσουμε να προσεγγίσουμε την έννοια της ταξινόμησης και την εφαρμογή της σε έναν πολύ συγκεκριμένο τομέα της επιστήμης, την ταξινόμηση ιατρικών δεδομένων. Στόχος μας είναι να επιλέξουμε το ιδανικότερο μοντέλο που θα μπορεί να ταξινομήσει αντικείμενα με τον πλέον βέλτιστο τρόπο. Για το σκοπό αυτό συγκρίνουμε διάφορα μοντέλα ταξινόμησης και αναπτύσσουμε μεθόδους σύγκρισης και αξιολόγησης των μοντέλων αυτών. Η τελική προσέγγιση του προβλήματος εκτός από την επιλογή του βέλτιστου μοντέλου θα μας προσδιορίσει και την περιοχή στην οποία μπορεί να κινηθεί το μοντέλο ώστε τα αποτελέσματα να είναι ακριβή σε κάθε περίπτωση.

## 1.1. Γενική περιγραφή του προβλήματος

Στο σημείο αυτό θα αποσαφηνιστεί ο όρος ταξινόμηση (classification) και θα προσδιοριστεί ο τρόπος χρήσης της στο συγκεκριμένο πρόβλημα.

### 1.1.1. Τι είναι Classification

Η στατιστική ταξινόμηση (statistical classification ή απλά classification), είναι μια διαδικασία κατά την οποία αντικείμενα χωρίζονται σε κατηγορίες (κλάσεις) με βάση τις τιμές ενός σαφώς ορισμένου συνόλου χαρακτηριστικών (features), καθώς και την πρότερη γνώση που έχει αποκτηθεί από προηγουμένως ορθά ταξινομημένα αντικείμενα. [1],[2]

Η διαδικασία αυτή γίνεται με τους ταξινομητές (classifiers). Με τους Classifiers πετυχαίνουμε την απεικόνιση ενός συνεχούς ή διακριτού πεδίου χαρακτηριστικών (feature space) σε ένα διακριτό και προκαθορισμένο σετ κατηγοριών (set of labels). Ο τρόπος με τον οποίο θα πραγματοποιηθεί αυτή η απεικόνιση, περιγράφεται στον classifier μέσω της διαδικασίας εκπαίδευσης, κατά την οποία μπορεί να αντλήσει

γνώση μέσω ενός συνόλου δεδομένων-αντικειμένων (training set) των οποίων η κατηγορία είναι από πριν γνωστή. Η διαδικασία αυτή καλείται Supervised Learning [3].

Υπάρχουν πολλοί τύποι ταξινομητών, καθένας από τους οποίους βασίζει τη λειτουργία του σε διαφορετικά μαθηματικά μοντέλα και μηχανισμούς. Απώτερος στόχος μας είναι να εντοπιστεί ο classifier με τον οποίο μπορεί να βελτιστοποιηθεί η ποιότητα αντιστοίχισης μεταξύ των χαρακτηριστικών ενός αντικειμένου (feature set) και της κατηγορίας στην οποία αυτό πραγματικά ανήκει.

Συνοψίζοντας, στόχος της διαδικασίας ταξινόμησης είναι ο εντοπισμός του καλύτερου δυνατού μοντέλου, το οποίο εφόσον εκπαιδευτεί με ένα κατάλληλα επιλεγμένο σετ δεδομένων (training data), θα μπορέσει να ταξινομήσει αξιόπιστα άλλα -ανεξάρτητα από το σετ εκπαίδευσης- δεδομένα.

### 1.1.2. Σε ποιους τομείς της επιστήμης χρησιμοποιείται το Classification

Η διαδικασία classification έχει χρησιμοποιηθεί σε πολλούς τομείς της επιστήμης όπως στη μηχανική όραση, ανίχνευση ναρκωτικών ουσιών, γεωσταστιστική, αναγνώριση φωνής, αναγνώριση γραφικού χαρακτήρα, αναγνώριση βιομετρικών χαρακτηριστικών, επεξεργασία φυσικής γλώσσας, μηχανές αναζήτησης κ.α.

Ο τομέας που εξετάζουμε στα πλαίσια αυτής της εργασίας είναι η επεξεργασία ιατρικών δεδομένων. Χρησιμοποιώντας ιατρικά δεδομένα τα οποία σχετίζονται με πληροφορίες και χαρακτηριστικά ενός προκαθορισμένου συνόλου ασθενών, θα εκπαιδεύσουμε ένα πλήθος διαφορετικών classifiers, τους οποίους στη συνέχεια θα χρησιμοποιήσουμε για να αξιολογήσουμε ένα σύνολο ασθενείς – διαφορετικό από αυτό της εκπαίδευσης. Η κατηγοριοποίηση των ασθενών ποικίλει ανάλογα με το πρόβλημα που εξετάζουμε κάθε φορά (π.χ. αν ο ασθενής είναι δεκτικός σε μία συγκεκριμένη θεραπεία ή όχι κλπ.)

Πιο συγκεκριμένα, τα ιατρικά δεδομένα που χρησιμοποιήθηκαν τόσο για την εκπαίδευση όσο και για την αξιολόγηση των classifiers αφορούσαν σε ασθενείς που είχαν κάποιο τύπο όγκου. Στόχος μας είναι να χρησιμοποιήσουμε την τεχνολογία

του classification προκειμένου να προσδιορίσουμε τον τύπο του όγκου που μπορεί να έχει κάποιος ασθενής, καθώς και τον εντοπισμό του classifier εκείνου που με την κατάλληλη εκπαίδευση μπορεί να μας δώσει τα πιο αξιόπιστα αποτελέσματα.

### 1.1.3. Classification στους τομείς Ovarian Tumor, Brain Tumor, Breast Tumor

Τα δεδομένα που χρησιμοποιήθηκαν (dataset) έχουν τη μορφή πινάκων. Κάθε γραμμή αντιστοιχεί σε έναν ασθενή – δείγμα, με κάθε στήλη να αντιπροσωπεύει ένα συγκεκριμένο χαρακτηριστικό του. Τέτοια μπορεί να είναι η ηλικία, το βάρος κλπ. Στα πλαίσια της εργασίας χρησιμοποιήθηκαν τρεις τέτοιοι πίνακες, οι οποίοι περιέχουν τα χαρακτηριστικά ασθενών που μπορεί να έχουν καλοήθεις ή κακοήθεις όγκους, στις ωοθήκες, στον εγκέφαλο και στο μαστό, αντίστοιχα.

Ο καρκίνος των ωοθηκών είναι η πέμπτη αιτία θανάτου από καρκίνο στις γυναίκες παγκοσμίως ενώ είναι η πρώτη αιτία θανάτου από γυναικολογικό καρκίνο. Μια γυναίκα ισοβίως και ανεξαρτήτου ηλικίας έχει 1.5% πιθανότητες να νοσήσει από καρκίνο των ωοθηκών, πράγμα που καθιστά τον καρκίνο των ωοθηκών δεύτερη αιτία κακοήθειας μετά από τον καρκίνο του μαστού. Ο καρκίνος των ωοθηκών δεν έχει ιδιαίτερα χαρακτηριστικά συμπτώματα, ειδικά στα πρώτα στάδια, και για το λόγο αυτό είναι πολύ δύσκολο να εντοπιστεί, [4]. Το γεγονός αυτό σε συνδυασμό με το ότι υπάρχουν τέσσερα διαφορετικά είδη όγκων στις ωοθήκες [5] καθιστά επιτακτική την ανάγκη για classification στον τομέα αυτό. Στόχος είναι με τα ιατρικά δεδομένα που έχουμε να αποφασίσουμε σε ποια κατηγορία όγκων ανήκει ο όγκος που έχει ένας συγκεκριμένος ασθενής ώστε να μπορεί να του χορηγηθεί κατάλληλη θεραπεία η οποία είναι διαφορετική κατά περίπτωση ασθενή και τύπου όγκου. Επιπλέον είναι σημαντικό να γνωρίζουν οι γιατροί κατά πόσον ο όγκος αυτός είναι εγχειρήσιμος ή όχι.

Τα δεδομένα του δεύτερου dataset που εξετάζουμε αφορούν σε εγκεφαλικούς όγκους. Οι εγκεφαλικοί όγκοι [6] είναι μη φυσιολογικές μάζες που αναπτύσσονται στο εσωτερικό του κρανίου και μπορεί να είναι καρκινικοί ή όχι. Προκαλούνται από μη φυσιολογική και ανεξέλεγκτη διαίρεση εγκεφαλικών κυττάρων, τους νευρώνες ή προέρχονται από καρκινικά κύτταρα που εξαπλώθηκαν στο ανθρώπινο σώμα από

άλλη μορφή καρκίνου. Και στην περίπτωση των εγκεφαλικών όγκων υπάρχουν διάφοροι τύποι όγκων, η ταυτοποίηση των οποίων συνοδεύεται από διάφορες νευρολογικές εξετάσεις και αναλύσεις εγκεφαλικού ιστού. Οι πληροφορίες από τις εξετάσεις αυτές χρησιμοποιούνται για την κατάταξη των όγκων σε λιγότερο επιθετικούς (καλοήθεις) και σε περισσότερο επιθετικούς (κακοήθεις) ώστε να μπορέσουν οι γιατροί να χορηγήσουν στον ασθενή τη θεραπεία που αρμόζει.

Τέλος, ο καρκίνος του μαστού [7] είναι ο δεύτερος συχνότερος τύπος καρκίνου με ποσοστό 10.4% και στα δύο φύλα και η πέμπτη αιτία θανάτου από καρκίνο παγκοσμίως. Στις γυναίκες είναι εκατό φορές πιο συχνό φαινόμενο, ενώ οι πιθανότητες επιβίωσης είναι ίδια σε άντρες και γυναίκες. Και σε αυτή την περίπτωση η έγκαιρη διάγνωση και ο προσδιορισμός του τύπου όγκου σε κάθε ασθενή μπορεί να αποβεί σωτήρια.

## 1.2. Στόχος Διπλωματικής Εργασίας

Ο στόχος της διπλωματικής εργασίας είναι η αξιολόγηση των επιδόσεων ενός πλήθους ταξινομητών (classifier performance) στον προσδιορισμό του τύπου του όγκου που μπορεί να έχει ένας ασθενής, προκειμένου να επιλέξουμε τον πιο κατάλληλο – αξιόπιστο για κάθε μία από τις τρεις ασθένειες.

Αρχικά επεξεργαστήκαμε τα διαθέσιμα ιατρικά δεδομένα, προκειμένου να διαμορφώσουμε το σύνολο που θα χρησιμοποιηθεί για την εκπαίδευση των classifiers (training data), καθώς και αυτό που θα χρησιμοποιηθεί για τη διαδικασία της αξιολόγησης (testing data).

Κατόπιν αξιολογήσαμε την επίδοση όλων των διαθέσιμων classifier (classifier performance), σε κάθε μία από τις τρεις ασθένειες. Κατά την αξιολόγηση χρησιμοποιήσαμε το ανεξάρτητο δείγμα αξιολόγησης (testing data) , το οποίο είναι κοινό για όλους τους classifiers σε κάθε dataset . Το δείγμα αξιολόγησης εφαρμόζεται στα training data τα οποία είναι μοιρασμένα σε δέκα τμήματα με συγκεκριμένη μέθοδο, για πρόσθετη αξιοπιστίας, λόγω μικρού αριθμού δειγμάτων εκπαίδευσης στο dataset. Έτσι καταλήγουμε να έχουμε δέκα διαφορετικές

αξιολογήσεις για τον ίδιο classifier, από τις οποίες προκύπτει ένας μέσος σαν τελικό συμπέρασμα της απόδοσης του classifier.

Τέλος, με τη χρήση της καμπύλης της λειτουργίας δείκτη (ROC), προσπαθήσαμε για κάθε classifier να εντοπίσουμε τις συνθήκες (το σημείο λειτουργίας) κάτω από τις οποίες φαίνεται να μεγιστοποιείται η ακρίβεια ή/και η ευαισθησία του.

## 1.3. Δομή της διπλωματικής εργασίας (Roadmap)

Οι εργασίες που πραγματοποιήθηκαν στα πλαίσια της διπλωματικής εργασίας περιγράφονται αναλυτικά στις τέσσερις ενότητες που ακολουθούν. Πιο συγκεκριμένα, στην ενότητα 2 αναλύεται το θεωρητικό υπόβαθρο που απαιτείται να έχει ο αναγνώστης για την κατανόηση του προβλήματος που εξετάζουμε, καθώς και η φύση των ιατρικών δεδομένων που χρησιμοποιήθηκαν. Στην ενότητα 3 περιγράφεται η μεθοδολογία που ακολουθήσαμε για επίλυση του προβλήματος, καθώς και των υποσυστημάτων από αναπτύχθηκαν για το σκοπό αυτό. Ακολουθεί η ενότητα 4, στην οποία εμφανίζουμε και αξιολογούμε την επίδοση του κάθε classifier, ενώ τέλος στην ενότητα 5 διατυπώνουμε τα συμπεράσματα που προέκυψαν στα πλαίσια της εργασίας καθώς και μελλοντικές κατευθύνσεις που θα μπορούσαν να υπάρξουν.

# 2. Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό γίνεται μια δομημένη προσέγγιση του προβλήματος, εξετάζουμε το θεωρητικό υπόβαθρο των όλων μεθόδων που χρησιμοποιήθηκαν και αναφέρουμε άλλες προσεγγίσεις που τυχόν έγιναν στο παρελθόν στο πεδίο που εξετάζουμε. Ποιο συγκεκριμένα θα δούμε τα είδη των classifiers που χρησιμοποιήσαμε , τους ίδιους τους classifiers, την χρησιμότητα του Confusion Matrix, τον τρόπο υπολογισμού της καμπύλης ROC και το συγκριτικό της πλεονέκτημα σε σχέση με άλλες μεθόδου αξιολόγησης ταξινομητών, την τυπική απόκλιση της καμπύλης ROC που χρησιμοποιείται ως εργαλείο μέτρησης αξιοπιστίας της ROC, τον τρόπο με τον οποίο επεξεργαστήκαμε τα ιατρικά δεδομένα και τις μεθόδους με τις οποίες δημιουργήσαμε τα training data και testing data (Stratified K – Folds Cross validation) , τα ίδια τα ιατρικά δεδομένα και τα χαρακτηριστικά τους.

## 2.1. Classifiers και Classification

Ο τύπος των ταξινομητών που χρησιμοποιήθηκαν είναι αρκετά σημαντικός για να κατανοήσουμε όχι μόνο τη δομή και λειτουργία του κάθε classifier ξεχωριστά αλλά και για να ερμηνεύσουμε την συμπεριφορά των μοντέλων αυτών αργότερα στη διαδικασία της αξιολόγησης.

### 2.1.1. Περιγραφή τύπων ταξινομητών

Linear [8]: Ο στόχος του linear classification είναι να ομαδοποιηθούν χαρακτηριστικά του dataset που έχουν όμοιες τιμές. Αυτό επιτυγχάνεται καθώς η απόφαση του classification βασίζεται στις τιμές του γραμμικού συνδυασμού των χαρακτηριστικών. Αν έχουμε είσοδο το διάνυσμα $\vec{x}$ η έξοδος θα είναι:

$$y = f(\vec{w} * \vec{x}) = f(\sum_j w_j x_j) ,$$

Όπου $\vec{w}$ είναι το διάνυσμα από βάρη και $f$ η συνάρτηση που συνδέει το εσωτερικό γινόμενο των δύο διανυσμάτων με την επιθυμητή έξοδο. Το $\vec{w}$ εκπαιδεύεται από ένα σετ δειγμάτων. Στην περίπτωσή μας που έχουμε δύο κλάσεις μπορούμε να δούμε την δράση του γραμμικού classifier σαν έναν διαχωρισμό των διαστάσεων της εισόδου με μία γραμμή όπου τα δείγματα από τη μία μεριά της γραμμής

κατατάσσονται στην κλάση ‘1’ και τα δείγματα από την άλλη μεριά της γραμμής στην κλάση ‘0’. Γραμμικοί classifiers είναι οι: klldc, ldc, nmsc, nmc, pcldc, fisherc

K-Nearest [9],[10]: Χρησιμοποιούν τον k-nearest neighbors αλγόριθμο, ο οποίος κατηγοριοποιεί τα αντικείμενα βασιζόμενος στο κοντινότερο παράδειγμα κατηγοριοποίησης που έχει από το δείγμα της εκπαίδευσης. Τα δείγματα εκπαίδευσης είναι διανύσματα σε ένα πολυδιάστατο feature space. Το διάστημα αυτό χωρίζεται σε περιοχές μέσα στις οποίες οροθετείται η μία κλάση ή άλλη. Ένα σημείο στο χώρο καθορίζει την συγκεκριμένη κλάση αν είναι η πιο συνήθης επιλογή μεταξύ των k nearest δειγμάτων εκπαίδευσης. Συνήθως, η Ευκλείδεια απόσταση χρησιμοποιείται σαν μετρική. K-Nearest classifiers είναι οι : knnc

Logistic regression [11]: Είναι μια μέθοδος classification , ένα δυαδικό μοντέλο , που χρησιμοποιείται για την πρόβλεψη της πιθανότητας να συμβεί ένα γεγονός, συνδυάζοντας τα δεδομένα σε μία εκθετική καμπύλη $f(x) = {}^{1}/_{1 + e^{-x}}$ . Είναι ένα είδος γραμμικού classification. Logistic classifiers: loglc

Naïve Bayes [12]: Είναι μια απλή υλοποίηση του θεωρήματος του Bayes με ισχυρά ανεξάρτητες αποφάσεις. Είναι γενικά μια μέθοδος που η απόφαση δεν εξαρτάται από τα χαρακτηριστικά του classifier. Δηλαδή συμπεριφέρεται σε όλα τα χαρακτηριστικά σαν να ήταν ίδιας σπουδαιότητας και τα λαμβάνει ως ισάξια για τον υπολογισμό της πιθανότητας για να κατηγοριοποιήσει σε ποια κλάση ανήκει το δείγμα. Classifiers: naivebc

Parzen [13]: Όμοια υλοποίηση με την k-nearest μέθοδο μόνο που αντί για αριθμό γειτονικών σημείων χρησιμοποιείται η smoothing παράμετρος η οποία μπορεί να δίνεται από το χρήστη ή να βελτιστοποιείται από τη μέθοδο leave -one-out κατά το training. Classifiers: parzen, parzendc

Polynomial Classifiers : Είναι μία μορφή γραμμικού classifier αλλά με μεγαλύτερο βαθμό. Προσθέτει πολυωνυμικά χαρακτηριστικά στο dataset και τρέχει τον untrained classifier claasf. Για δευτέρου βαθμού πολυώνυμα δημιουργούνται κανόνες προτεραιότητας. Για μεγαλύτερο βαθμό όχι. Polynomial Classifiers: polyc

Normal Density based Classifiers : Οι classifiers της κατηγορίας αυτής χρησιμοποιούν την συνάρτηση κανονικής κατανομής και αυτό εφαρμόζεται σε γραμμικούς, quadratic αλλά και μη γραμμικούς classifiers. Classifiers: ldc, qdc, udc

Quadratic Classifiers : Καθορίζει ένα τετραγωνικό κατώφλι σύμφωνα με το οποίο αποφασίζει τη μία ή την άλλη κλάση. Έχει παρόμοια υλοποίηση με τον γραμμικό classifier. Η περιοχή καθορίζεται από ένα σετ διανύσματα το οποίο θεωρείται ως training set. Το πρόβλημα εστιάζεται στο να αποφασίζουμε για κάθε νέο σετ διανυσμάτων ποια είναι η σωστή επιλογή κλάσης. Αυτό δίνεται $y = x^T Ax + b^T + c$ . Classifiers: quadrc

Neural Network Classifiers [14]: Χρησιμοποιεί ένα μαθηματικό μοντέλο βασισμένο σε νευρωνικά δίκτυα. Αποτελείται από ένα γκρουπ διασυνδεμένων τεχνιτών νευρώνων και επεξεργάζεται τις πληροφορίες μέσω της συνάφειας. Το μοντέλο αυτό προσαρμόζεται και αλλάζει τη δομή του ανάλογα με τις εξωτερικές ή εσωτερικές πληροφορίες που παίρνει κατά την εκπαίδευση. Classifiers: rbnc

Decision Trees [15]: Ένα δενδρικό διάγραμμα είναι ένα μοντέλο λήψης αποφάσεων που χρησιμοποιεί την γραφική δενδρική δομή και τις πιθανές επιπτώσεις συνυπολογίζοντας τις πιθανότητες, τα κόστη και τη χρησιμότητα μιας απόφασης. Classifiers: treec

### 2.1.2. Ταξινομητές που χρησιμοποιήθηκαν

Έπειτα από της περιγραφή των τύπων ταξινομητών που χρησιμοποιήθηκαν ακολουθεί η παρουσίαση των ταξινομητών αυτών. Οι δεκαεπτά classifiers [16] που χρησιμοποιήθηκαν είναι οι εξής : Fisherc (Fisher's Least Square Linear Classifier), Klldc (Linear classifier built on the KL expansion of the common covariance matrix) , Knnc (K-Nearest Neighbor Classifier, Ldc (Linear Bayes Normal Classifier), Loglc (Logistic linear classifier), Naivebc (Naive Bayes classifier), Nmc (Nearest Mean Classifier) , Nmsc (Nearest Mean Scaled Classifier) , Parzenc (Optimisation of the Parzen classifier) , Parzendc (Parzen density based classifier) , Pcldc (Linear classifier using PC expansion on the joint data) , Polyc (Polynomial Classification) , Qdc (Quadratic Bayes Normal Classifier) , Quadrc (Quadratic Discriminant Classifier) , Rbnc (Radial basis function neural network classifier), Treec (Build a decision tree classifier) , Udc (Uncorrelated normal based quadratic Bayes classifier).

## 2.2. Confusion Matrix

Το Confusion Matrix είναι ένα εργαλείο αναπαράστασης που χρησιμοποιείται στον τομέα supervised learning, στην διαδικασία δηλαδή της πρόβλεψης της τιμής μιας συνάρτησης αφού έχουν προηγηθεί οι τιμές της συνάρτησης για τα δεδομένα εκπαίδευσης. Είναι δηλαδή μια άλλη μετρική για να προσδιοριστεί το πόσο καλά αναγνωρίζει ή όχι το μοντέλο ταξινόμησης την κλάση στην οποία ανήκει ένα αντικείμενο.

Το Confusion Matrix χρησιμοποιείται γιατί πολλές φορές το Error Rate δεν μας δίνει αντιπροσωπευτικά αποτελέσματα για την επίδοση ενός classifier, ειδικά στις περιπτώσεις που η συντριπτική πλειοψηφία δειγμάτων ανήκει σε μία συγκεκριμένη κλάση. Όταν συμβαίνει αυτό, αν για παράδειγμα έχουμε ένα dataset με 1000 δείγματα από τα οποία τα 990 ανήκουν στην κλάση A και μόνο 10 στην κλάση B, αν ο classifier αποφασίσει να κατατάξει όλα τα δείγματα στην κλάση A τότε το Error Rate θα είναι ίσο με 99%, όπου αυτό όμως δεν είναι ικανοποιητική ένδειξη για τον classifier αφού στην πραγματικότητα η απόδοση θα είναι 100% για την κλάση A και 0% για τη B. Για αυτούς τους λόγους χρειαζόμαστε πιο ενδεικτικές μετρικές όπως το Confusion Matrix.

Το Confusion Matrix αποτελεί έναν πίνακα του οποίου κάθε κολώνα αναπαριστά τα προβλεπόμενα δείγματα μιας συγκεκριμένης κλάσης ενώ κάθε γραμμή του αναπαριστά τα πραγματικά δείγματα κάθε κλάσης. Στον πίνακα 1 βλέπουμε την μορφή του Confusion Matrix:

| | Actual P | Actual N | total |
|---|---|---|---|
| Predicted P | **True Positive (TP)** | **False Positive (FP)** | |
| Predicted N | **False Negative(FN)** | **True Negative(TN)** | |
| total | | | |

Πίνακας 1 : Πίνακας Confusion Matrix

Ο πίνακας αυτός αποτελείται από τα εξής στοιχεία:

**TP**: Τα δείγματα που στην πραγματικότητα ανήκουν στην κλάση A (positive) και αναγνωρίστηκαν σωστά, σαν δείγματα της κλάσης A

**FP**: Τα δείγματα που στην πραγματικότητα ανήκουν στην κλάση A (positive) και αναγνωρίστηκαν λάθος, σαν δείγματα της κλάσης B

**FN**: Τα δείγματα που στην πραγματικότητα ανήκουν στην κλάση B (negative) και αναγνωρίστηκαν λάθος, σαν δείγματα της κλάσης A

**TN**: Τα δείγματα που στην πραγματικότητα ανήκουν στην κλάση B (negative) και αναγνωρίστηκαν σωστά, σαν δείγματα της κλάσης B

## 2.3. ROC

Η καμπύλη ROC είναι το βασικότερο εργαλείο που χρησιμοποιήθηκε για την επιλογή του βέλτιστου μοντέλου ταξινόμησης, όπως θα δούμε και παρακάτω, στον τρόπο υπολογισμού της, η ROC αποτελείται κατά μία έννοια από πολλά Confusion Matrices. Η χρήση της καμπύλης ROC είναι αναγκαία για την αξιολόγηση ενός μοντέλου ταξινόμησης όχι μόνο γιατί οδηγεί σε σαφή και αξιόπιστα αποτελέσματα για την απόδοση ενός μοντέλου ταξινόμησης ή την επιλογή του βέλτιστου μοντέλου κατά τη σύγκριση αυτών, αλλά και μπορεί να προσδιορίσει τα χαρακτηριστικά σημεία ενός ταξινομητή στα οποία η λειτουργία του και η απόδοσή του

μεγιστοποιείται. Στη συνέχεια του κεφαλαίου έχουμε τον τρόπο υπολογισμού της καμπύλης ROC και τα διαστήματα στα οποία η καμπύλη ROC βελτιστοποιείται.

### 2.3.1. ROC Curve

Η καμπύλη ROC [17],[18],[19],[20] (**Receiver Operating Characteristic**) είναι μία γραφική απεικόνιση του sensitivity vs. (1- specificity) για έναν δυαδικό classifier, καθώς το discrimination threshold του ποικίλει.

Sensitivity και specificity είναι στατιστικές μετρικές της απόδοσης ενός δυαδικού classifier. Το sensitivity [21] μετρά την αναλογία των δειγμάτων που ανήκουν στην πρώτη κλάση και σωστά αναγνωρίστηκαν ότι ανήκουν σε αυτή την κλάση και το specificity [21] μετρά την αναλογία των δειγμάτων που ανήκαν στη δεύτερη κλάση και αυτά επίσης αναγνωρίστηκαν να ανήκουν στη δεύτερη κλάση.

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{number\ of\ True\ Positive}}{\boldsymbol{number\ of\ True\ Positive + number\ of\ False\ Negative}}$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{number\ of\ True\ Negative}}{\boldsymbol{number\ of\ True\ Negative + number\ of\ False\ Positive}}$$

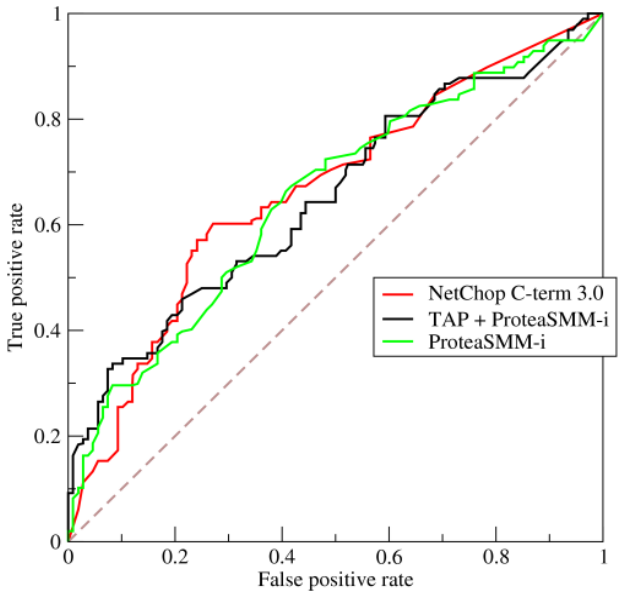
Το discrimination threshold είναι ένα σαφώς ορισμένο σημείο, το οποίο διχοτομεί ένα διάστημα - μία διάσταση, επιτυγχάνοντας το διαχωρισμό αντικειμένων μεταξύ αυτών, που απεικονίζονται στο ένα ή το άλλο μέρος του διαστήματος.

Η ROC μπορεί να σχεδιαστεί και από κλάσμα του TPR (True Positive Rate) vs. FPR (False Positive Rate). Το TPR καθορίζει κατά πόσον ο classifier ή το διαγνωστικό μας μοντέλο μπορεί να αναγνωρίσει τα αντικείμενα που ανήκουν στη θετική κλάση ανάμεσα σε όλα τα θετικά δείγματα κατά τη διάρκεια το testing. Από την άλλη μεριά του FPR μας δείχνει πόσα λάθος αναγνωρισμένα θετικά αποτελέσματα έχουμε ανάμεσα σε όλα τα αρνητικά δείγματα που προκύπτουν κατά τη διάρκεια του testing. Η ROC καμπύλη σχεδιάζεται σε ένα διάγραμμα με άξονες x(FPR) και y (TPR) στην ουσία απεικονίζει το σχετικές εναλλαγές μεταξύ του true positive

(πλεονέκτημα μεθόδου) και false positive (κόστος μεθόδου). Από τον ορισμό του TPR και FPR όπως θα δούμε και παρακάτω, κάθε πρόβλεψη που κάνουμε στο Confusion Matrix ουσιαστικά είναι ένα σημείο της καμπύλης ROC.

Η ανάλυση με χρήση της καμπύλης ROC οδηγεί με μεγάλη πιθανότητα στην επιλογή του βέλτιστου discrimination threshold και παράλληλα στην απόρριψη άλλων τυχόν υποβέλτιστων, ανεξάρτητα από τα κριτήρια διαχωρισμού ή την κατανομή που αυτά ακολουθούν. Η ROC είναι μία αξιόπιστη μέθοδος ανάλυσης στον κλάδο του diagnostic decision making.

Στην εικόνα 1 βλέπουμε μια ROC καμπύλη.




Εικόνα 1 : Μορφή Καμπύλης ROC

Κάθε σημείο της καμπύλης υπολογίζεται από τους παρακάτω τύπους:

ROC (x,y)=(TPR, FPR)

**TPR**: True Positive Rate, TPR = TP / P = TP / (TP + FN)

**FPR**: False Positive Rate, FPR = FP / N = FP / (FP + TN)

**ACC:** Accuracy, $ACC = (TP + TN) / (P + N)$

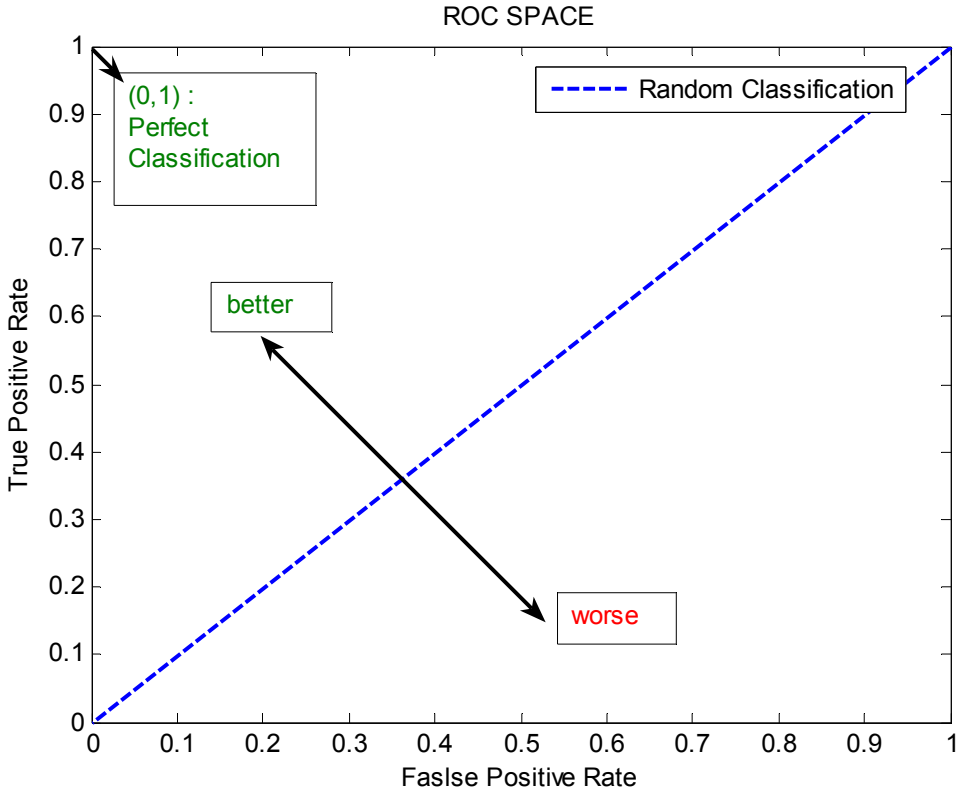
**SPC**: Specificity, SPC = TN / (FP + TN) = 1 – FPR

Όπως παρατηρούμε και από τους τύπους υπολογισμού της καμπύλης ROC , κάθε σημείο της είναι ένα Confusion Matrix.

### 2.3.2. ROC Space

Ας κατανοήσουμε τον χώρο της ROC για να δούμε τον τρόπο με τον οποίο αξιολογούμε το αποτέλεσμα μιας ROC καμπύλης. Η καλύτερη δυνατή πρόβλεψη που μπορεί να γίνει είναι σημείο (0,1) που αυτό πρακτικά σημαίνει 100% sensitivity, δηλαδή ότι όλα τα θετικά δείγματα βρέθηκαν και αναγνωρίστηκαν από το μοντέλο σαν θετικά και 100% specificity, δηλαδή ότι δεν βρέθηκε κανένα λάθος αναγνωρισμένο θετικό δείγμα.

Όπως βλέπουμε και στο παρακάτω σχήμα, για τα σημεία που πέφτουν πάνω στην ευθεία y=x έχει γίνει τυχαία ταξινόμηση. Όσο η καμπύλη ROC πλησιάζει την ευθεία αυτή, τόσο πιο τυχαία γίνεται η επιλογή κατάταξης των δειγμάτων. Για όλα τα σημεία που βρίσκονται κάτω από την ευθεία αυτή σημαίνει ότι η ταξινόμηση έφερε το ακριβώς αντίθετο αποτέλεσμα από το επιθυμητό. Η καλύτερη καμπύλη ROC πρέπει να πλησιάζει το σημείο (0,1) το οποίο θεωρητικά είναι και το σημείο στο οποίο η ταξινόμηση είχε 100% ακριβές αποτέλεσμα.




Εικόνα 2 : Χώρος καμπύλης ROC – ROC Space

## 2.4. Confidence Intervals της καμπύλης ROC

Με την ROC έχουμε κάποια πολύ κατατοπιστικά αποτελέσματα για την απόδοση των μοντέλων που εκπαιδεύσαμε. Χρειαζόμαστε όμως έναν ακόμα βαθμό βεβαιότητας της απόδοσης του μοντέλου που εκπαιδεύσαμε. Με άλλα λόγια χρειαζόμαστε τα διαστήματα εμπιστοσύνης (Confidence Intervals) , μία μετρική με την οποία να προσδιορίσαμε το εύρος που μπορεί να έχει μία καμπύλη ROC, δηλαδή την απόσταση μεταξύ της ROC που υπολογίζουμε και την καλύτερη αναμενόμενη ROC ή αντίστοιχα τη χειρότερη ROC που μπορούμε να πάρουμε. Η απάντηση δηλαδή στο ερώτημα ποιο είναι το βέλτιστο μοντέλο ταξινόμησης δεν προκύπτει μόνο με την καλύτερη ROC.

Τη λύση μας δίνει η τυπική απόκλιση ή Standard Deviation. Η τυπική απόκλιση είναι το μέτρο της απόκλισης μιας τυπικής μέτρησης από τη μέση τιμή ενός συνόλου επαναλαμβανόμενων μετρήσεων. Έχει αποδειχτεί ότι 68% των μετρήσεων βρίσκονται μέσα στο διάστημα μιας τυπικής απόκλισης από τη μέση τιμή. Επομένως υπολογίζοντας την τυπική απόκλιση της καμπύλης ROC μπορούμε να εξάγουμε

συμπεράσματα για τα όρια μέσα στα οποία μπορεί να κινηθεί η ROC. Με αυτόν τον τρόπο θα πάρουμε μια πιο σαφή εικόνα για την απόδοση του classifier.

Η τυπική απόκλιση υπολογίζεται ως εξής:

$$\sigma_{\chi} = \sqrt{\frac{\sum_{\iota=1}^{N}(\chi_{\iota} - \bar{\chi})^2}{N}}$$

## 2.5. Επεξεργασία Δεδομένων

Στο σημείο αυτό παρουσιάζουμε τις μεθόδους τις οποίες χρησιμοποιήσαμε για την επεξεργασία των δεδομένων μας. Η επεξεργασία των δεδομένων περιλαμβάνει τον διαχωρισμό του dataset σε training set και testing set. Το γεγονός ότι τα σετ δεδομένων εκπαίδευσης , τα δείγματα του κάθε dataset είναι σχετικά λίγα σε πλήθος καθώς και το γεγονός ότι η συντριπτική πλειοψηφία των δειγμάτων ανήκουν στη μία από τις δύο κλάσεις, μας οδήγησαν σε τεχνικές επεξεργασίας και διαχωρισμού δεδομένων όπως η Cross Validation που περιγράφεται παρακάτω.

### 2.5.1. Cross Validation

Κατά τη διαδικασία του classification και μετά το training θέλουμε να διασφαλίσουμε ότι τα αποτελέσματα που πήραμε και η εκπαίδευση που ήδη κάναμε στον classifier μπορούν να χρησιμοποιηθούν με την ίδια επιτυχία σε κάποιο ανεξάρτητο δείγμα. Για το λόγο αυτό χρησιμοποιούμε την τεχνική Cross Validation[24][25].

Η Cross Validation είναι μια τεχνική αξιολόγησης του κατά πόσον τα αποτελέσματα που πήραμε από τη στατιστική ανάλυση μπορούν να γενικευθούν και να χρησιμοποιηθούν σε ένα ανεξάρτητο dataset. Με άλλα λόγια χρησιμοποιείται για να καθοριστεί η ακρίβεια σε ένα μοντέλο προβλέψεων που χρησιμοποιείται στην πράξη.

Η διαδικασία περιλαμβάνει την διάσπαση των δειγμάτων σε συμπληρωματικά υποσύνολα, η ανάλυση - εκπαίδευση γίνεται σε ένα από αυτά τα υποσύνολα , το οποίο ονομάζεται training set, και η αξιολόγηση - έλεγχος ορθότητας των αποτελεσμάτων γίνεται με ένα υποσύνολο το οποίο ονομάζεται evaluation set ή

test set. Για ακόμα μεγαλύτερη αξιοπιστία αποτελεσμάτων και για να περιορίσουμε το εύρος της εκτίμησης των αποτελεσμάτων. Η διαδικασία του Cross Validation μπορεί να χωριστεί σε επιμέρους στάδια χρησιμοποιώντας σε κάθε ένα από αυτά διαφορετικό υποσύνολο του dataset για test set και διαφορετικό για training set. Το τελικό αποτέλεσμα της αξιολόγησης προκύπτει από το μέσο όρο των φάσεων που πραγματοποιήθηκαν κατά το Cross Validation.

Ας υποθέσουμε τώρα ότι έχουμε ένα μοντέλο με μία ή περισσότερες παραμέτρους που πρέπει να προσδιοριστούν ώστε ένα dataset από το οποίο προήλθε το μοντέλο. Η διαδικασία κάνει το μοντέλο να ταιριάζει και να έχει όσο το δυνατόν καλύτερα αποτελέσματα για το συγκεκριμένο training set. Αν τώρα κρατήσουμε ένα ανεξάρτητο δείγμα του dataset από την εκπαίδευση, θα δούμε ότι το μοντέλο δεν συμπεριφέρεται το ίδιο καλά στο ανεξάρτητο δείγμα όπως και στο training set. Αυτό το φαινόμενο ονομάζεται overfitting και συμβαίνει κυρίως όταν το μέγεθος του training data είναι μικρό και οι παράμετροι του dataset είναι πολλές σε πλήθος. Η Cross Validation είναι μία μέθοδος να αποφύγουμε το φαινόμενο αυτό.

Το Linear Regression μας δείχνει με ποιον τρόπο το overfitting συμβαίνει. Έχουμε τιμές $Y_1, \ldots, Y_n$, και διανύσματα συνδιακύμανσης $X_1, \ldots, X_p$. Με τη χρήση των ελαχίστων τετραγώνων μπορούμε να θέσουμε ένα όριο – κατώφλι $a + b_1X_1 + \ldots + b_pX_p$ μεταξύ X και Y και μετά να υπολογίσουμε το fit από το mean squared error (MSE).

$$\sum_i (Y_i - a - b_1X_{1i} - \cdots - b_p X_{pi})^2/n,$$

Όπου $X_{ji}$ είναι τιμές της μεταβλητής $X_j$ της $i^{th}$ τιμής του $Y_i$.

Μπορούμε να δείξουμε ότι η αναμενόμενη τιμή του MSE για το training set είναι $(n - p - 1)/n < 1$ φορές η αναμενόμενη τιμή MSE για το validation set. Για το λόγο αυτό υπολογίζουμε το MSE για το training set και μετά παίρνουμε πρόβλεψη του πώς το μοντέλο θα συμπεριφερθεί στο independent set.

### 2.5.2. K-Folds Cross Validation

Στην μέθοδο K-FOLD Cross Validation, το αρχικό dataset διασπάται σε K- υποσύνολα τα οποία ονομάζονται folds. Ένα από αυτά τα K-υποσύνολα εξαιρείται από τη

διαδικασία του training και χρησιμοποιείται σαν το ανεξάρτητο δείγμα αξιολόγησης του μοντέλου. Τα υπόλοιπα K-1 υποσύνολα χρησιμοποιούνται στο training.

Η διαδικασία του Cross Validation επαναλαμβάνεται K φορές, όσες δηλαδή και τα folds, με το κάθε ένα από αυτά τα K- υποσύνολα να χρησιμοποιείται μία ακριβώς φορά σαν validation data. Από τα K αποτελέσματα που προκύπτουν από αυτή την επανάληψη της διαδικασίας μπορούμε να πάρουμε τον μέσο όρο σαν οριστικό αποτέλεσμα της μεθόδου.

Το πλεονέκτημα της μεθόδου K-FOLD Cross Validation σε σχέση με την κλασική μέθοδο Cross Validation, που διαλέγει ένα τυχαίο κομμάτι από το dataset για validation data, είναι ότι κάθε δείγμα χρησιμοποιείται και για την εκπαίδευση αλλά και για την αξιολόγηση του μοντέλου ακριβώς μία φορά. Την προτιμούμε από την απλή μέθοδο Cross Validation λόγω αξιοπιστίας και από άλλες παραλλαγές της μεθόδου Cross Validation όπως η leave-one-out γιατί τα dataset που χρησιμοποιούμε έχουν σχετικά μικρό αριθμό δειγμάτων και δεν έχουμε την πολυτέλεια να αποκλείουμε δείγματα από τη διαδικασία.

### 2.5.3. Stratified K-Folds Cross Validation

Σε περιπτώσεις που τα δείγματα σε ένα dataset ανήκουν κατά συντριπτική πλειοψηφία σε μία από τις δύο κλάσεις δημιουργείται το πρόβλημα ότι διασπώντας με τυχαίο τρόπο το dataset για να δημιουργήσουμε τα K-Folds μπορεί να συμβεί να μην συμπεριλάβουμε σε κάποιο από τα folds κανένα δείγμα από την κλάση η οποία έχει τα μειοψηφικά δείγματα. Για αυτό ακριβώς το λόγο θα πρέπει να εξασφαλίσουμε ότι κάθε υποσύνολο – fold του dataset το οποίο δημιουργούμε θα περιέχει αναλογικά το ίδιο ποσοστό δειγμάτων που ανήκουν στη μία και στην άλλη κλάση. Η διαδικασία αυτή περιγράφεται σαν stratified K-FOLD Cross Validation.

Κατά τη μέθοδο Stratification, ομαδοποιούμε τα δείγματα που ανήκουν σε κάθε κλάση πριν κάνουμε δειγματοληψία και έπειτα κάνουμε τη δειγματοληψία σε κάθε μέρος (stratum) ξεχωριστά. Θα πρέπει να εξασφαλίσουμε ότι κάθε δείγμα ανήκει σε ένα μόνο stratum και ότι κανένα από τα δείγματα δεν θα αποκλειστεί από τη δειγματοληψία. Η τυχαία δειγματοληψία τώρα θα γίνει ανάμεσα στα δείγματα κάθε stratum. Αυτό μας βοηθά να περιορίσουμε τα λάθη στη δειγματοληψία.

## 2.6. Datasets – Τα δεδομένα που χρησιμοποιήθηκαν

Στο κεφάλαιο αυτό αναλύουμε την προέλευση και την μορφή των αρχικών δεδομένων του πειράματός μας. Βλέπουμε όλα τα datasets και τα χαρακτηριστικά τους.

### 2.6.1. Ovarian Tumor Dataset

Το συγκεκριμένο dataset [22] και που εξετάζουμε προέρχεται από μία μελέτη που έγινε το διεθνές κέντρο ανάλυσης του καρκίνου των ωοθηκών. Το dataset αποτελείται από 1066 ασθενείς οι οποίοι έχουν τουλάχιστον όγκο στην περιοχή των ωοθηκών. Από αυτούς τους ασθενείς οι 800 έχουν καλοήθη όγκο, οι 169 έχουν κακοήθη όγκο που όμως επιδέχεται επέμβαση, οι 55 έχουν κακοήθη όγκο μη εγχειρήσιμο και οι 42 έχουν μεταστατικό κακοήθη όγκο που επίσης δεν είναι εγχειρήσιμος. Αυτό που μας ενδιαφέρει κατά το classification σε αυτή την περίπτωση είναι να εκπαιδευτεί ο classifier ώστε να δίνει με ασφάλεια αποτελέσματα στους γιατρούς για το αν ο όγκος που παρουσιάζεται στον ασθενή είναι εγχειρήσιμος ή όχι. Επομένως το dataset μας τελικά έχει 2 κλάσεις οι οποίες έχουν 969 και 97 δείγματα αντίστοιχα. Η έρευνα έγινε με δείγματα ασθενών από 9 Ευρωπαϊκά κέντρα ογκολογίας, τα οποία είναι: Μάλμο Σουηδίας, Λέβεν Βέλγιο, Ρώμη, Μιλάνο, Παρίσι στο κέντρο Baucicout, Παρίσι στο κέντρο Maurepas, Λονδίνο, Μόνζα Ιταλίας, Νάπολη. Τα δείγματα ασθενών και οι τύποι όγκων είναι κατανεμημένα όπως δείχνει ο παρακάτω πίνακας.

**Table 1.** Histologic Outcome of Adnexal Tumors by Participating Center

| | Total | | Benign | | Malignant | | Primary Invasive | | Borderline | | Metastatic | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Center | No. | %* | No. | %† | No. | %† | No. | %‡ | No. | %‡ | No. | %‡ |
| Malmö | 315 | 29.5 | 247 | 78.4 | 68 | 21.6 | 40 | 58.8 | 17 | 25.0 | 11 | 16.2 |
| Leuven | 263 | 24.7 | 170 | 64.6 | 93 | 35.4 | 62 | 66.7 | 14 | 15.1 | 17 | 18.3 |
| Rome | 126 | 11.8 | 81 | 64.3 | 45 | 35.7 | 23 | 51.1 | 12 | 26.7 | 10 | 22.2 |
| Milano | 87 | 8.2 | 79 | 90.8 | 8 | 9.2 | 6 | 75.0 | 1 | 12.5 | 1 | 12.5 |
| Paris§ | 80 | 7.5 | 71 | 88.8 | 9 | 11.2 | 7 | 77.8 | 2 | 22.2 | 0 | 0 |
| Paris‖ | 64 | 6.0 | 57 | 89.1 | 7 | 10.9 | 6 | 85.7 | 1 | 14.3 | 0 | 0 |
| London | 54 | 5.1 | 38 | 70.4 | 16 | 29.6 | 10 | 62.5 | 4 | 25.0 | 2 | 12.5 |
| Monza | 46 | 4.3 | 29 | 63.0 | 17 | 37.0 | 12 | 70.6 | 4 | 23.5 | 1 | 5.9 |
| Naples | 31 | 2.9 | 28 | 90.3 | 3 | 9.7 | 3 | 100.0 | 0 | 0 | 0 | 0 |
| All | 1,066 | 100 | 800 | 75.0 | 266 | 25.0 | 169 | 63.5 | 55 | 20.7 | 42 | 15.8 |

*Of study total.
†Of center total.
‡Of malignant tumors.
§Paris, Baucicout.
‖Paris, Maurepas.

Εικόνα 3 : Δείγματα ασθενών Καρκίνου Ωοθηκών ανά Ογκολογικό Κέντρο

Τα χαρακτηριστικά που έχει το dataset για κάθε δείγμα είναι τα εξής:

- Προσωπικό ιστορικό καρκίνου ωοθηκών

  (ναι/ όχι)
- Ορμονική θεραπεία

  (ναι/ όχι)
- Ηλικία

  (σε έτη)
- Μέγιστη διάμετρος κάκωσης

  (Σε millimeters)
- Πόνος

  (ναι/ όχι)
- Ασκίτης, Συγκέντρωση κίτρινων υγρών στην περιτονιακή κοιλότητα (Ascites = abnormal accumulation of clear yellow fluid in the peritoneal cavity)

  (ναι/ όχι)
- Ροή αίματος στη στέρεα θηλώδη προβολή (Blood Flow within a solid papillary projection)

  (ναι/ όχι)
- Παρουσία τμήματος του όγκου που είναι τελείως συμπαγές (στέρεο)

  (ναι/ όχι)
- Μέγιστη διάμετρος τμήματος όγκου που είναι συμπαγής

  (Σε millimeters)
- Προβληματικά εσωτερικά τοιχώματα στην κύστη

  (ναι/ όχι)
- Acoustic Shadows (Περιοχές που ο ήχος δε μπορεί να μεταδοθεί λόγω εμποδίων)

  (ναι/ όχι)
- Χρωματική απεικόνιση της ροής του αίματος στην περιοχή εντός του όγκου (Color score of intratumoral blood flow)

  (1,2,3,4)

### 2.6.2. Brain Tumor Dataset

Το Brain Tumor dataset [23], προέρχεται από δεδομένα ασθενών που διέθεσαν οι C. L. Blake, C. J. Merz. Τα δείγματα ασθενών προέρχονται από ογκολογικό κέντρο της

Γερμανίας [23] και αφορούν σε ασθενείς με όγκους στον εγκέφαλο . Το dataset περιέχει 1000 δείγματα ασθενών και δύο κλάσεις την Α και Β. Οι ασθενείς που ανήκουν στην κλάση Α έχουν καλοήθη όγκο και είναι συνολικά 700, ενώ οι υπόλοιποι 300 ασθενείς ανήκουν στην δεύτερη κλάση και εμφανίζουν κακοήθη, υποτροπιάζον όγκο που δεν επιδέχεται εγχείρηση. Το dataset έχει 20 χαρακτηριστικά τα οποία είναι φύλο, ηλικία, προγενέστερη θεραπεία , ακτινοβολίες, μέγεθος όγκου κλπ, αλλά περιέχει κυρίως χαρακτηριστικά που προσδιορίζουν την περιοχή του εγκεφάλου στην οποία βρίσκεται ο όγκος.

### 2.6.3. Breast Tumor Dataset

Το dataset αυτό προέρχεται από το University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia με δεδομένα που διέθεσαν οι M.Zwitter και M.Soklic. [26] Το dataset αφορά όγκους καρκίνου του μαστού. Αποτελείται από δύο κλάσεις Α,Β. Στην πρώτη κλάση ανήκουν τα δείγματα όγκων στα οποία είναι σε αρχικό στάδιο και μπορεί να είναι εγχειρήσιμα ενώ στη δεύτερη κλάση ανήκουν οι υποτροπιάζοντες όγκοι που δεν είναι εγχειρήσιμοι. Χρησιμοποιήθηκαν συνολικά 277 δείγματα ασθενών από τα οποία 196 ανήκουν στην πρώτη κλάση και 81 στη δεύτερη.

Τα χαρακτηριστικά που έχει το dataset για κάθε δείγμα είναι τα εξής:

- Τύπος όγκου
  (No-recurrence-events, recurrence-events)
- Ηλικία
  (10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99)
- Εμμηνόπαυση
  (Πριν τα 40, Μετά τα 40, Όχι)
- Μέγεθος όγκου
  (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59)
- Αδένες:
  (0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39)
- Node-caps:

(Ναι/Όχι)

- Περιοχή κακοήθειας

  (1,2,3)

- Μαστός

  (Δεξιός , Αριστερός)

- Τεταρτημόριο μαστού

  (Πάνω αριστερό, Κάτω αριστερό, Πάνω δεξιό, Κάτω Δεξιό)

- Θεραπεία με Ακτινοβολία

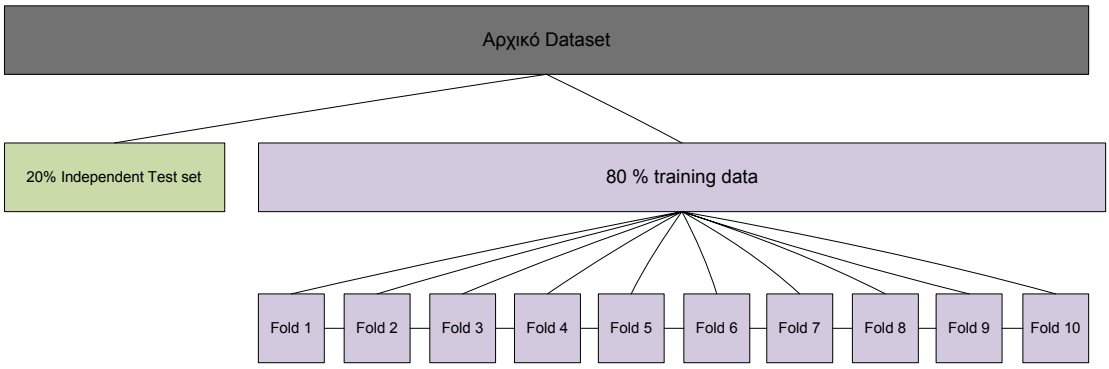
  (Ναι/ Όχι)

# 3. Υλοποίηση

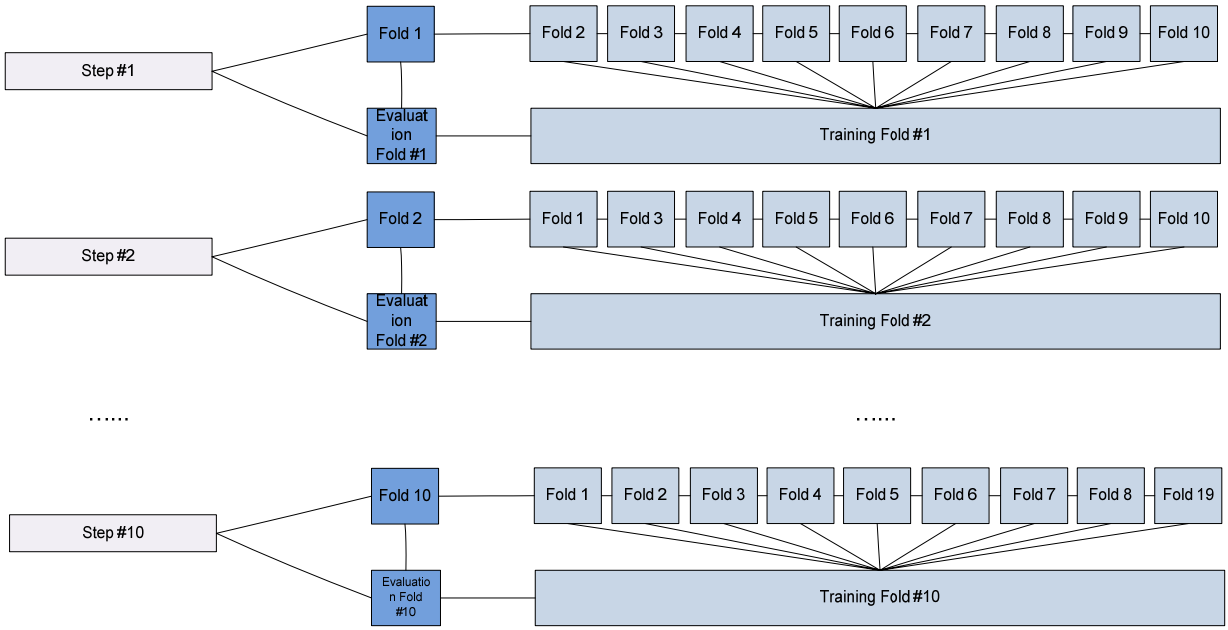
Στο κεφάλαιο αυτό θα δούμε τη σύνδεση των θεωρητικών δεδομένων και μεθόδων που περιγράψαμε στο προηγούμενο κεφάλαιο με την υλοποίηση που έγινε στα πλαίσια της διπλωματικής εργασίας και την προσαρμογή των δεδομένων στο δικό μας πρόβλημα.

## 3.1. Datasets, Training Sets, Testing Sets

Για την επεξεργασία των dataset χρησιμοποιήσαμε Stratified K-Folds Cross Validation. Πιο συγκεκριμένα επιλέξαμε Stratified 10-Folds Cross Validation. Η διαδικασία ήταν κοινή και για τα τρία datasets που χρησιμοποιήσαμε. Από το αρχικό dataset αφαιρούμε ρο 20% των δειγμάτων, το οποίο το χρησιμοποιούμε σαν ανεξάρτητο test set. Αυτό το πρώτο μέρος της διαδικασίας γίνεται έχοντας εξασφαλίσει ότι στο testing set έχουμε κάνει stratified sampling και άρα το ποσοστό δειγμάτων που ανήκουν στην μία ή στην άλλη κλάση είναι το ίδιο με το αντίστοιχο ποσοστό που είχε το αρχικό δείγμα. Στο υπόλοιπο 80% του αρχικού dataset εφαρμόζουμε Stratified K-Folds Cross Validation και το χωρίζουμε σε 10 υποσύνολα. Με αυτά τα δέκα υποσύνολα, που αντιπροσωπεύουν και τις 10 φορές που θα εκτελεστεί το πείραμά μας, δημιουργούμε για κάθε ένα από τα δέκα βήματα ένα training set (fold) και ένα evaluation set. Στο πρώτο βήμα το πρώτο υποσύνολο του dataset (το πρώτο από τα υποσύνολα που δημιουργήθηκαν με το 80% του αρχικού dataset) το χρησιμοποιούμε σαν evaluation set (evaluation set #1) ενώ τα υπόλοιπα εννέα υποσύνολα τα χρησιμοποιούμε σαν training set (fold #1). Με ανάλογο τρόπο στο δεύτερο βήμα το δεύτερο υποσύνολο του dataset το χρησιμοποιούμε σαν evaluation set (evaluation set #2) ενώ τα υπόλοιπα εννέα υποσύνολα τα χρησιμοποιούμε σαν training set (fold #2), η ίδια ακριβώς διαδικασία επαναλαμβάνεται 10 φορές μέχρι το δέκατο βήμα όπου εκεί το δέκατο υποσύνολο του dataset το χρησιμοποιούμε σαν evaluation set (evaluation set #10) ενώ τα υπόλοιπα εννέα υποσύνολα τα χρησιμοποιούμε σαν training set (fold #10). Η Εικόνα 4 δίνει μια πιο σαφή εικόνα για το πώς διαμορφώθηκαν τα δεδομένα μας μετά τη μέθοδο 10-Folds Cross Validation.




Εικόνα 4: Τρόπος επεξεργασίας των datasets με τη μέθοδο K-Folds Cross Validation




Εικόνα 5 : Δημιουργία Training Sets , Evaluation Sets, Independent Test Set για τα 10 – Folds από τα Datasets

Στην Εικόνα 5 βλέπουμε με ποιον τρόπο από τα τρία datasets δημιουργήθηκαν τα δεδομένα για τα δέκα βήματα διαδικασίας, ας δούμε αναλυτικά το μέγεθος και τη μορφή των δειγμάτων αυτών.

### 3.1.1. Μέγεθος και μορφή δεδομένων Testing και Training για Ovarian Tumor Dataset

Στο Ovarian Tumor Dataset [22] έχουμε 1066 δείγματα συνολικά με 969 στην κλάση #1 και 97 στην κλάση #2. Στον Πίνακα 2 φαίνονται αναλυτικά, ο αριθμός των

δειγμάτων ανά κλάση στο αρχικό dataset καθώς και το πώς διαμορφώθηκαν όλα αυτά μετά το 10-Fold Cross Validation.

| Ovarian Tumor | Total | Class 1 | Class 2 |
|---|---|---|---|
| Αχρικό Dataset | 1066 | 969 | 97 |
| Testing Set 80% | 214 | 194 | 20 |
| Folds | | | |
| Evaluation Set | 86 | 78 | 8 |
| Training Set | 766 | 697 | 69 |

Πίνακας 2: Πλήθος Δεδομένων ανά test και train set , Δείγματα ανά κλάση στο Ovarian Tumor Dataset

### 3.1.2. Μέγεθος και μορφή δεδομένων Testing και Training για Brain Tumor Dataset

Στο Brain Tumor Dataset [] έχουμε 1000 δείγματα συνολικά με 700 στην κλάση #1 και 300 στην κλάση #2. Στον Πίνακα 3 φαίνονται αναλυτικά, ο αριθμός των δειγμάτων ανά κλάση στο αρχικό dataset καθώς και το πώς διαμορφώθηκαν όλα αυτά μετά το 10-Fold Cross Validation.

| Brain Tumor | Total | Class 1 | Class 2 |
|---|---|---|---|
| Αχρικό Dataset | 1000 | 700 | 300 |
| Testing Set 80% | 200 | 140 | 60 |
| Folds | | | |
| Evaluation Set | 80 | 56 | 24 |
| Training Set | 720 | 504 | 216 |

Πίνακας 3: Πλήθος Δεδομένων ανά test και train set , Δείγματα ανά κλάση στο Brain Tumor Dataset

### 3.1.3. Μέγεθος και μορφή δεδομένων Testing και Training για Brain Tumor Dataset

Στο Breast Tumor dataset [26] έχουμε 277 δείγματα συνολικά με 196 στην κλάση #1 και 81 στην κλάση #2. Στον Πίνακα 4 φαίνονται αναλυτικά, ο αριθμός των δειγμάτων ανά κλάση στο αρχικό dataset καθώς και το πώς διαμορφώθηκαν όλα αυτά μετά το 10-Fold Cross Validation.

| Dsiota | Total | Class 1 | Class 2 |
|---|---|---|---|
| Αχρικό Dataset | 277 | 196 | 81 |
| Testing Set 80% | 57 | 40 | 17 |
| Folds | | | |
| Evaluation Set | 21 | 15 | 6 |
| Training Set | 189 | 135 | 54 |

Πίνακας 4: Πλήθος Δεδομένων ανά test και train set , Δείγματα ανά κλάση στο Breast Tumor Dataset

Παρατηρούμε ότι το dataset το οποίο είναι πιο δύσκολο να επεξεργαστούμε και να βγάλουμε συμπεράσματα για την απόδοση των μοντέλων εκπαίδευσης είναι το Breast Cancer Tumor και αυτό γιατί περιέχει τα λιγότερα δείγματα από όλα τα datasets και κατά συνέπεια τα δείγματα εκπαίδευσης και επαλήθευσης είναι πολύ μικρότερα σε αυτό από ότι στα άλλα datasets. Επίσης παρατηρούμε το γεγονός ότι στο Ovarian Tumor dataset το ποσοστό δειγμάτων που ανήκουν στην κλάση #1 είναι 90,9% , διατηρεί δηλαδή η κλάση #1 την συντριπτική πλειοψηφία δειγμάτων, σε σχέση με το αντίστοιχο ποσοστό δειγμάτων της κλάσης #1 του Brain Tumor Dataset και Breast Tumor Dataset που είναι 70% και 61% αντίστοιχα.

## 3.2. Classification

Στη διαδικασία του classification και σε κάθε ένα από τα δέκα βήματα της διαδικασίας ακολουθούμε την ίδια ακριβώς ρουτίνα με την οποία εξάγουμε τα mappings, ελέγχουμε το classifier performance, εξάγουμε το συνολικό Error Rate, βλέπουμε πόσα δείγματα ήταν hit και πόσα miss ανά κλάση. Αυτή είναι η πρώτη μετρική για να διαπιστώσουμε πόσο καλά εκπαιδεύτηκε το μοντέλο μας.

Κατά την εκπαίδευση λοιπόν, τρέχουμε τη ρουτίνα για κάθε fold ώστε να δημιουργήσουμε δέκα mappings. Η διαδικασία αυτή επαναλαμβάνεται για τους 17 classifiers που θέλουμε να αξιολογήσουμε και για τα τρία datasets που χρησιμοποιήθηκαν.

Αφού δημιουργηθούν τα δέκα mappings ακολουθεί το testing στο οποίο χρησιμοποιούμε σαν είσοδο το mapping κάθε fold και το αντίστοιχο evaluation test set κάθε fold, ενώ στην έξοδο παίρνουμε το Error Rate του classifier καθώς και έναν 1X2 πίνακα ο οποίος έχει τις λανθασμένες αποφάσεις ανά κλάση.

Από τις λανθασμένες αποφάσεις ανά κλάση σε συνδυασμό με τα πραγματικά αποτελέσματα μπορούμε να υπολογίσουμε σε πρώτη φάση το Confusion Matrix ως εξής: Ας υποθέσουμε ότι το testing dataset αποτελείται από M δείγματα, $M_1$ από τη πρώτη κλάση που ορίζουμε σας positive και $M_2$ από την δεύτερη κλάση που ορίζουμε σαν negative. Επίσης από την έξοδο του testing θα πάρουμε $F_1$ λανθασμένα δείγματα στην πρώτη κλάση και $F_2$ λανθασμένα δείγματα στη δεύτερη. Έτσι το Confusion Matrix υπολογίζεται ως εξής:

True Positive (TP) = $M_1$-$F_1$

False Positive (FP) = $F_2$

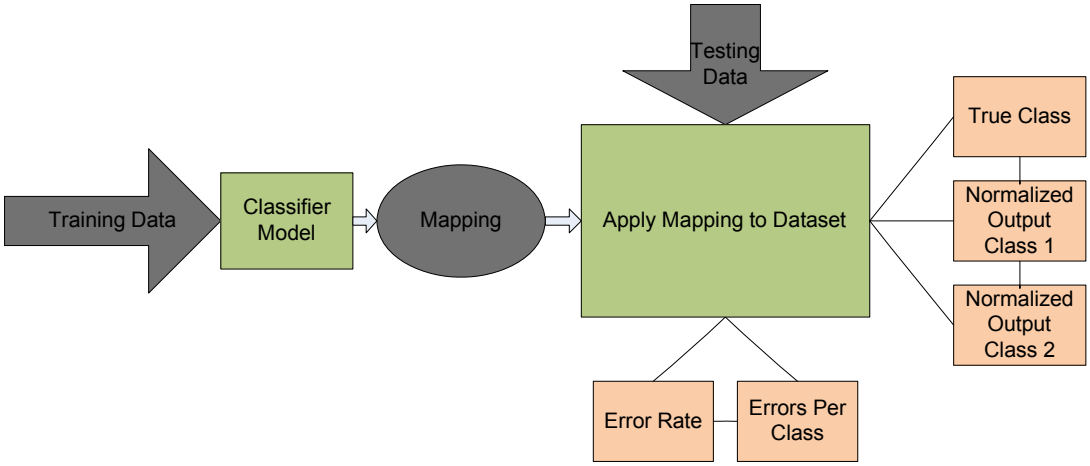
True Negative (TN) = $M_2$-$F_2$

False Negative (FN) = $F_1$

Για να μπορέσουμε βέβαια να συγκρίνουμε το μοντέλο μας σε κάθε dataset και επειδή αυτά δεν έχουν τον ίδιο αριθμό δειγμάτων μετατρέπουμε τους καθαρούς αριθμούς σε ποσοστό.

Αυτό το mapping που δημιουργήθηκε το εφαρμόζουμε στο evaluation test set και στη ουσία σε αυτό το σημείο εφαρμόζουμε το μοντέλο που εκπαιδεύσαμε προηγουμένως στο test set μας ώστε να διαπιστώσουμε αν η εκπαίδευση ήταν επιτυχής.

Στη συνέχεια, μετασχηματίζουμε το mapping σε ώστε να πάρουμε στην έξοδο τελικά το επιθυμητό, την τελική απάντηση του πόσο καλά λειτούργησε ο classifier μας στο evaluation set. Με άλλα λόγια η έξοδος είναι ένας πίνακας που κάθε γραμμή του απεικονίζει και ένα δείγμα του evaluation test set. Η πρώτη στήλη του μας προσφέρει πληροφορία για την κλάση στην οποία ανήκε το δείγμα στην πραγματικότητα ενώ οι επόμενες δύο στήλες μας δίνουν σε κανονικοποιημένη μορφή την πιθανότητα να κατατάξει το μοντέλο το δείγμα μας στην μία κλάση και στην άλλη. Σε κάθε δείγμα βεβαία το άθροισμα των πιθανοτήτων αυτών ισούται με ένα.




Εικόνα 6: Διαδικασία και αποτελέσματα Classification για κάθε FOLD

Στην εικόνα 6 βλέπουμε τη διαδικασία που ακολουθήθηκε κατά το Classificationσε κάθε ένα από τα 10 Folds . Με γκρι απεικονίζονται οι είσοδοι, με πράσινο οι διαδικασίες, ενώ με τα πορτοκαλί οι έξοδοι. Η διαδικασία που φαίνεται στην εικόνα επαναλαμβάνεται δέκα φορές για κάθε fold κατά τη διαδικασία της αξιολόγησης με το evaluation set για κάθε classifier, έπειτα ακολουθεί ο έλεγχος της απόδοσης του μοντέλου με το ανεξάρτητο δείγμα που κρατήσαμε εκτός της εκπαίδευσης για κάθε dataset και επαναλαμβάνουνε τη διαδικασία άλλες δέκα φορές. Τελικά τα αποτελέσματα που παίρνουμε σε αυτή τη φάση για το performance είναι:

- Αξιολογώντας με Evaluation set
  - Error Rate για κάθε fold
  - Λανθασμένες αποφάσεις ανά κλάση για κάθε fold
  - Ποσοστιαίο Confusion Matrix για κάθε fold (TP,TN,FP,FN)
  - Μέσος όρος Error Rate για όλα τα folds
  - Μέσος όρος λανθασμένων αποφάσεων ανά κλάση
  - Μέσο ποσοστιαίο Confusion Matrix
- Αξιολογώντας με το ανεξάρτητο δείγμα
  - Error Rate για κάθε fold
  - Λανθασμένες αποφάσεις ανά κλάση για κάθε fold
  - Ποσοστιαίο Confusion Matrix για κάθε fold (TP,TN,FP,FN)
  - Μέσος όρος Error Rate για όλα τα folds
  - Μέσος όρος λανθασμένων αποφάσεων ανά κλάση
  - Μέσο ποσοστιαίο Confusion Matrix

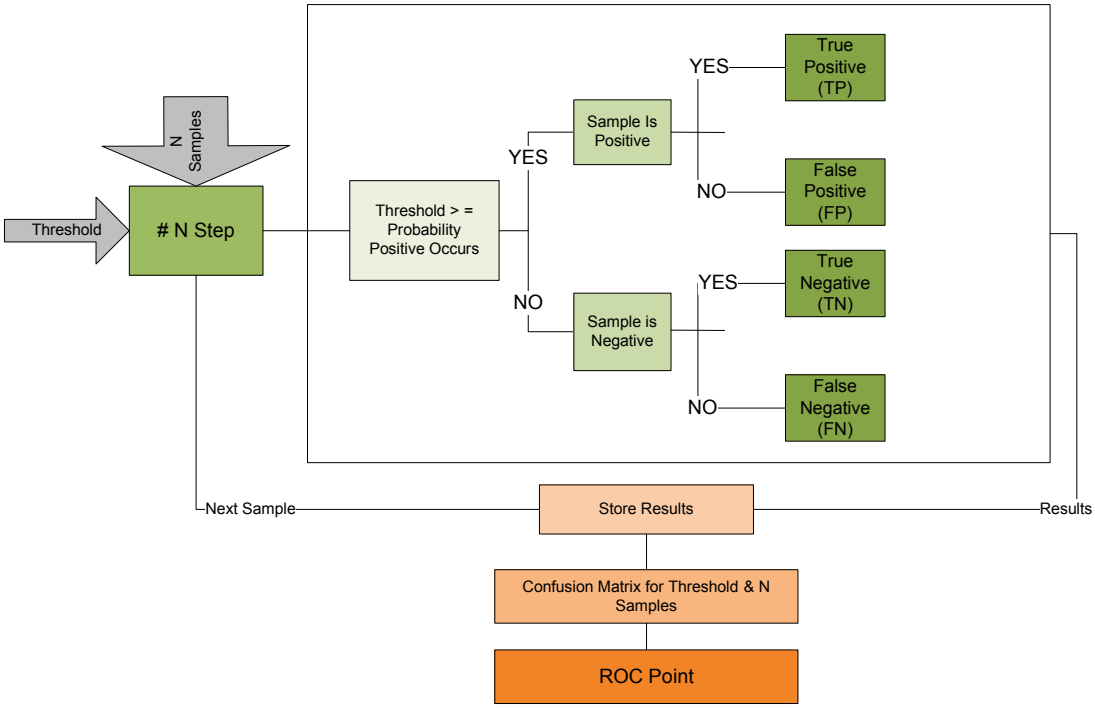
## 3.3. Σχεδίαση ROC

Όπως είδαμε και στο γενικό ορισμό της καμπύλης ROC, έτσι και για τους δυαδικούς classifiers που εξετάζουμε η ROC μπορεί να σχεδιαστεί και από κλάσμα του TPR (True Positive Rate) vs. FPR (False Positive Rate), διάγραμμα με άξονες x(FPR) και y (TPR). Κάθε σημείο της καμπύλης ROC προκύπτει από το Confusion Matrix για το σημείο αυτό. Πιο συγκεκριμένα, για κάθε δείγμα του testing μας ο classifier παράγει μία πιθανότητα - έναν βαθμό βεβαιότητας - που αντιπροσωπεύει το αν το δείγμα ανήκει στη μία κλάση ή στην άλλη, όπως είχαμε δει και στη διαδικασία classification. Με αυτές τις τιμές θέτουμε ένα κατώφλι σύμφωνα με το οποίο αποφασίζουμε σε ποια κλάση θα κατατάξουμε το δείγμα. Αν δηλαδή οι πιθανότητα αυτή είναι ίση ή μεγαλύτερη από το κατώφλι μας τότε αποφασίζουμε ότι το δείγμα ανήκει στη θετική κλάση, αν η πιθανότητα είναι μικρότερη από το κατώφλι τότε αποφασίζουμε αρνητική κλάση. Με αυτό τον τρόπο δημιουργούμε ένα Confusion Matrix για κάθε δείγμα. Σχεδιάζοντας το ROC σημείο για κάθε δυνατό κατώφλι έχουμε τελικά την καμπύλη ROC.

Για να επιλέξουμε τα κατώφλια αυτά έγιναν πολλά πειράματα σε όλο το διάστημα [0,1] που μπορεί να ανήκει ένα κατώφλι. Αρχικά πήραμε όλα τα σημεία από 0 έως ένα 1 με βήμα 0,1. Στις περιοχές της καμπύλης όμως που παρουσιάζεται μεγαλύτερο ενδιαφέρον, εκεί δηλαδή που φαίνεται να ελαχιστοποιείται η απόσταση μεταξύ της ROC και του σημείου (0,1) που είναι το ιδανικό classification, το βήμα επιλογής του operating point ήταν 0,001.

Επομένως για κάθε κατώφλι και για ένα-ένα τα δείγματα που έχουμε για testing ελέγχουμε αν η πιθανότητα να είναι το συγκεκριμένο δείγμα θετικό είναι μεγαλύτερη ή ίση από το κατώφλι. Αν είναι μεγαλύτερη αποφασίζουμε το δείγμα σαν positive. Έπειτα ελέγχουμε την πραγματική κλάση στην οποία ανήκει το δείγμα. Αν η κλάση στην οποία ανήκει το δείγμα είναι όντως positive τότε έχουμε ένα True Positive ενώ αν η κλάση που ανήκει το δείγμα δεν είναι positive έχουμε ένα False Positive. Αν τώρα η πιθανότητα να είναι το συγκεκριμένο δείγμα θετικό είναι μικρότερη από το κατώφλι μας τότε το δείγμα θεωρείται negative. Έπειτα ελέγχουμε την πραγματική κλάση στην οποία ανήκει το δείγμα. Αν η κλάση στην

οποία ανήκει το δείγμα είναι όντως negative τότε έχουμε ένα True Negative ενώ αν η κλάση που ανήκει το δείγμα δεν είναι positive έχουμε ένα False Negative. Αφού περάσουν όλα τα δείγματα από τον έλεγχο αυτό δημιουργείται ένα Confusion Matrix μεγέθους όσο και τα δείγματα του testing set. Από το Confusion Matrix αυτό υπολογίζουμε το TPR και FPR και αυτό μας δίνει ένα σημείο πάνω στη ROC. Επαναλαμβάνοντας την ίδια διαδικασία για όλα τα κατώφλια έχουμε ολόκληρη την ROC καμπύλη, όπως βλέπουμε και στην εικόνα 7.




Εικόνα 7: Block Diagram Υπολογισμού των σημείων της καμπύλης ROC

Επειδή η διαδικασία έχει γίνει με τη μέθοδο των 10 Folds αρχικά θα υπολογίσουμε με τον τρόπο που δείξαμε παραπάνω μία καμπύλη ROC για κάθε fold. Για να υπολογίσουμε την μέση ROC που θα αντιπροσωπεύει συνολικά το μοντέλο μας, κάθε classifier, θα πάρουμε έναν μέσο όρο των ROC υπολογίσαμε σε κάθε fold. Αυτό θα γίνει ως εξής: Κάθε σημείο της ROC αντιπροσωπεύει τις τιμές του Confusion Matrix για το συγκεκριμένο κατώφλι στο συγκεκριμένο fold. Άρα έχουμε δέκα συνολικά στοιχεία για κάθε κατώφλι. Από αυτά τα δέκα σημεία υπολογίζουμε το μέσο TPR και το μέσο FPR και από αυτά προκύπτει το νέο σημείο της μέσης καμπύλης ROC για το κατώφλι αυτό.

Σε κάθε σημείο της ROC που θα προκύψει υπολογίζουμε και τη μικρότερη απόσταση της καμπύλης με το σημείο τέλειας ταξινόμησης (0,1). Αυτό θα μας βοηθήσει στην σύγκριση των μοντέλων.

## 3.4. Standard Deviation της Καμπύλης ROC

Για να επιτύχουμε όμως ακόμα μεγαλύτερη ακρίβεια στον υπολογισμό της τυπικής απόκλισης, δεν ακλουθούμε τον κλασικό ορισμό αλλά μια παραλλαγή αυτού ειδικά προσαρμοσμένη στα δεδομένα μας.

Έστω λοιπόν ότι έχουμε ένα τυχαίο σημείο πάνω στη μέση ROC το M($x_m$ ,$y_m$) , εμείς θέλουμε να βρούμε πάνω στην ευθεία που ενώνει το σημείο M με το σημείο (0,1) - γιατί αυτό είναι το πεδίο ενδιαφέροντός μας, η τυπική απόκλιση από το τέλειο classification – άλλα δύο σημεία το B, M+Std και το C, M-Std. Τα νέα σημεία που θα προκύψουν θα προσδιορίσουν δύο νέες καμπύλες που θα απεικονίζουν τις τυπικές αποκλίσεις της ROC προς το καλύτερο classification και προς το χειρότερο.

Ας δούμε πώς υπολογίζουμε τις συντεταγμένες των σημείων αυτών:

Για το σημείο B έχουμε:

$$X_b = X_m - \frac{d_m}{D} X_m \ , \ Y_b = Y_m + \frac{d_m}{D} * (1 - Y_m)$$

Για το σημείο B έχουμε:

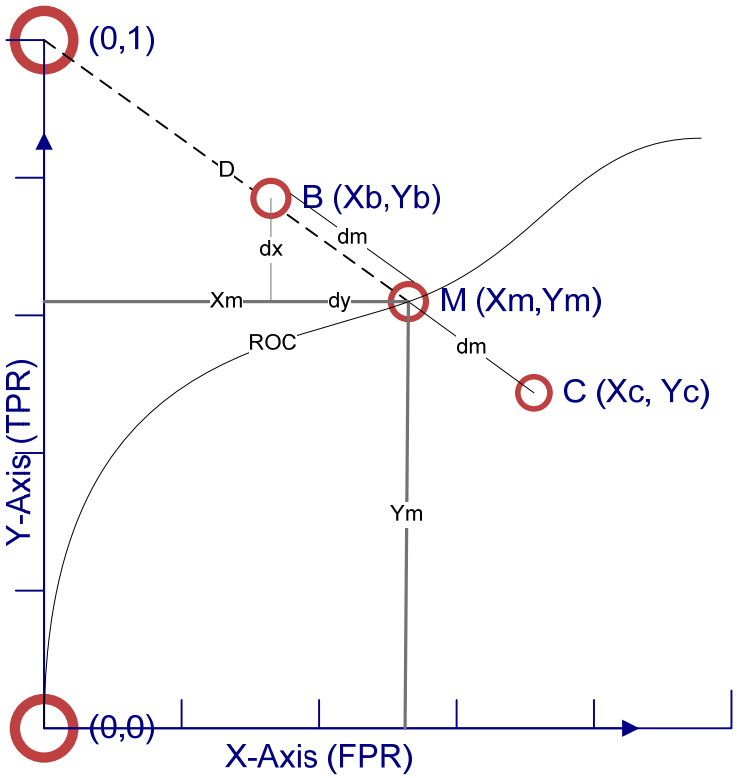
$$X_c = X_m + \frac{d_m}{D} X_m \, , \, Y_c = Y_m - \frac{d_m}{D} * (1 - Y_m)$$

D είναι η ευκλείδεια απόσταση του M από το (0,1) : $D = \sqrt{X_m^2 + (1 - Y_m)^2}$

Και $d_m$ είναι το standard deviation για το συγκεκριμένο σημείο M. Το $d_m$ υπολογίζεται ως εξής : Αφού το σημείο M προκύπτει από τη μέση απόσταση των δέκα σημείων που προέκυψαν από της 10 ROC, μία για κάθε fold, το std του συγκεκριμένου σημείο θα προκύψει υπολογίζοντας την μέση απόσταση και των δέκα σημείων από το μέσο M. Αν δηλαδή $d_1$ η απόσταση του σημείου της ROC για το fold 1 από το σημείο M, $d_2$ η απόσταση του σημείου της ROC για το fold 2 από το

σημείο M, …. Και $d_{10}$ η απόσταση του σημείου της ROC για το fold 10 από το σημείο M τότε το std $d_m$: $d_m = \sqrt{\frac{d_1^2+d_2^2+d_3^2+d_4^2+d_5^2+d_6^2+d_7^2+d_8^2+d_9^2+d_{10}^2}{10}}$

Ας δούμε το παρακάτω σχήμα για να κατανοήσουμε και γεωμετρικά την εύρεση του std




Εικόνα 8 Υπολογισμός του Standard Deviation για κάθε σημείο της καμπύλης ROC

# 4. Αξιολόγηση

Στο κεφάλαιο αυτό θα δούμε τα αποτελέσματα όλων των μεθόδων που χρησιμοποιήθηκαν κατά την υλοποίηση. Αρχικά η αξιολόγηση θα γίνει για το τυπικό Error Rate προκύπτει από το Classification, έπειτα έχουμε τις καμπύλες ROC για τα 10 Folds, τέλος παρουσιάζουμε την μέση καμπύλη ROC και την τυπική της απόκλιση που είναι και η τελική αξιολόγηση των μοντέλων εκπαίδευσης. Εκεί θα γίνει και η επιλογή του classifier που αποτελεί το βέλτιστο μοντέλο ταξινόμησης για κάθε dataset ξεχωριστά και η επιλογή του classifier που υλοποιεί ένα ικανοποιητικό μοντέλο ταξινόμησης και για τα τρία datasets – και για τους τρείς τομείς ιατρικών δεδομένων που εξετάσαμε.

## 4.1. Classification

Για την παρουσίαση των αποτελεσμάτων του Classification χρησιμοποιήσαμε την εξής δομή για κάθε dataset ξεχωριστά χωρίζουμε τα αποτελέσματα σε αποτελέσματα που προέκυψαν από το Evaluation Set σε κάθε Fold και σε αποτελέσματα που προέκυψαν από Independent Test Set . Για τις δύο αυτές κατηγορίες υπάρχει (i) ένας συγκεντρωτικός πίνακας με όλους τους classifiers ο οποίος παρουσιάζει το Error Rate, το ποσοστό λανθασμένων αποφάσεων ανά κλάση, και το ποσοστό True Positives, False Positives, True Negatives, False Negatives , (ii) ακολουθούν ενδεικτικά για επιλεγμένους classifiers διαγράμματα συμπεριφοράς του Error Rate για κάθε fold , (iii) Το συγκεντρωτικό διάγραμμα Error Rate για όλους τους Classifiers με σειρά από την καλύτερη απόδοση προς την χειρότερη. (iv) το συγκεντρωτικό διάγραμμα συμπεριφοράς όλων των Classifiers σε κάθε fold και την επιλογή ενός κατωφλίου μέσα στο οποίο κινούνται οι ικανοποιητικές αποδόσεις σε Error Rate.

### 4.1.1. Classification Ovarian Tumor Dataset

Ο πίνακας 5 και 6 περιγράφουν όλες τις πληροφορίες που προέκυψαν από την διαδικασία του Classification και συγκρίνουν και τα 17 μοντέλα εκπαίδευσης για Evaluation Set και Independent Set.

| average folds<br>Ovarian Cancer Dataset | | | | | | Errors per Class | |
|---|---|---|---|---|---|---|---|
| EVALUATON SET | Error Rate | TP | FP | TN | FN | class1 | class2 |
| Parzenc | 0,09033516 | 90,97% | 0,00% | 0,00% | 9,03% | 0,00% | 100,00% |
| Rbnc | 0,09148394 | 90,50% | 0,47% | 0,35% | 8,68% | 0,51% | 96,07% |
| Fisherc | 0,09149795 | 90,50% | 0,47% | 0,35% | 8,80% | 0,52% | 96,25% |
| Polyc | 0,09149795 | 90,50% | 0,47% | 0,35% | 8,68% | 0,52% | 96,25% |
| Knnc | 0,09266074 | 90,73% | 0,23% | 0,00% | 9,15% | 0,26% | 100,00% |
| Klldc | 0,09376783 | 89,80% | 1,17% | 0,83% | 8,33% | 1,29% | 90,71% |
| Ldc | 0,09376783 | 89,80% | 1,17% | 0,83% | 8,33% | 1,29% | 90,71% |
| Pcldc | 0,09376783 | 89,80% | 1,17% | 0,83% | 8,21% | 1,29% | 90,71% |
| Loglc | 0,09380985 | 90,27% | 0,70% | 0,35% | 8,80% | 0,77% | 96,25% |
| Treec | 0,12088561 | 87,09% | 3,88% | 0,82% | 8,21% | 4,26% | 90,71% |
| Qdc | 0,12297603 | 86,52% | 4,44% | 1,18% | 7,86% | 4,89% | 86,96% |
| Quadrc | 0,12297603 | 86,52% | 4,44% | 1,18% | 7,86% | 4,89% | 86,96% |
| Parzendc | 0,13123933 | 84,99% | 5,97% | 1,88% | 7,15% | 6,58% | 78,93% |
| Naivebc | 0,13940688 | 84,65% | 6,32% | 1,41% | 7,62% | 6,96% | 84,29% |
| Udc | 0,14750472 | 83,01% | 7,96% | 2,24% | 6,79% | 8,76% | 74,82% |
| Nmsc | 0,15342877 | 82,89% | 8,08% | 1,77% | 7,26% | 8,89% | 80,00% |
| Nmc | 0,40816657 | 55,54% | 35,43% | 3,64% | 5,39% | 38,97% | 59,29% |

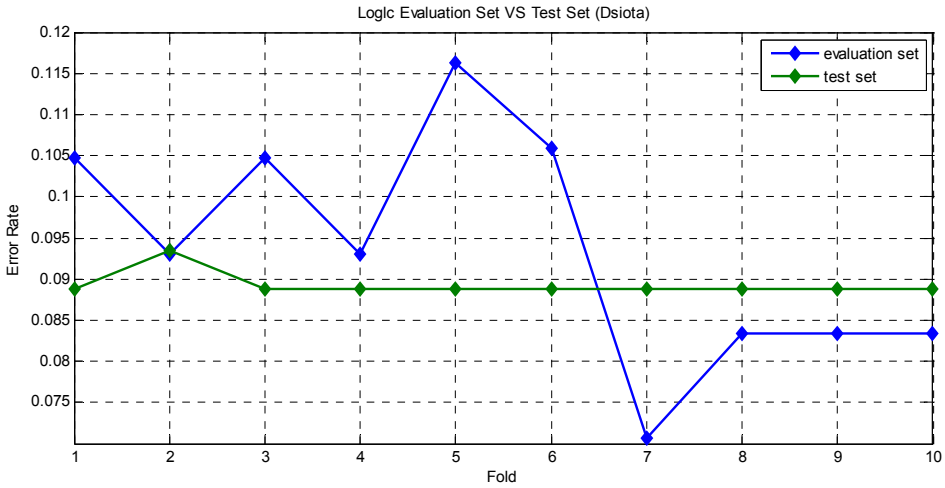
**Πίνακας 5: Ovarian Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Evaluation Set**

| average folds<br>Ovarian Cancer Dataset | | | | | | Errors per Class | |
|---|---|---|---|---|---|---|---|
| TEST SET | Error Rate | TP | FP | TN | FN | class1 | class2 |
| Loglc | 0,08925234 | 90,19% | 0,47% | 0,89% | 8,46% | 0,52% | 90,50% |
| Parzenc | 0,09345794 | 90,65% | 0,00% | 0,00% | 9,35% | 0,00% | 100,00% |
| Knnc | 0,09392523 | 90,61% | 0,05% | 0,00% | 9,35% | 0,05% | 100,00% |
| Klldc | 0,0953271 | 89,53% | 1,12% | 0,93% | 8,41% | 1,24% | 90,00% |
| Ldc | 0,0953271 | 89,53% | 1,12% | 0,93% | 8,41% | 1,24% | 90,00% |
| Pcldc | 0,0953271 | 89,53% | 1,12% | 0,93% | 8,41% | 1,24% | 90,00% |
| Fisherc | 0,09626168 | 89,95% | 0,70% | 0,42% | 8,93% | 0,77% | 95,50% |
| Polyc | 0,09626168 | 89,95% | 0,70% | 0,42% | 8,93% | 0,77% | 95,50% |
| Rbnc | 0,09766355 | 90,23% | 0,42% | 0,00% | 9,35% | 0,46% | 100,00% |
| Qdc | 0,12429907 | 86,03% | 4,63% | 1,54% | 7,80% | 5,10% | 83,50% |
| Quadrc | 0,12429907 | 86,03% | 4,63% | 1,54% | 7,80% | 5,10% | 83,50% |
| Udc | 0,12897196 | 82,94% | 7,71% | 4,16% | 5,19% | 8,51% | 55,50% |
| Treec | 0,13738318 | 84,21% | 6,45% | 2,06% | 7,29% | 7,11% | 78,00% |
| Naivebc | 0,14485981 | 81,17% | 9,49% | 4,35% | 5,00% | 10,46% | 53,50% |
| Parzendc | 0,15420561 | 82,66% | 7,99% | 1,92% | 7,43% | 8,81% | 79,50% |
| Nmsc | 0,15607477 | 80,61% | 10,05% | 3,79% | 5,56% | 11,08% | 59,50% |
| Nmc | 0,39485981 | 55,23% | 35,42% | 5,28% | 4,07% | 39,07% | 43,50% |

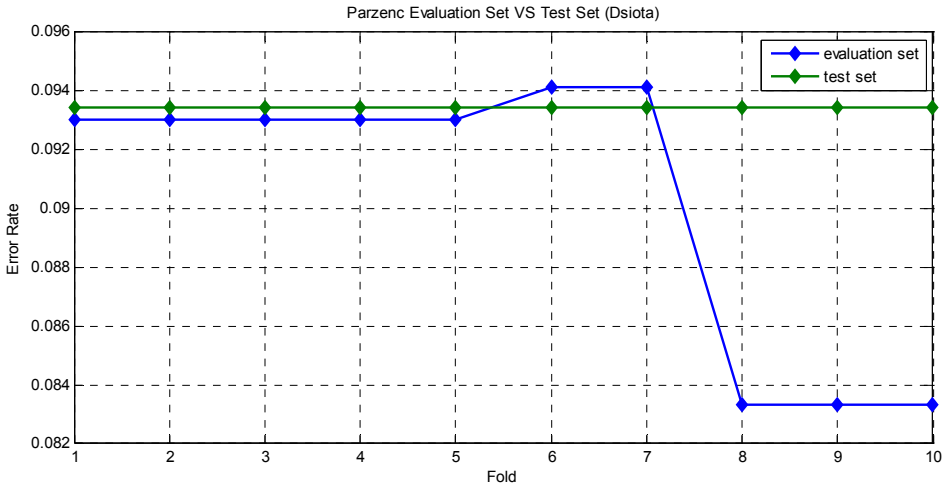
**Πίνακας 6: Ovarian Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Independent Set**

Παρατηρούμε και στους δύο πίνακες ότι οι Classifiers που έχουν το μικρότερο Error Rate και είναι η επιλογή μας προς το παρόν για το καλύτερο performance , έχουν ποσοστό αποτυχίας αναγνώρισης των δειγμάτων της δεύτερης κλάσης κοντά στο 100% και αυτό δηλώνει ξεκάθαρα το μειονέκτημα της επιλογής βέλτιστου μοντέλου ταξινόμησης βάσει Error Rate και την αναγκαιότητα της καμπύλης ROC σαν μετρική για το performance.

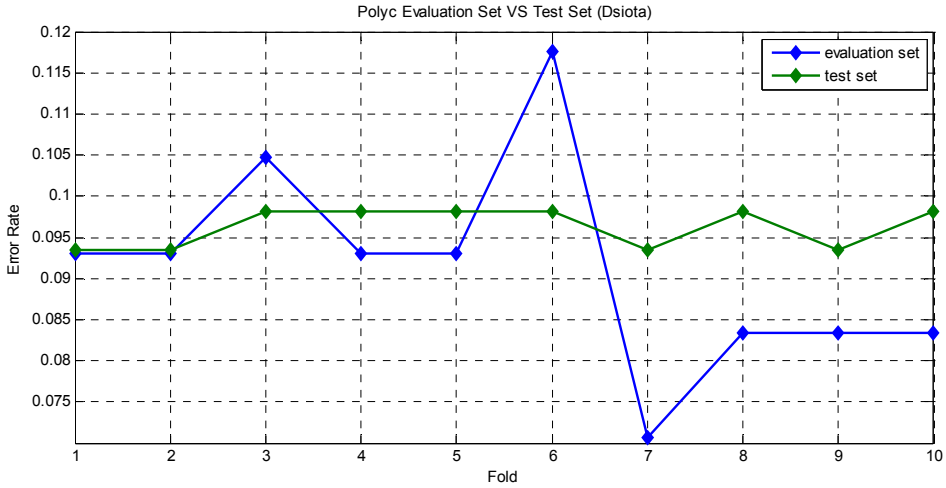
Στη συνέχεια έχουμε τα διάγραμμα συμπεριφοράς Error Rate σε κάθε fold για evaluation set και independent test set. Βλέπουμε τη συμπεριφορά του Error Rate για κάθε fold για τρία αντιπροσωπευτικά δείγματα classifier.




**Εικόνα 9: Ovarian Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Loglc Classifier**

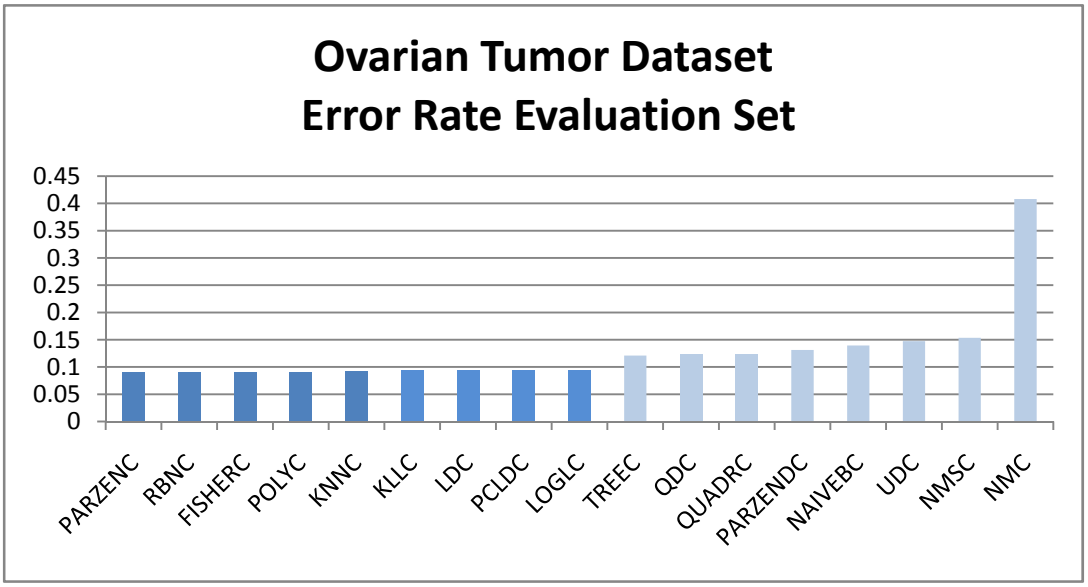



**Εικόνα 10: Ovarian Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Parzenc Classifier**



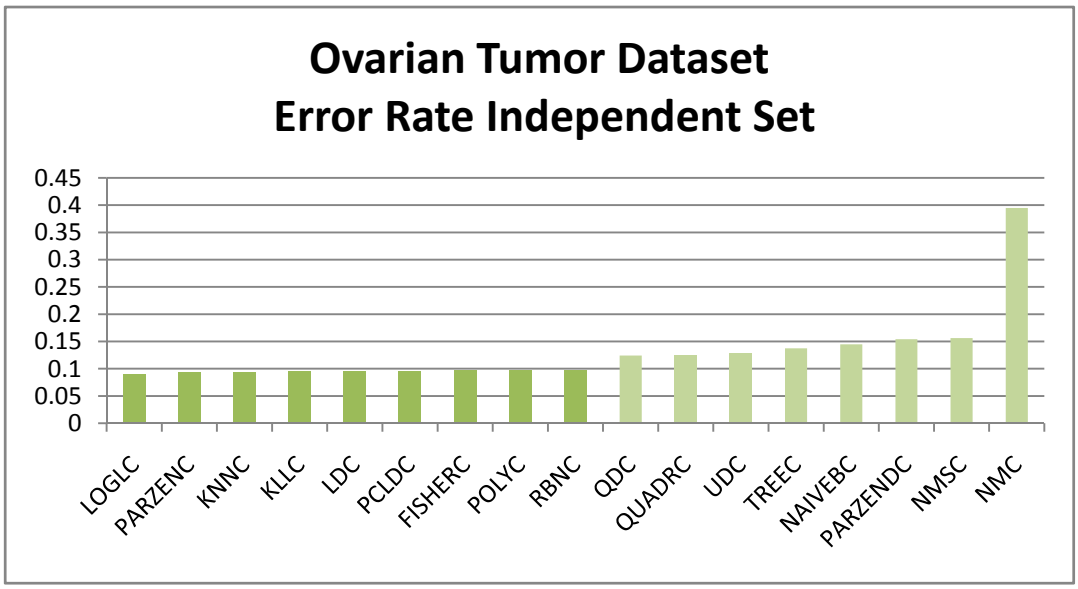

**Εικόνα 11: Ovarian Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Polyc Classifier**

Στα συνολικά διαγράμματα σύγκρισης μέσου Error Rate, βλέπουμε το διάγραμμα μέσου Error Rate για όλους τους classifiers για το Evaluation Set, η σειρά είναι από την καλύτερη επίδοση προς την χειρότερη. Η αλλαγή χρώματος οριοθετεί το κατώφλι επιλογής classifier με βάση την τιμή του Error Rate.



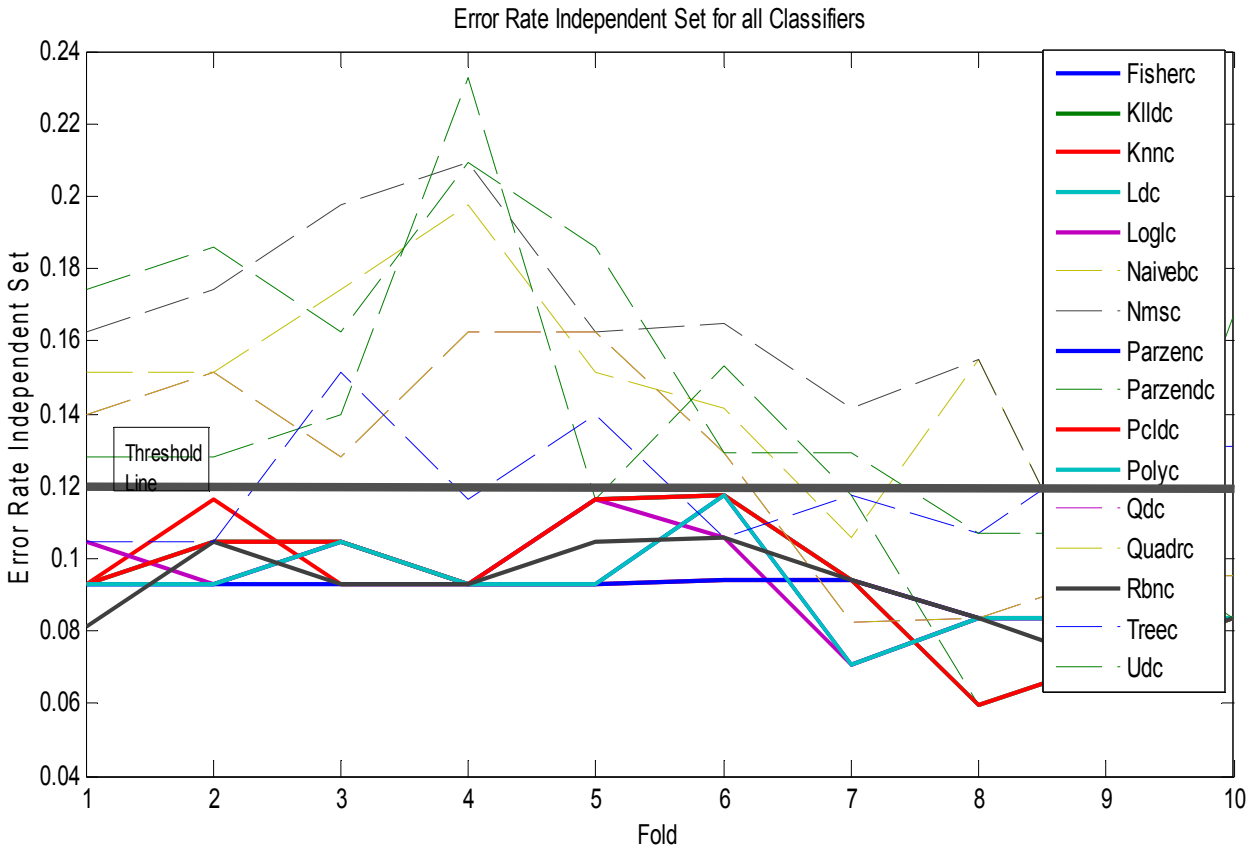

**Εικόνα 12: Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers - Ovarian Tumor Dataset – Evaluation Set**

Το διάγραμμα μέσου Error Rate για όλους τους classifiers για το Independent Test Set, η σειρά είναι από την καλύτερη επίδοση προς την χειρότερη.



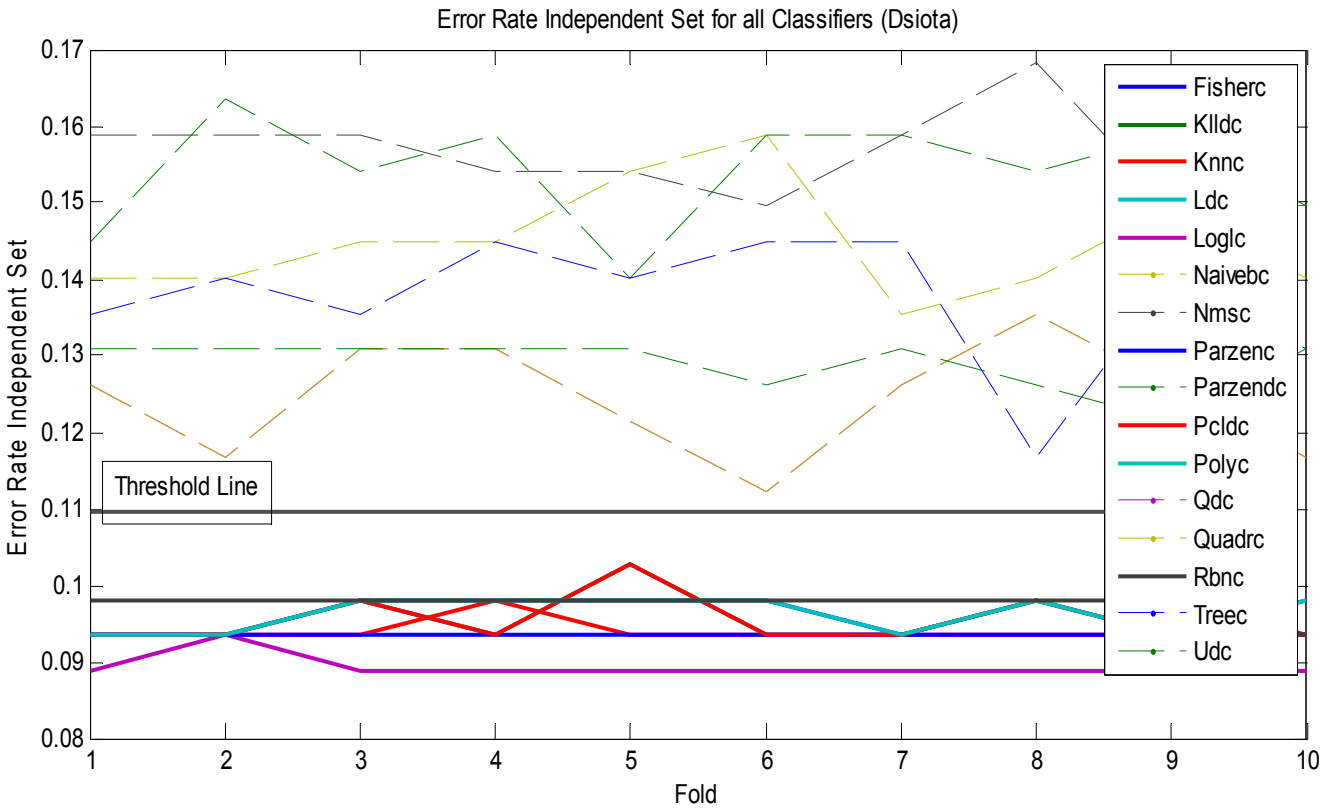

**Εικόνα 13: Ovarian Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers - Independent Set**

Τα συνολικά διαγράμματα performance των καλύτερων σε επιδόσεις classifiers στο Evaluation Set παρουσιάζονται στην εικόνα 13 . Βλέπουμε τις επιδόσεις Error Rate σε κάθε fold. Με έντονο χρώμα είναι οι Classifiers που έχουν τις καλύτερες επιδόσεις. Μπορούμε να πούμε πως μια ικανοποιητική εικόνα παρουσιάζουν οι Fisherc, Klldc, Knnc, Ldc, Loglc, Parzenc, Pcldc,Polyc, Rbnc.




Εικόνα 14 : Ovarian Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Evaluation Set

Τα συνολικά διαγράμματα performance των καλύτερων σε επιδόσεις classifiers στο Independent Test Set παρουσιάζονται στην εικόνα 14. Βλέπουμε τις επιδόσεις Error Rate σε κάθε fold. Με έντονο χρώμα είναι οι Classifiers που έχουν τις καλύτερες επιδόσεις. Μπορούμε να πούμε πως μια ικανοποιητική εικόνα παρουσιάζουν οι Fisherc, Klldc, Knnc, Ldc, Loglc, Parzenc, Pcldc,Polyc, Rbnc.




Εικόνα 15: Ovarian Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Independent Set

#### 4.1.1.1. Επιλογή μοντέλου ταξινόμησης για Ovarian Tumor Dataset

Το χαμηλότερο Error Rate βρέθηκε για τους Parzenc και Loglc classifiers. Αυτό όμως απαιτεί περαιτέρω διερεύνηση καθώς στον Parzenc το ποσοστιαίο σφάλμα αναγνώρισης της κλάσης 2 είναι 100%, δηλαδή δεν αναγνωρίστηκε σωστά κανένα από τα δείγματα της δεύτερης κλάσης.

### 4.1.2. Classification Brain Cancer Dataset

#### 4.1.2.1. Παρουσίαση αποτελεσμάτων Brain Cancer Dataset

Ο πίνακας 7 και 8 περιγράφουν όλες τις πληροφορίες που προέκυψαν από την διαδικασία του Classification και συγκρίνουν και τα 17 μοντέλα εκπαίδευσης για Evaluation Set και Independent Set.

| average folds – Brain Tumor Dataset EVALUATON SET | Error Rate | TP | FP | TN | FN | Errors per Class class1 | class2 |
|---|---|---|---|---|---|---|---|
| Klldc | 0,2425 | 61,38% | 8,63% | 14,38% | 15,50% | 12,32% | 52,08% |
| Ldc | 0,2425 | 61,38% | 8,63% | 14,38% | 15,50% | 12,32% | 52,08% |
| Pcldc | 0,2425 | 61,38% | 8,63% | 14,38% | 15,50% | 12,32% | 52,08% |
| Loglc | 0,24875 | 61,25% | 8,75% | 13,88% | 16,00% | 12,50% | 53,75% |
| Fisherc | 0,25 | 62,25% | 7,75% | 12,75% | 17,13% | 11,07% | 57,50% |
| Polyc | 0,25 | 62,25% | 7,75% | 12,75% | 17,13% | 11,07% | 57,50% |
| Nmsc | 0,25375 | 58,88% | 11,13% | 15,75% | 13,75% | 15,89% | 47,50% |
| Parzenc | 0,25375 | 64,63% | 5,38% | 10,00% | 19,75% | 7,68% | 66,67% |
| Naivebc | 0,255 | 62,25% | 7,75% | 12,25% | 17,75% | 11,07% | 59,17% |
| Knnc | 0,2575 | 63,00% | 7,00% | 11,25% | 18,38% | 10,00% | 62,50% |
| Qdc | 0,27 | 56,25% | 13,75% | 16,75% | 12,38% | 19,64% | 44,17% |
| Quadrc | 0,27 | 56,25% | 13,75% | 16,75% | 12,38% | 19,64% | 44,17% |
| Parzendc | 0,27625 | 61,63% | 8,38% | 10,75% | 18,88% | 11,96% | 64,17% |
| Udc | 0,28125 | 54,75% | 15,25% | 17,13% | 12,13% | 21,79% | 42,92% |
| Rbnc | 0,29875 | 70,00% | 0,00% | 0,13% | 29,88% | 0,00% | 99,58% |
| Nmc | 0,30375 | 47,25% | 22,75% | 22,38% | 7,50% | 32,50% | 25,42% |
| Treec | 0,33 | 61,13% | 8,88% | 5,88% | 24,00% | 12,68% | 80,42% |

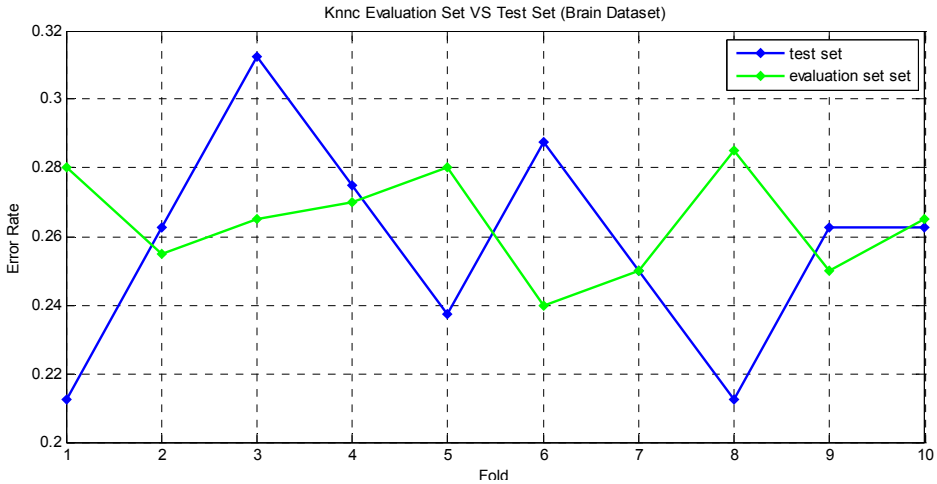
**Πίνακας 7 : Brain Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Evaluation Set**

| average folds –Brain Cancer TEST SET | Error Rate | TP | FP | TN | FN | Errors per Class class1 | class2 |
|---|---|---|---|---|---|---|---|
| Nmsc | 0,231 | 61,70% | 8,30% | 15,20% | 14,80% | 11,86% | 49,33% |
| Udc | 0,2375 | 57,85% | 12,15% | 18,40% | 11,60% | 17,36% | 38,67% |
| Naivebc | 0,251 | 62,85% | 7,15% | 12,05% | 17,95% | 10,21% | 59,83% |
| Loglc | 0,252 | 62,10% | 7,90% | 12,70% | 17,30% | 11,29% | 57,67% |
| Klldc | 0,2555 | 62,05% | 7,95% | 12,40% | 17,60% | 11,36% | 58,67% |
| Ldc | 0,2555 | 62,05% | 7,95% | 12,40% | 17,60% | 11,36% | 58,67% |
| Pcldc | 0,2555 | 62,05% | 7,95% | 12,40% | 17,60% | 11,36% | 58,67% |
| Knnc | 0,264 | 63,15% | 6,85% | 10,45% | 19,55% | 9,79% | 65,17% |
| Fisherc | 0,2645 | 62,35% | 7,65% | 11,20% | 18,80% | 10,93% | 62,67% |
| Polyc | 0,2645 | 62,35% | 7,65% | 11,20% | 18,80% | 10,93% | 62,67% |
| Parzenc | 0,2835 | 65,10% | 4,90% | 6,55% | 23,45% | 7,00% | 78,17% |
| Parzendc | 0,2835 | 62,70% | 7,30% | 8,95% | 21,05% | 10,43% | 70,17% |
| Qdc | 0,2835 | 56,85% | 13,15% | 14,80% | 15,20% | 18,79% | 50,67% |
| Quadrc | 0,2835 | 56,85% | 13,15% | 14,80% | 15,20% | 18,79% | 50,67% |
| Rbnc | 0,3 | 70,00% | 0,00% | 0,00% | 30,00% | 0,00% | 100,00% |
| Treec | 0,3045 | 64,55% | 5,45% | 5,00% | 25,00% | 7,79% | 83,33% |
| Nmc | 0,334 | 45,45% | 24,55% | 21,15% | 8,85% | 35,07% | 29,50% |

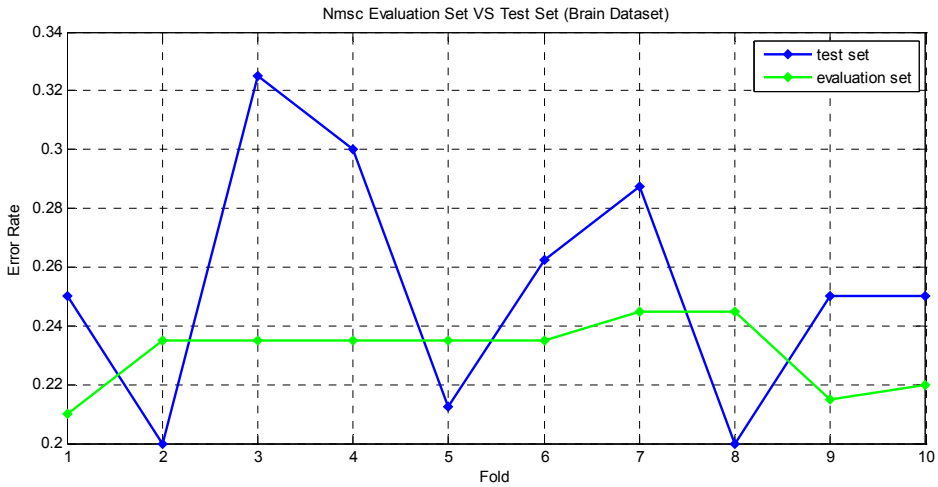
**Πίνακας 8: Brain Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Independent Set**

Σε αυτό το dataset δεν παρατηρείται το φαινόμενο της 100% αποτυχίας αναγνώρισης των δειγμάτων της κλάσης #2. Αυτό συμβαίνει γιατί το dataset έχει πιο ομοιόμορφα κατανεμημένο τον αριθμό δειγμάτων και στις δύο κλάσεις. (Είναι 70% του dataset στην κλάση#1 και 30% στην κλάση #2 ενώ το Ovarian Tumor Dataset 90,9% του dataset στην κλάση#1 και 9,1% στην κλάση #2)

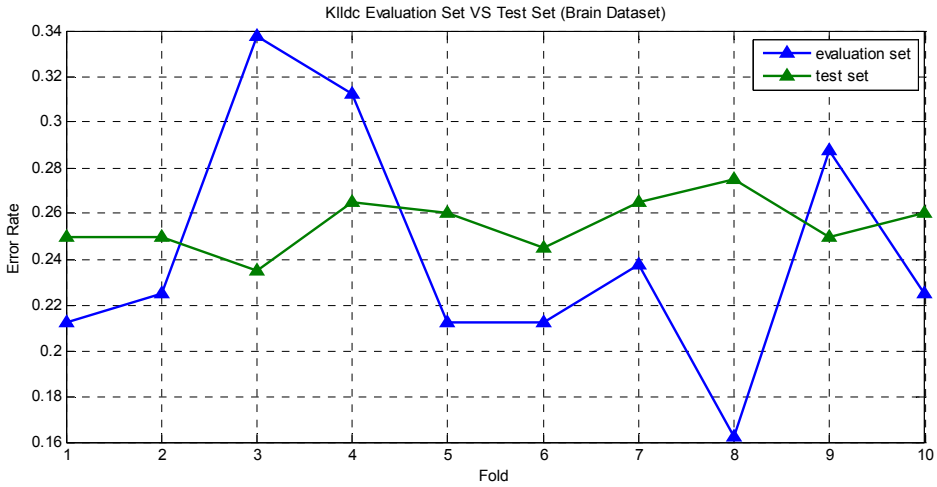
Στη συνέχεια έχουμε τα διάγραμμα συμπεριφοράς Error Rate σε κάθε fold για evaluation set και independent test set. Βλέπουμε τη συμπεριφορά του Error Rate για κάθε fold για τρία αντιπροσωπευτικά δείγματα classifier.




**Εικόνα 17: Brain Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Knnc Classifier**

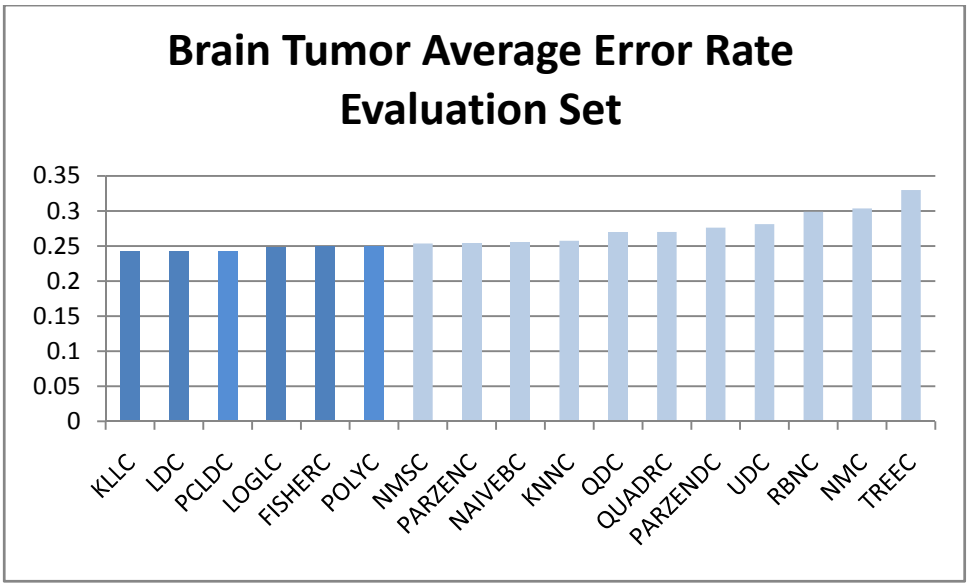



**Εικόνα 16: Brain Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Nmsc Classifier**



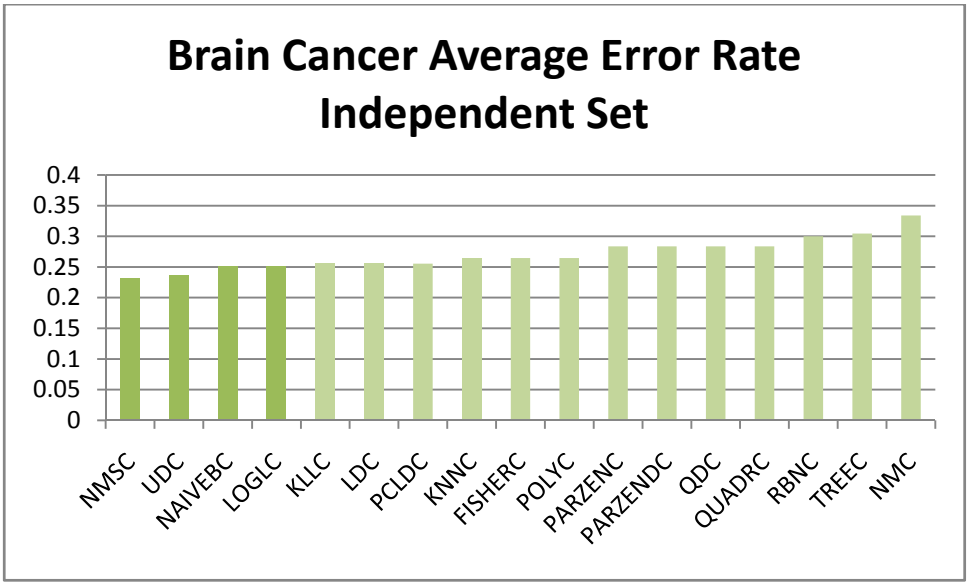

**Εικόνα 18 : Brain Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Klldc Classifier**

Στα συνολικά διαγράμματα σύγκρισης μέσου Error Rate, βλέπουμε το διάγραμμα μέσου Error Rate για όλους τους classifiers για το Evaluation Set, η σειρά είναι από την καλύτερη επίδοση προς την χειρότερη. Η αλλαγή χρώματος οριοθετεί το κατώφλι επιλογής classifier με βάση την τιμή του Error Rate.



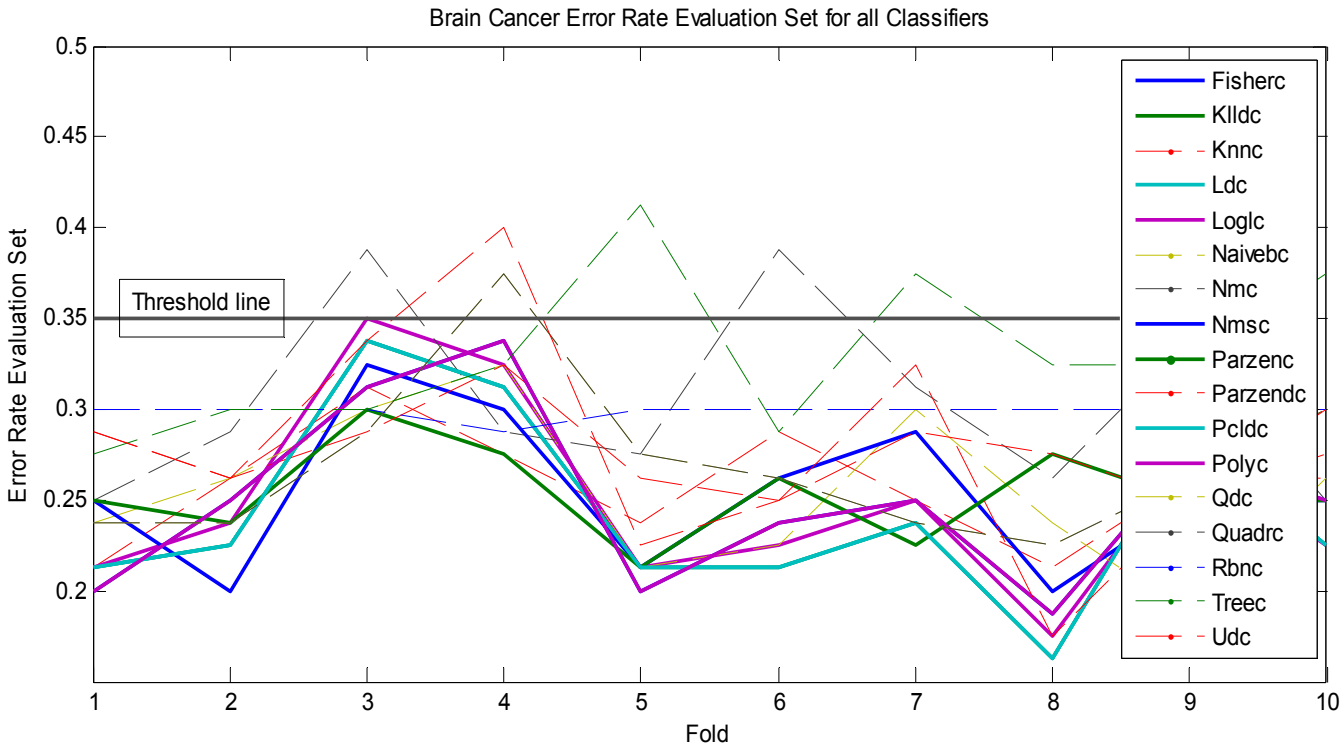

**Εικόνα 19: Brain Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers – Evaluation Set**

Ας δούμε το διάγραμμα μέσου Error Rate για όλους τους classifiers για το Independent Test Set, η σειρά είναι από την καλύτερη επίδοση προς την χειρότερη, Εικόνα 19.




Εικόνα 20 : Brain Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers – Independent Set

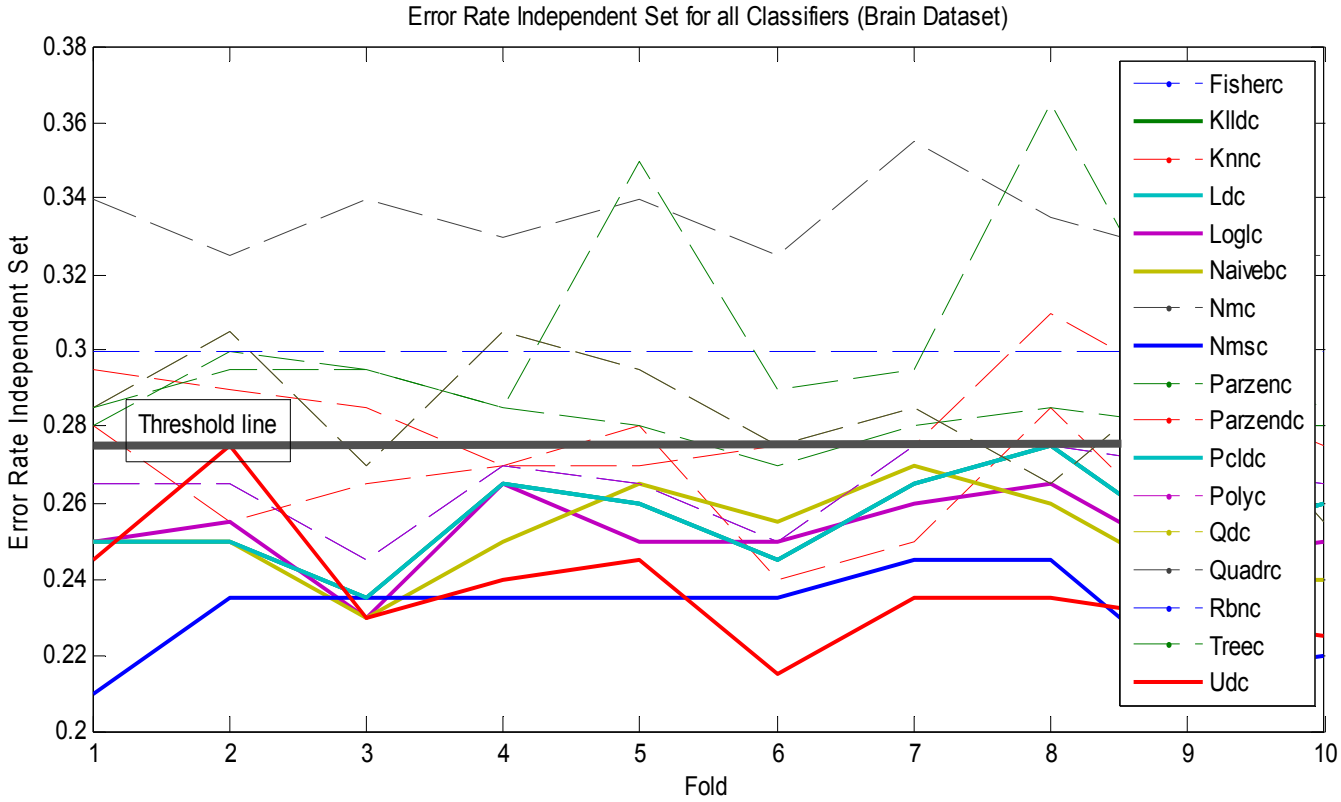
Τα συνολικά διαγράμματα performance των καλύτερων σε επιδόσεις classifiers στο Evaluation Set παρουσιάζονται στην εικόνα 20. Βλέπουμε τις επιδόσεις Error Rate σε κάθε fold. Με έντονο χρώμα είναι οι Classifiers που έχουν τις καλύτερες επιδόσεις. Μπορούμε να πούμε πως μια ικανοποιητική εικόνα παρουσιάζουν οι Fisherc, Klldc, Ldc, Loglc, Nmsc, Parzenc, Pcldc,Polyc.






Εικόνα 21 : Brain Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Evaluation Set

Τα συνολικά διαγράμματα performance των καλύτερων σε επιδόσεις classifiers στο Independent Test Set παρουσιάζονται στην εικόνα 21. Βλέπουμε τις επιδόσεις Error Rate σε κάθε fold. Με έντονο χρώμα είναι οι Classifiers που έχουν τις καλύτερες επιδόσεις. Μπορούμε να πούμε πως μια ικανοποιητική εικόνα παρουσιάζουν οι Klldc, Ldc, Loglc, Naivebc, Nmsc, Pcldc,Polyc, Udc.




Εικόνα 22: Brain Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Independent Set

#### 4.1.2.2. Επιλογή μοντέλου ταξινόμησης για Brain Cancer Dataset

Το χαμηλότερο Error Rate βρέθηκε για τον Nmsc classifier.

### 4.1.3. Classification Breast Cancer Dataset

#### 4.1.3.1. Παρουσίαση αποτελεσμάτων Breast Cancer Dataset

Ο πίνακας 9 και 10 περιγράφουν όλες τις πληροφορίες που προέκυψαν από την διαδικασία του Classification και συγκρίνουν και τα 17 μοντέλα εκπαίδευσης για Evaluation Set και Independent Set.

| average folds - Breast Tumor Dataset | | | | | | Errors per Class | |
|---|---|---|---|---|---|---|---|
| EVALUATON SET | Error Rate | TP | FP | TN | FN | class1 | class2 |
| Naivebc | 0,28571429 | 71,43% | 2,86% | 4,76% | 25,71% | 4,00% | 33,75% |
| Treec | 0,29047619 | 73,81% | 0,48% | 1,90% | 28,57% | 0,67% | 37,50% |
| Nmsc | 0,31904762 | 66,67% | 7,62% | 6,19% | 23,81% | 10,67% | 31,88% |
| Loglc | 0,32380952 | 67,14% | 7,14% | 5,24% | 24,29% | 10,00% | 33,13% |
| Klldc | 0,32380952 | 68,10% | 6,19% | 4,29% | 25,71% | 8,67% | 34,38% |
| Ldc | 0,32380952 | 68,10% | 6,19% | 4,29% | 25,71% | 8,67% | 34,38% |
| Pcldc | 0,32380952 | 68,10% | 6,19% | 4,29% | 25,71% | 8,67% | 34,38% |
| Fisherc | 0,32380952 | 68,57% | 5,71% | 3,81% | 26,19% | 8,00% | 35,00% |
| Parzenc | 0,32380952 | 69,05% | 5,24% | 3,33% | 26,67% | 7,33% | 35,63% |
| Polyc | 0,32380952 | 68,57% | 5,71% | 3,81% | 26,19% | 8,00% | 35,00% |
| Rbnc | 0,33809524 | 68,57% | 5,71% | 2,38% | 28,10% | 8,00% | 36,88% |
| Knnc | 0,33809524 | 65,71% | 8,57% | 5,24% | 24,76% | 12,00% | 33,13% |
| Parzendc | 0,34285714 | 66,67% | 7,62% | 3,81% | 26,67% | 10,67% | 35,00% |
| Qdc | 0,35714286 | 57,62% | 16,67% | 11,43% | 19,52% | 23,33% | 25,00% |
| Quadrc | 0,35714286 | 57,62% | 16,67% | 11,43% | 19,52% | 23,33% | 25,00% |
| Udc | 0,40952381 | 50,48% | 23,81% | 13,33% | 17,14% | 33,33% | 22,50% |
| Nmc | 0,43333333 | 42,86% | 31,43% | 18,57% | 11,90% | 44,00% | 15,63% |

Πίνακας 9 : Breast Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Evaluation Set

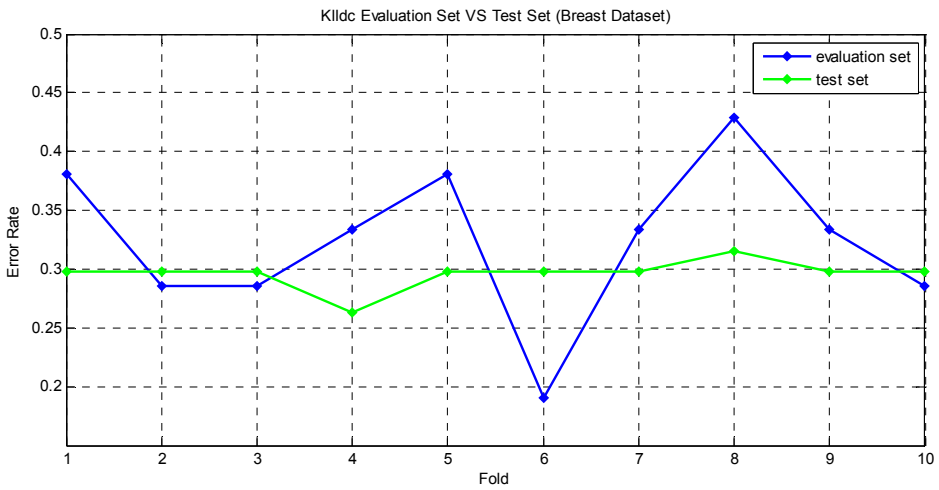
| average folds – Breast Tumor Dataset | | | | | | Errors per Class | |
|---|---|---|---|---|---|---|---|
| TEST SET | Error Rate | TP | FP | TN | FN | class1 | class2 |
| Parzenc | 0,27894737 | 69,47% | 0,70% | 2,63% | 27,19% | 1,00% | 91,18% |
| Klldc | 0,29649123 | 67,89% | 2,28% | 2,46% | 27,37% | 3,25% | 91,76% |
| Ldc | 0,29649123 | 67,89% | 2,28% | 2,46% | 27,37% | 3,25% | 91,76% |
| Pcldc | 0,29649123 | 67,89% | 2,28% | 2,46% | 27,37% | 3,25% | 91,76% |
| Rbnc | 0,29824561 | 70,18% | 0,00% | 0,00% | 29,82% | 0,00% | 100,00% |
| Treec | 0,3 | 70,00% | 0,18% | 0,00% | 29,82% | 0,25% | 100,00% |
| Naivebc | 0,30175439 | 69,30% | 0,88% | 0,53% | 29,30% | 1,25% | 98,24% |
| Fisherc | 0,30877193 | 67,89% | 2,28% | 1,23% | 28,60% | 3,25% | 95,88% |
| Polyc | 0,30877193 | 67,89% | 2,28% | 1,23% | 28,60% | 3,25% | 95,88% |
| Knnc | 0,31052632 | 66,32% | 3,86% | 2,63% | 27,19% | 5,50% | 91,18% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Loglc | 0,31754386 | 65,79% | 4,39% | 2,46% | 27,37% | 6,25% | 91,76% |
| Parzendc | 0,31929825 | 65,26% | 4,91% | 2,81% | 27,02% | 7,00% | 90,59% |
| Nmsc | 0,32280702 | 65,79% | 4,39% | 1,93% | 27,89% | 6,25% | 93,53% |
| Qdc | 0,41754386 | 52,98% | 17,19% | 5,26% | 24,56% | 24,50% | 82,35% |
| Quadrc | 0,41754386 | 52,98% | 17,19% | 5,26% | 24,56% | 24,50% | 82,35% |
| Udc | 0,45263158 | 48,77% | 21,40% | 5,96% | 23,86% | 30,50% | 80,00% |
| Nmc | 0,52280702 | 37,54% | 32,63% | 10,18% | 19,65% | 46,50% | 65,88% |

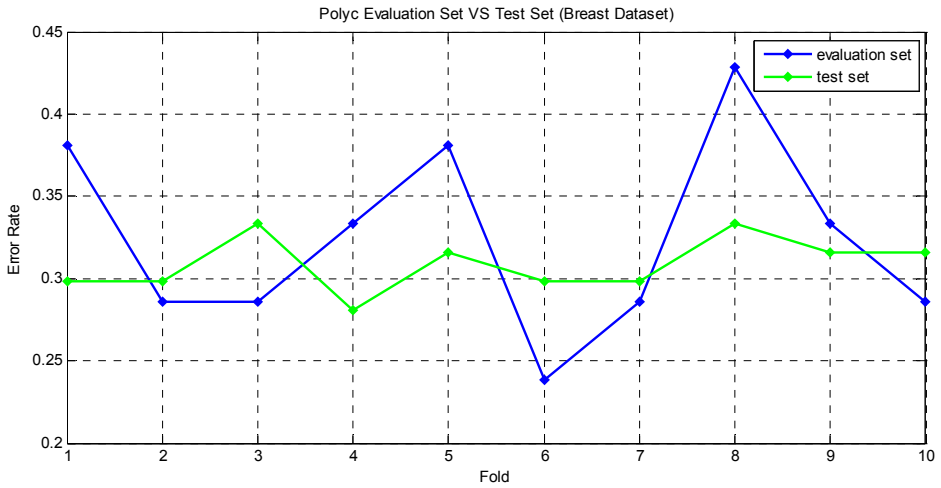
**Πίνακας 10 : Breast Tumor Dataset-Μέσο Error Rate για όλους τους Classifiers , Ποσοστιαίο Confusion Matrix , Errors per Class - 17 Classifiers – Evaluation Set**

Σε αυτό το dataset δεν παρατηρείται το φαινόμενο της 100% αποτυχίας αναγνώρισης των δειγμάτων της κλάσης #2. Αυτό συμβαίνει γιατί το dataset έχει πιο ομοιόμορφα κατανεμημένο τον αριθμό δειγμάτων και στις δύο κλάσεις. (Είναι 61% του dataset στην κλάση#1 και 39% στην κλάση #2 ενώ το Ovarian Tumor Dataset 90,9% του dataset στην κλάση#1 και 9,1% στην κλάση #2)

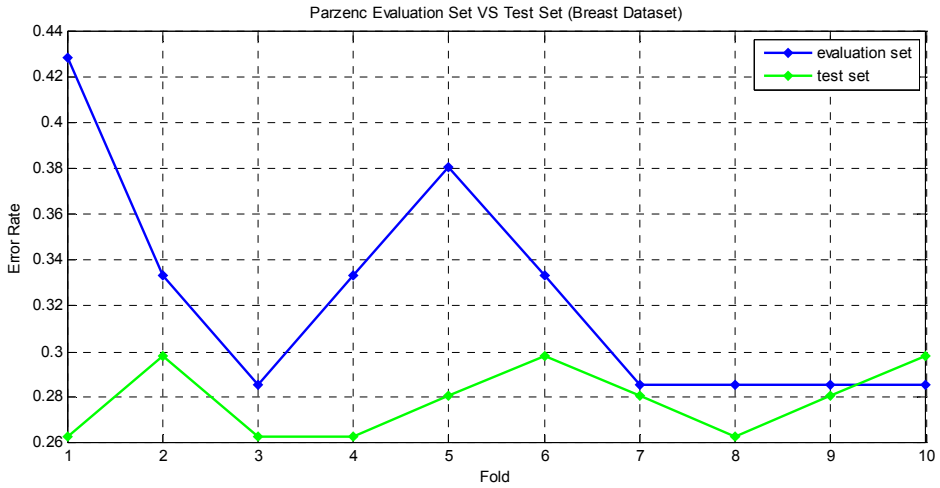
Στη συνέχεια έχουμε τα διάγραμμα συμπεριφοράς Error Rate σε κάθε fold για evaluation set και independent test set. Βλέπουμε τη συμπεριφορά του Error Rate για κάθε fold για τρία αντιπροσωπευτικά δείγματα classifier.




**Εικόνα 23: Breast Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Klldc Classifier**

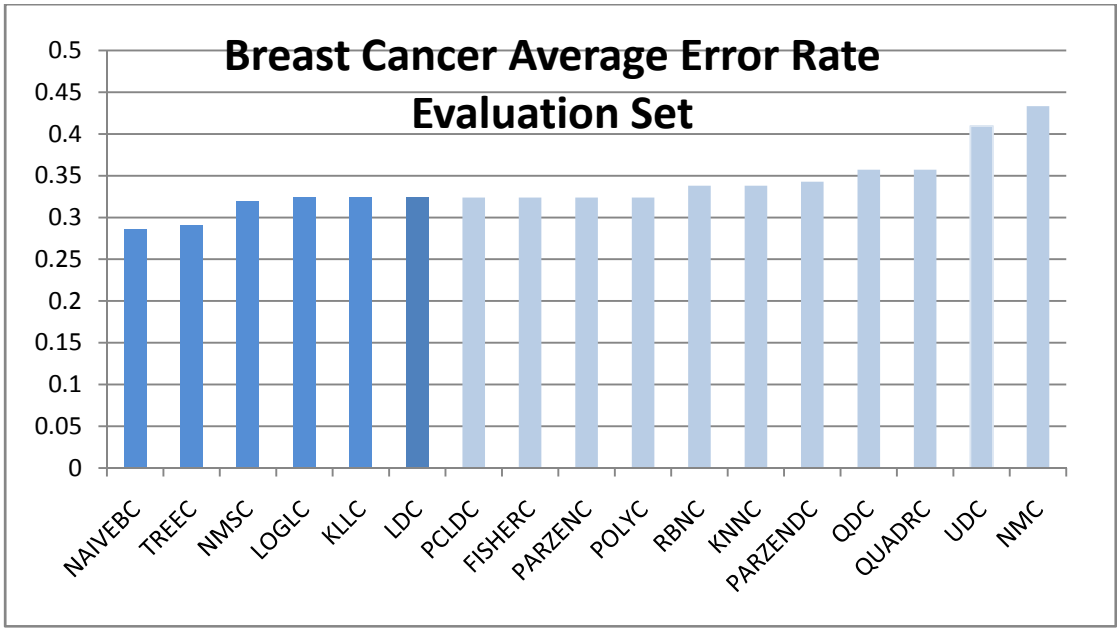



**Εικόνα 25: Breast Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Polyc Classifier**



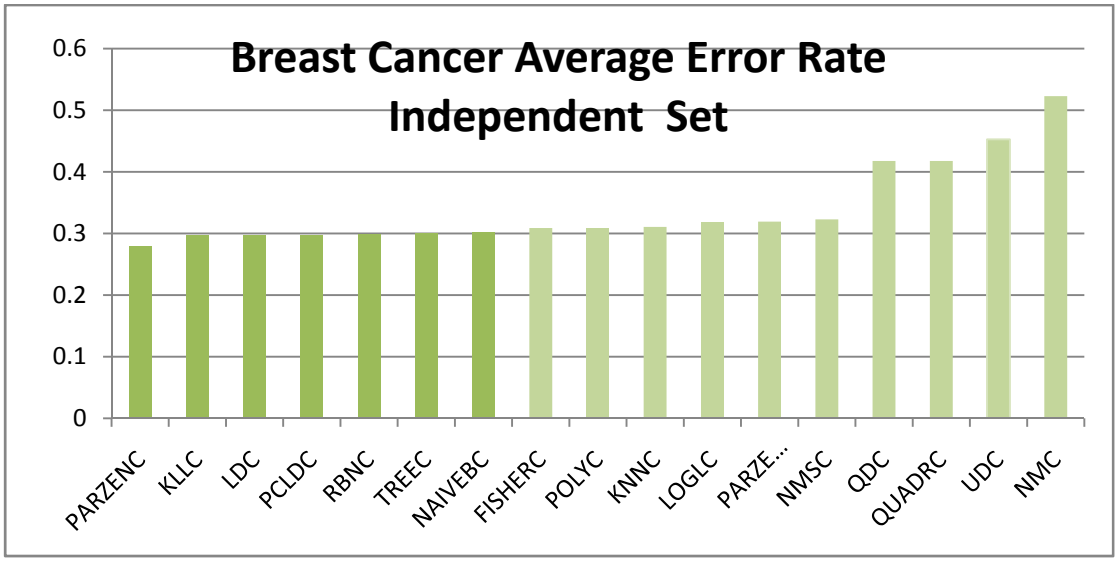

**Εικόνα 24: Breast Tumor Dataset- Συμπεριφορά Error Rate σε κάθε Fold – Evaluation Set, Independent Set – Parzenc Classifier**

Στα συνολικά διαγράμματα σύγκρισης μέσου Error Rate, βλέπουμε το διάγραμμα μέσου Error Rate για όλους τους classifiers για το Evaluation Set, η σειρά είναι από την καλύτερη επίδοση προς την χειρότερη. Η αλλαγή χρώματος οριοθετεί το κατώφλι επιλογής classifier με βάση την τιμή του Error Rate.



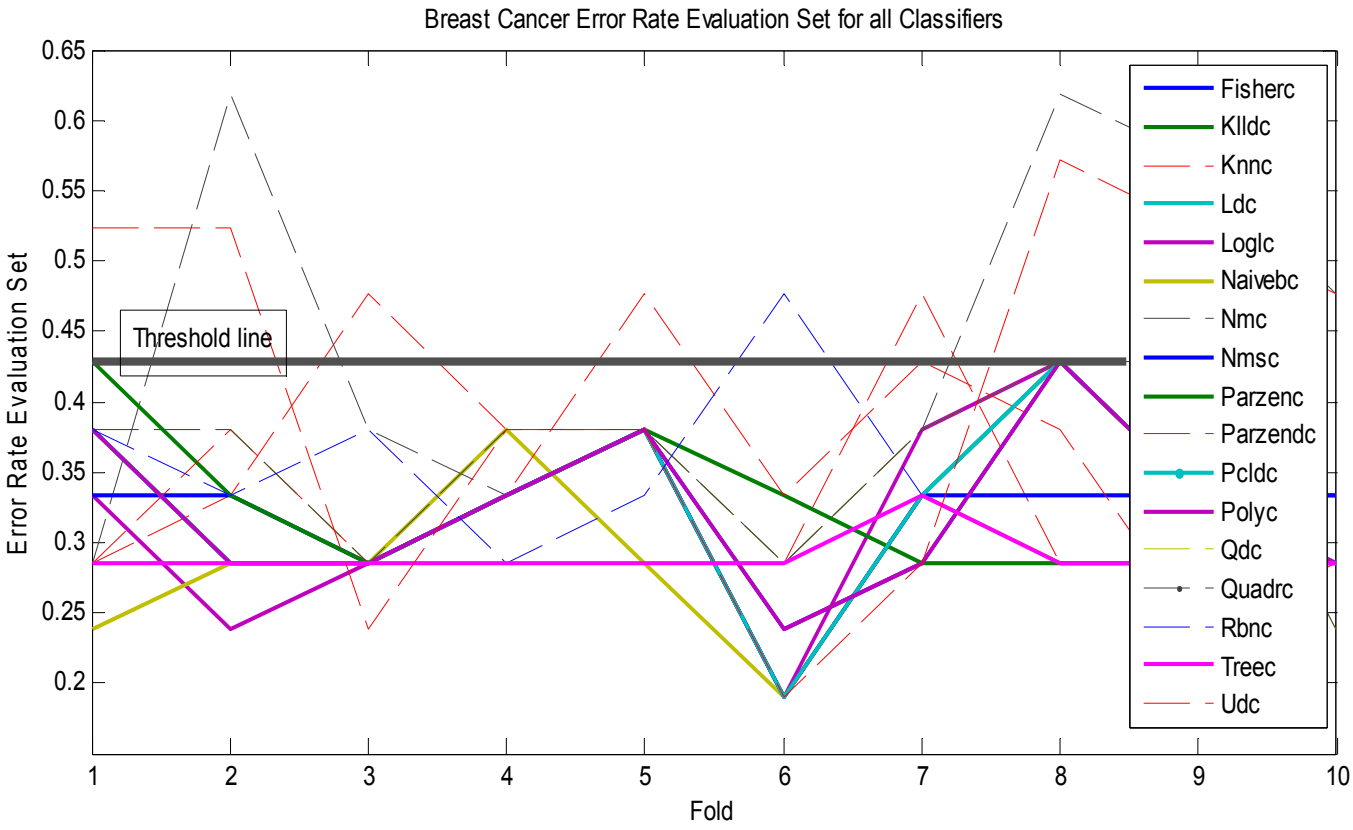

Εικόνα 26: Breast Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers – Evaluation Set

Ας δούμε το διάγραμμα μέσου Error Rate για όλους τους classifiers για το Independent Test Set, η σειρά είναι από την καλύτερη επίδοση προς την χειρότερη.



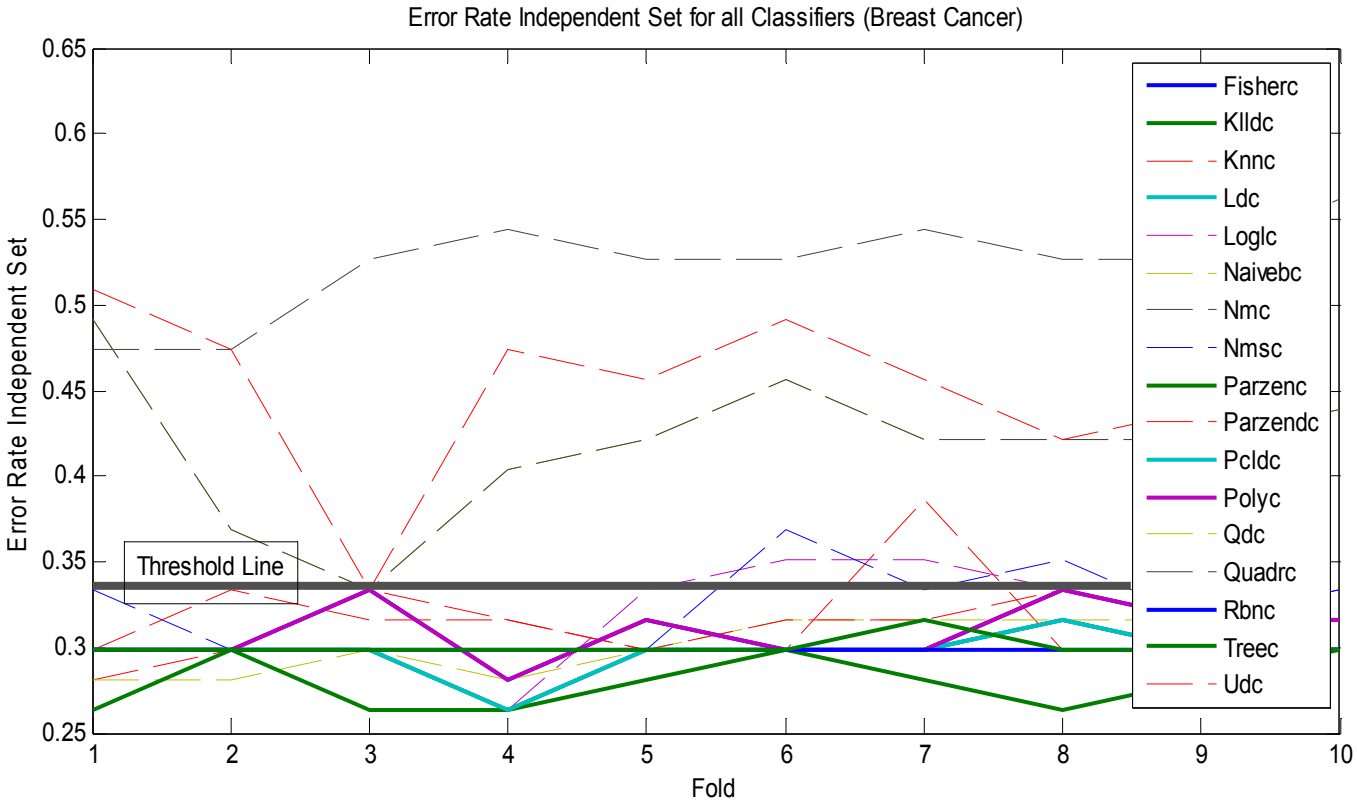

Εικόνα 27: Breast Tumor Dataset – Απόδοση με βάση το μικρότερο Error Rate - 17 Classifiers – Independent Set

Τα συνολικά διαγράμματα performance των καλύτερων σε επιδόσεις classifiers στο Evaluation Set παρουσιάζονται στο παρακάτω διάγραμμα. Βλέπουμε τις επιδόσεις Error Rate σε κάθε fold. Με έντονο χρώμα είναι οι Classifiers που έχουν τις καλύτερες επιδόσεις. Μπορούμε να πούμε πως μια ικανοποιητική εικόνα παρουσιάζουν οι Fisherc, Klldc, Ldc, Loglc, Naivebc, Nmsc, Parzenc, Pcldc, Polyc, Treec.




Εικόνα 28 : Breast Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Evaluation Set

Τα συνολικά διαγράμματα performance των καλύτερων σε επιδόσεις classifiers στο Independent Test Set παρουσιάζονται στο παρακάτω διάγραμμα. Βλέπουμε τις επιδόσεις Error Rate σε κάθε fold. Με έντονο χρώμα είναι οι Classifiers που έχουν τις καλύτερες επιδόσεις. Μπορούμε να πούμε πως μια ικανοποιητική εικόνα παρουσιάζουν οι Klldc, Ldc, Parzenc ,Polyc, Rbnc, Treec.




Εικόνα 29 : Breast Tumor Dataset - Error Rate ανά Fold - 17 Classifiers – Independent Set

#### 4.1.3.2. Επιλογή μοντέλου ταξινόμησης για Breast Cancer Dataset

Το χαμηλότερο Error Rate βρέθηκε για τους Parzenc classifier. Αυτό όμως απαιτεί περαιτέρω διερεύνηση καθώς στον Parzenc το ποσοστιαίο σφάλμα αναγνώρισης της κλάσης 2 είναι 91,18%, δηλαδή δεν αναγνωρίστηκε σωστά η συντριπτική πλειοψηφία από τα δείγματα της δεύτερης κλάσης.

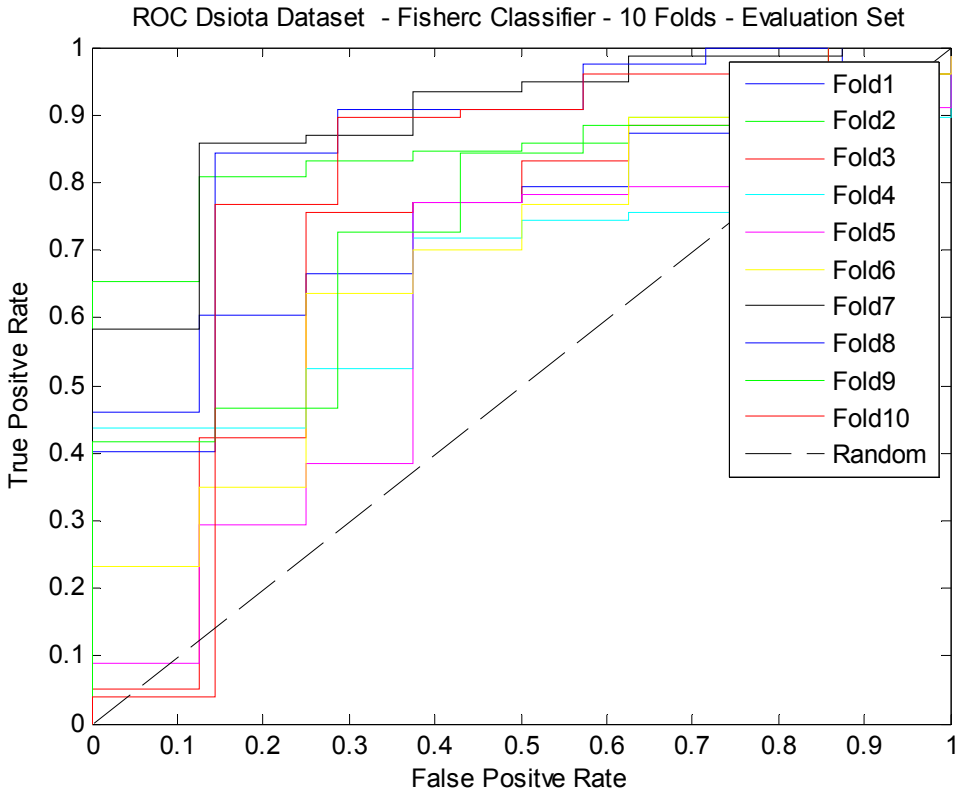
## 4.2. Καμπύλες ROC

### 4.2.1. Καμπύλες ROC 10-Folds

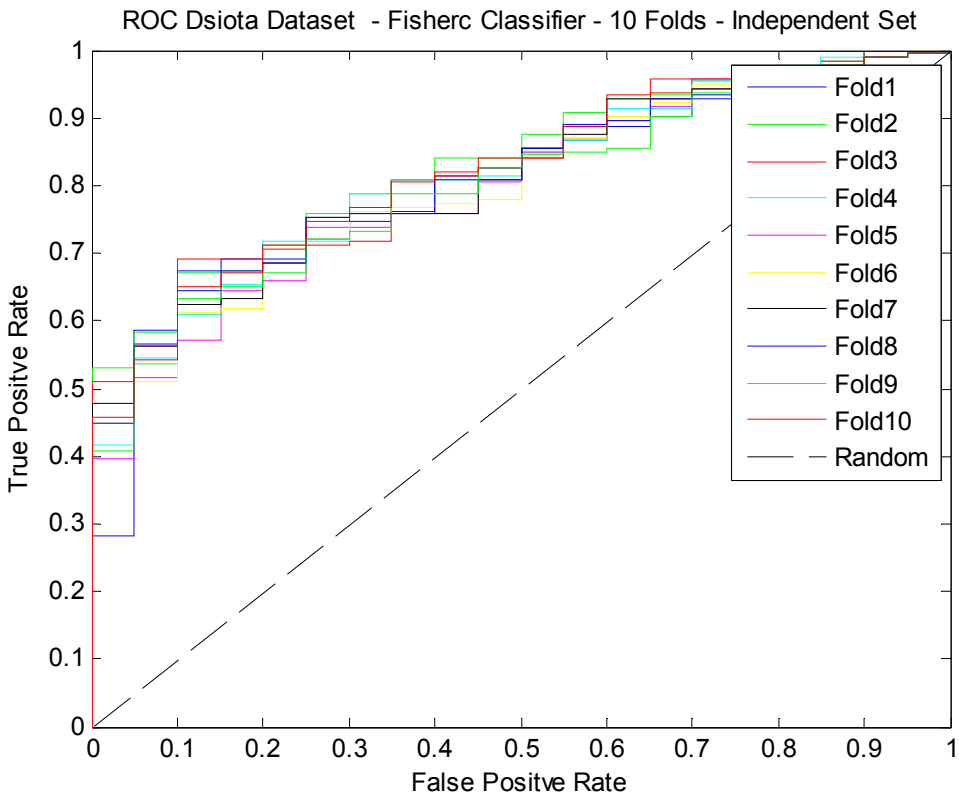
Αρχικά θα παρουσιάσουμε ενδεικτικά καμπύλες ROC, μία για κάθε FOLD σε ένα διάγραμμα ανά classifier. Από αυτές τις καμπύλες προέκυψαν οι μέσες καμπύλες ROC που θα δούμε στο τέλος και από αυτές θα αποφασιστεί το καλύτερο μοντέλο για κάθε dataset.

#### 4.2.1.1. Καμπύλες ROC 10 –Folds για Ovarian Tumor Dataset

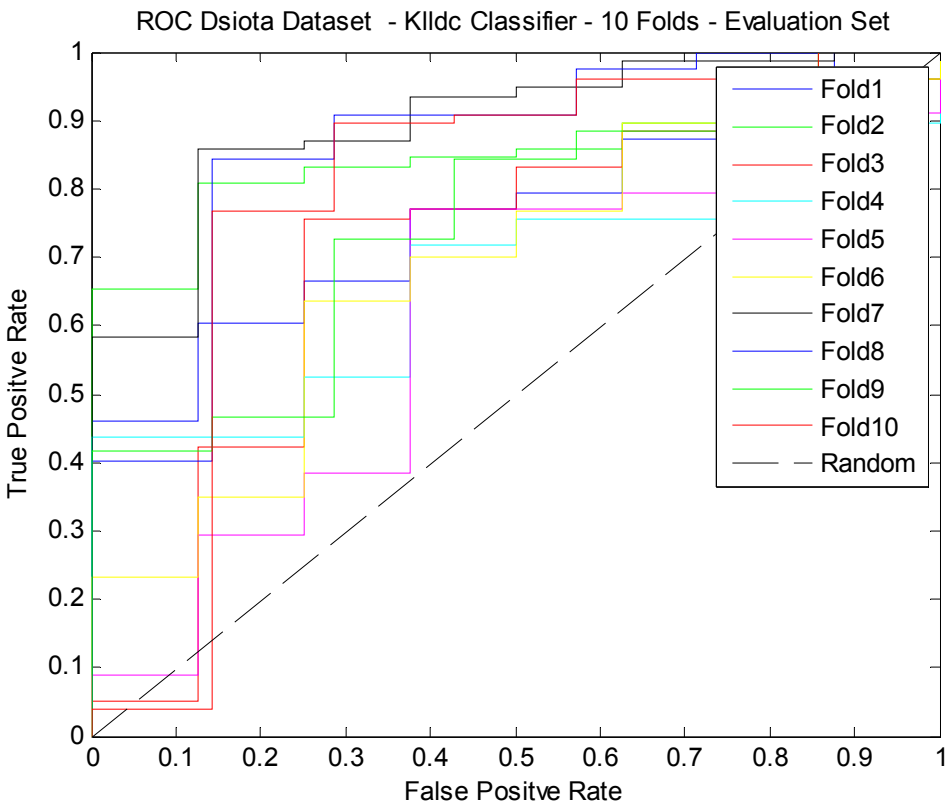
Για το Ovarian Tumor Dataset επιλέξαμε να δούμε τις ROC και από τα 10 Folds μαζί για τους classifiers Fisherc και Klldc. Βλέπουμε παρακάτω δύο διαγράμματα ένα για το Evaluation Set και ένα για το Independent Test Set.



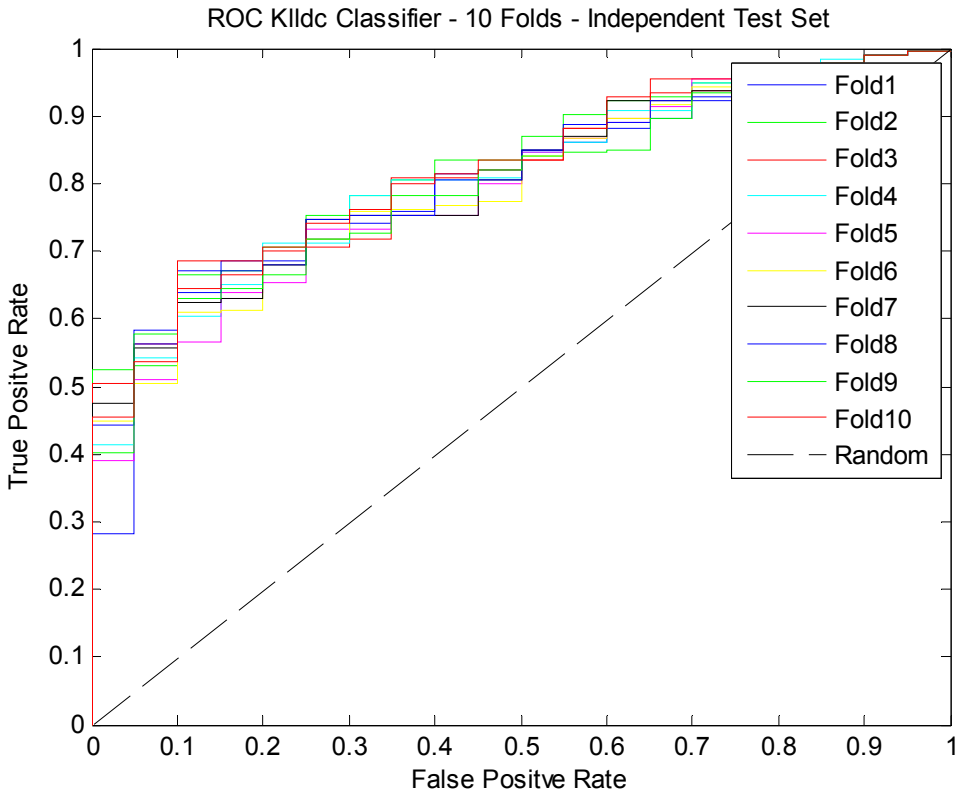

Εικόνα 30 : Ovarian Tumor Dataset – Roc 10 Folds – Fisherc Classifier – Evaluation Set




Εικόνα 31 : Ovarian Tumor Dataset – Roc 10 Folds – Fisherc Classifier – Independent Set




Εικόνα 32: Ovarian Tumor Dataset – Roc 10 Folds – Klldc Classifier – Evaluation Set

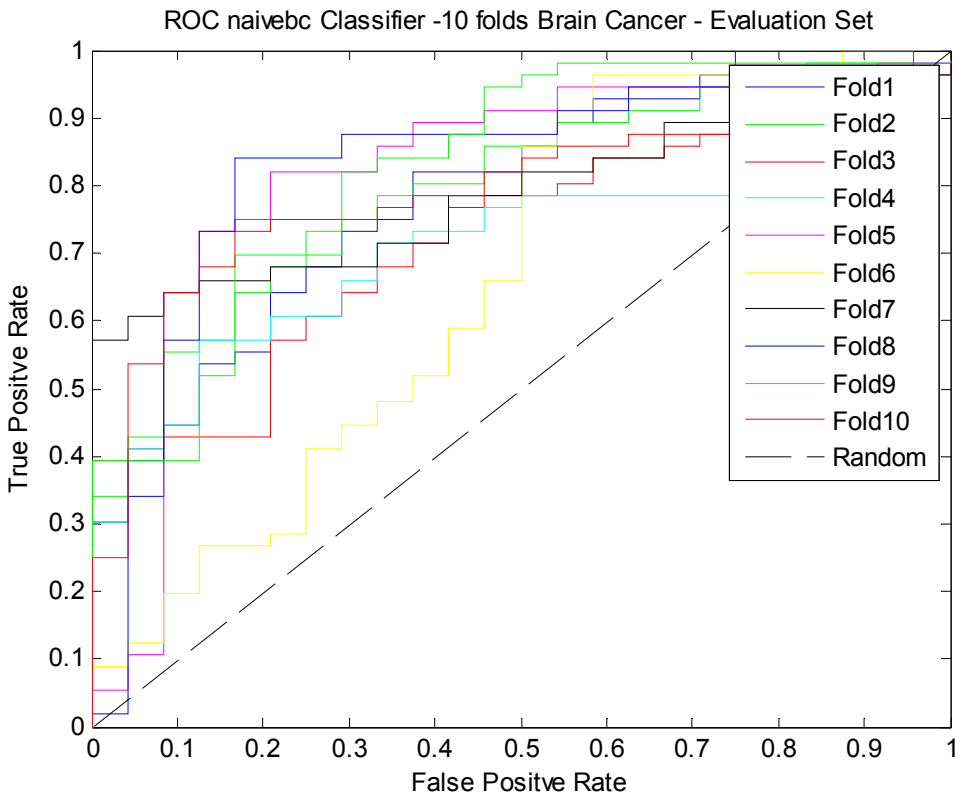



Εικόνα 33 : Ovarian Tumor Dataset – Roc 10 Folds – Klldc Classifier – Independent Set

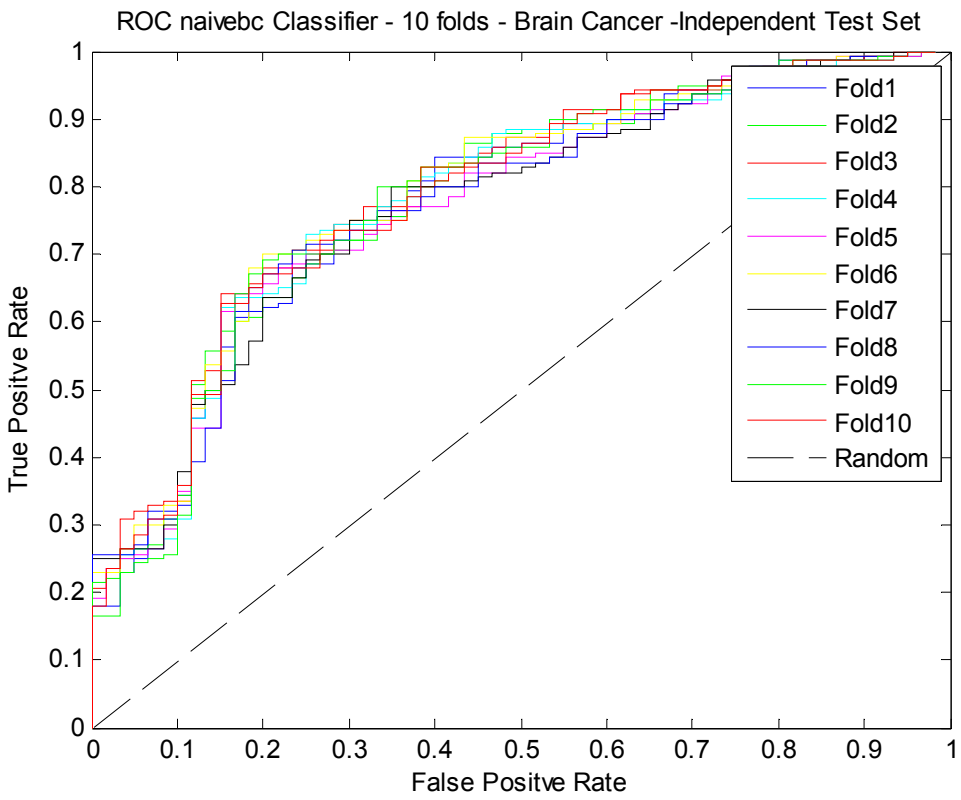
Παρατηρούμε και στις δύο περιπτώσεις αλλά και στους υπόλοιπους classifiers, ότι οι ROC καμπύλες για κάθε fold στο Evaluation Set απέχουν αρκετά μεταξύ τους, ενώ στο Independent Test Set τα δείγματα είναι περισσότερα αναλογικά και οι καμπύλες είναι πιο ομαλές και πιο κοντά η μία με την άλλη. Επίσης το Evaluation Set σε κάθε Fold είναι ένα διαφορετικό δείγμα , ενώ το Independent Test Set είναι κοινό για όλα τα Folds, έτσι οι Roc καμπύλες για το Independent Set σε κάθε Fold τείνουν να συμπίπτουν ή να βρίσκονται πολύ κοντά η μία στην άλλη.

#### 4.2.1.2. Καμπύλες ROC 10 –Folds για Brain Cancer Dataset

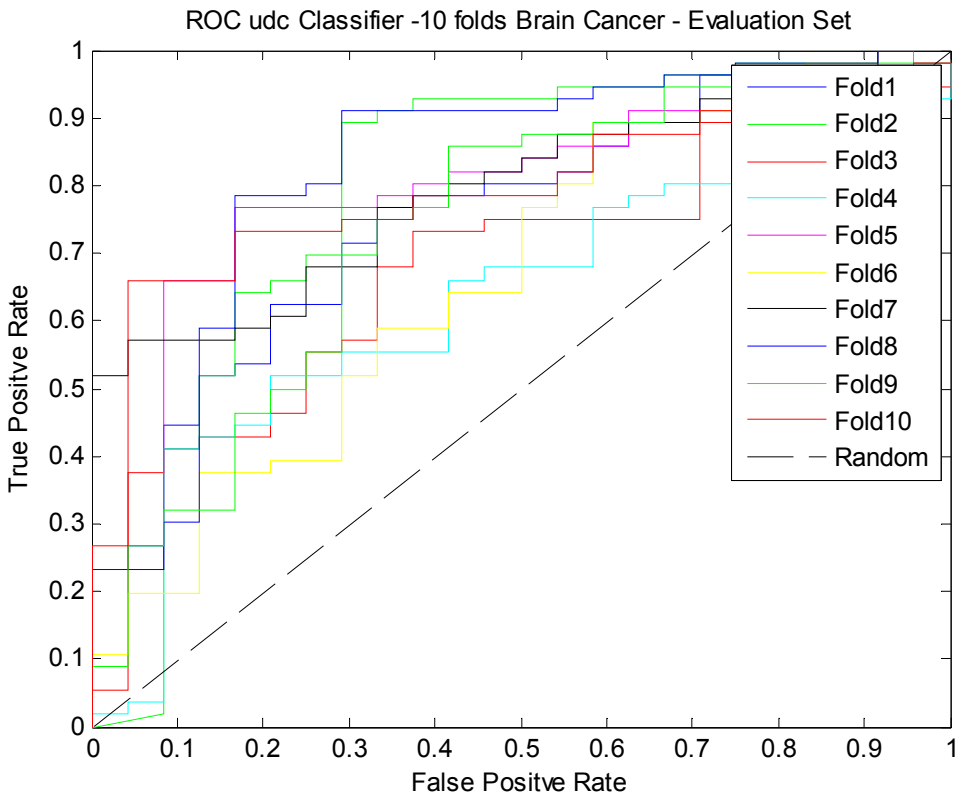
Για το Brain Cancer Dataset επιλέξαμε να δούμε τις ROC και από τα 10 Folds μαζί για τους classifiers Naivebc και Udc. Βλέπουμε παρακάτω δύο διαγράμματα ένα για το Evaluation Set και ένα για το Independent Test Set.



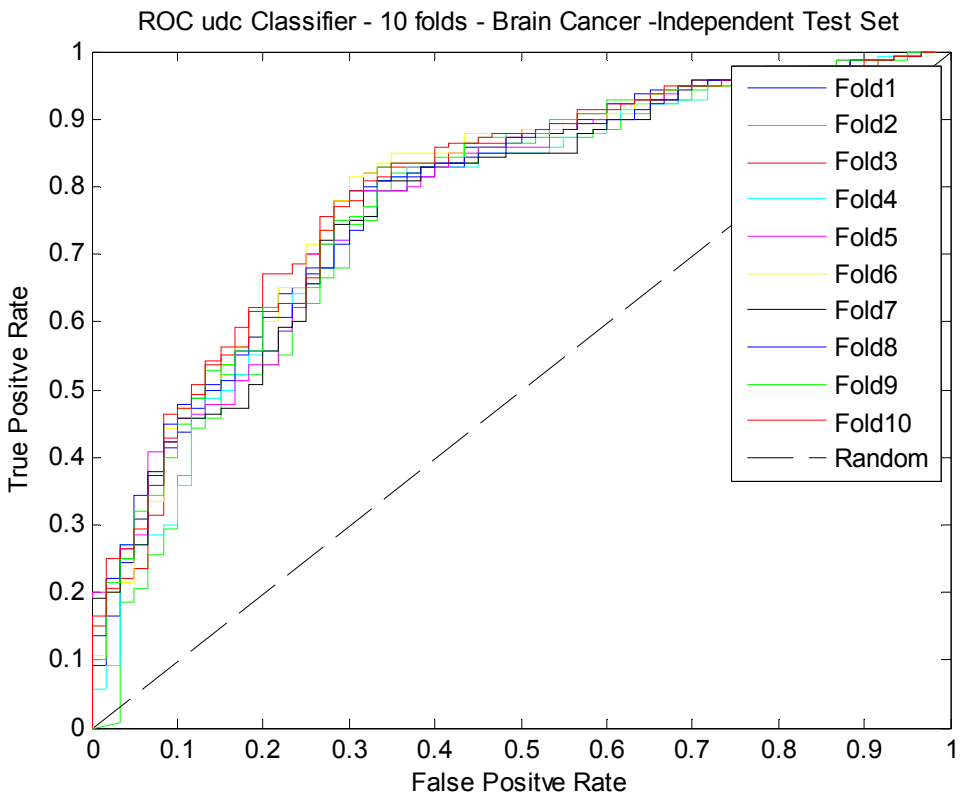

Εικόνα 34 : Brain Tumor Dataset – Roc 10 Folds – Naïvebc Classifier – Evaluation Set




Εικόνα 35 : Brain Tumor Dataset – Roc 10 Folds – Naïvebc Classifier – Independent Set




Εικόνα 36 : Brain Tumor Dataset – Roc 10 Folds – Udc Classifier – Evaluation Set

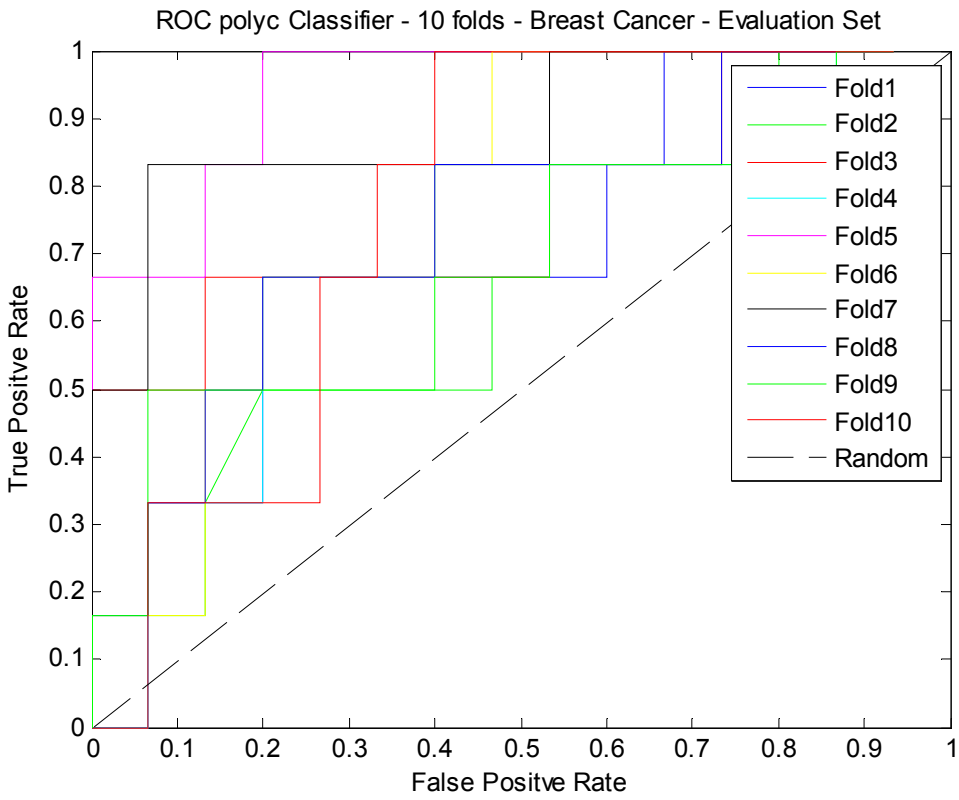



Εικόνα 37 : Brain Tumor Dataset – Roc 10 Folds – Udc Classifier – Independent Set

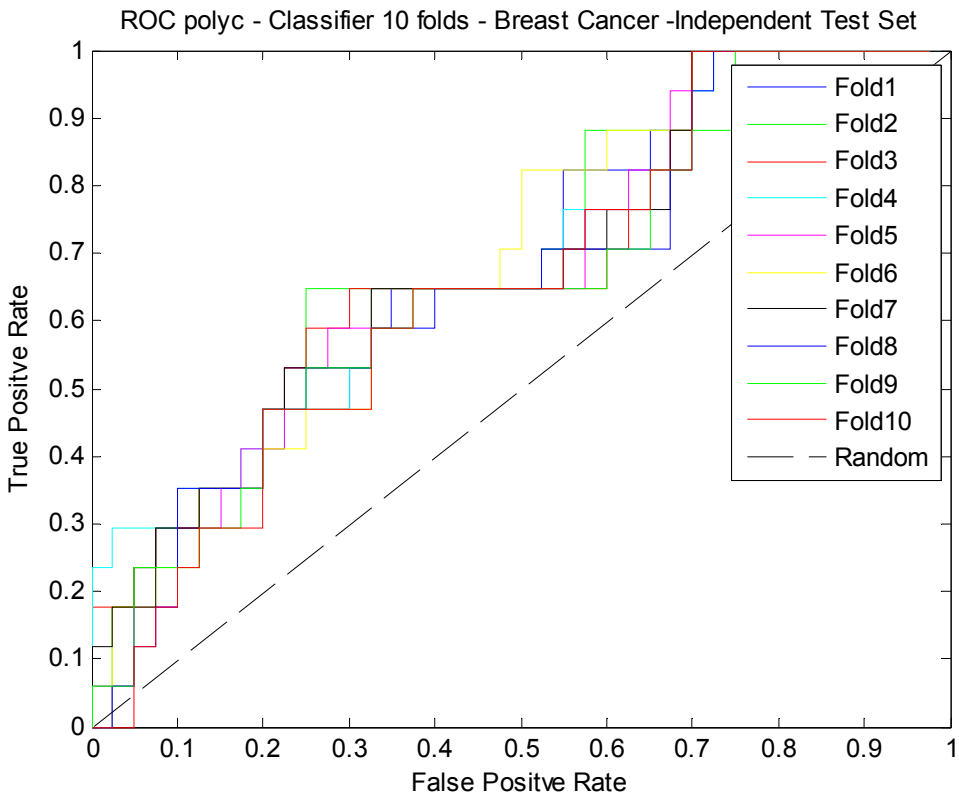
Και σε αυτό το dataset παρατηρούμε και στις δύο περιπτώσεις classifiers αλλά και στους υπόλοιπους classifiers, ότι οι ROC καμπύλες για κάθε fold στο Evaluation Set απέχουν αρκετά μεταξύ τους, ενώ στο Independent Test Set τα δείγματα είναι περισσότερα αναλογικά και οι καμπύλες είναι πιο ομαλές και πιο κοντά η μία με την άλλη. Επίσης το Evaluation Set σε κάθε Fold είναι ένα διαφορετικό δείγμα , ενώ το Independent Test Set είναι κοινό για όλα τα Folds, έτσι οι Roc καμπύλες για το Independent Set σε κάθε Fold τείνουν να συμπίπτουν ή να βρίσκονται πολύ κοντά η μία στην άλλη.

#### 4.2.1.3. Καμπύλες ROC 10 –Folds για Breast Cancer Dataset

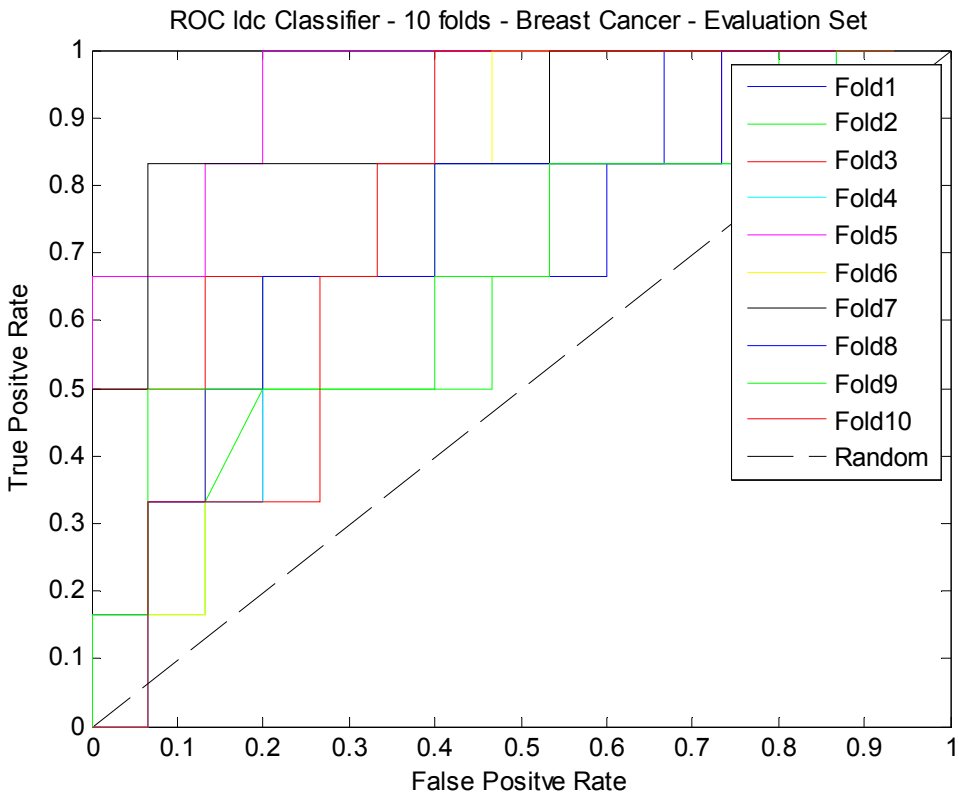
Για το Brain Cancer Dataset επιλέξαμε να δούμε τις ROC και από τα 10 Folds μαζί για τους classifiers Polyc και Ldc. Βλέπουμε παρακάτω δύο διαγράμματα ένα για το Evaluation Set και ένα για το Independent Test Set.



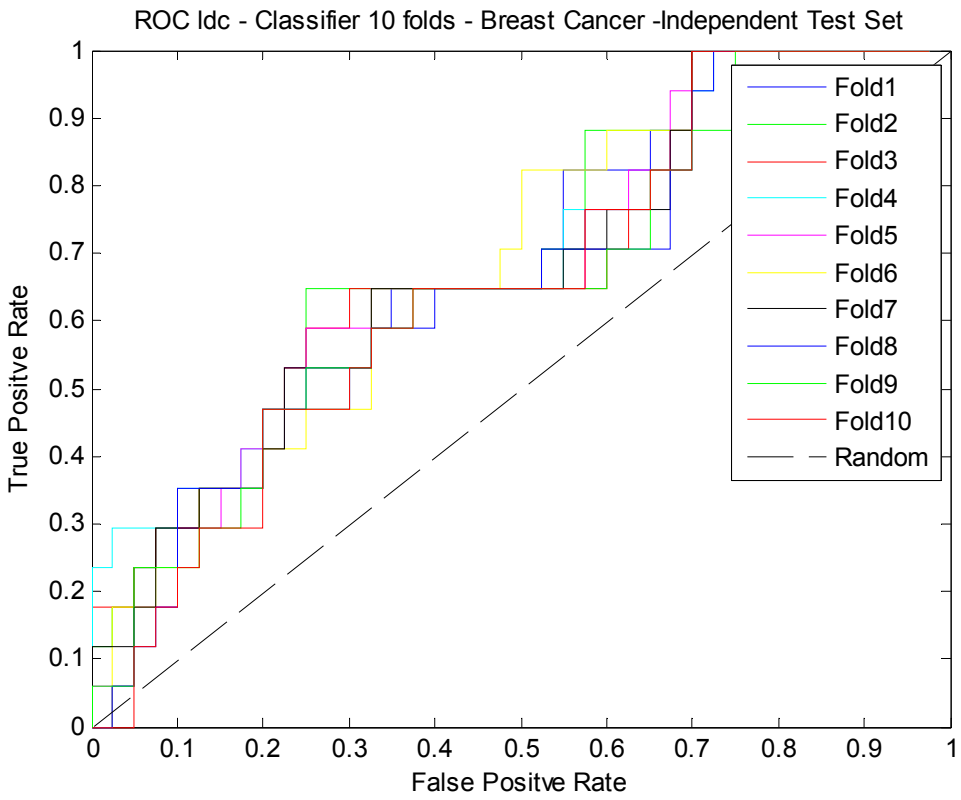

Εικόνα 38 : Breast Tumor Dataset – Roc 10 Folds – Polyc Classifier – Evaluation Set




Εικόνα 39 : Breast Tumor Dataset – Roc 10 Folds – Polyc Classifier – Independent Set




Εικόνα 40 : Breast Tumor Dataset – Roc 10 Folds – Ldc Classifier – Evaluation Set




Εικόνα 41 : Breast Tumor Dataset – Roc 10 Folds – Ldc Classifier – Independent Set

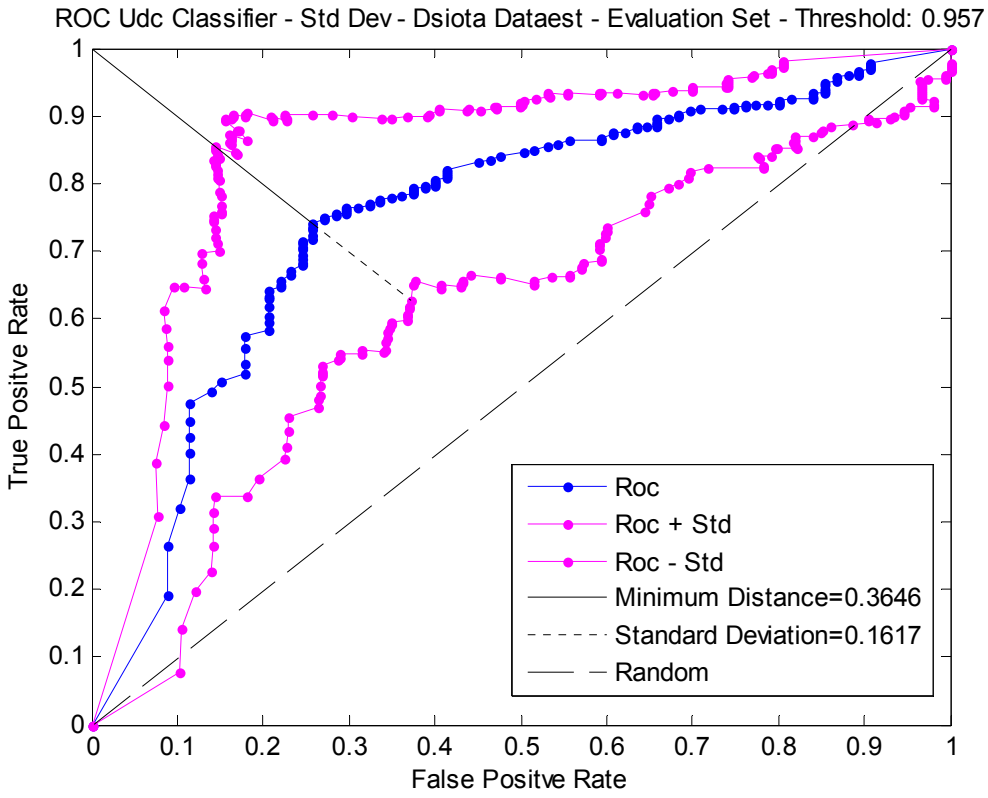
Και σε αυτό το dataset παρατηρούμε και στις δύο περιπτώσεις classifiers αλλά και στους υπόλοιπους classifiers, ότι οι ROC καμπύλες για κάθε fold στο Evaluation Set απέχουν αρκετά μεταξύ τους, ενώ στο Independent Test Set τα δείγματα είναι περισσότερα αναλογικά και οι καμπύλες είναι πιο ομαλές και πιο κοντά η μία με την άλλη. Επίσης το Evaluation Set σε κάθε Fold είναι ένα διαφορετικό δείγμα , ενώ το Independent Test Set είναι κοινό για όλα τα Folds, έτσι οι Roc καμπύλες για το Independent Set σε κάθε Fold τείνουν να συμπίπτουν ή να βρίσκονται πολύ κοντά η μία στην άλλη. Ακόμα, σε αυτό το dataset παρατηρούμε πολύ λιγότερο καλές επιδόσεις στα μοντέλα ταξινόμησης. Το γεγονός αυτό οφείλεται στον πολύ μικρό αριθμό δειγμάτων που έχει το συγκεκριμένο dataset σε σχέση με τα άλλα δύο που εξετάσαμε.

## 4.3. Καμπύλες Μέσης ROC με Standard Deviation

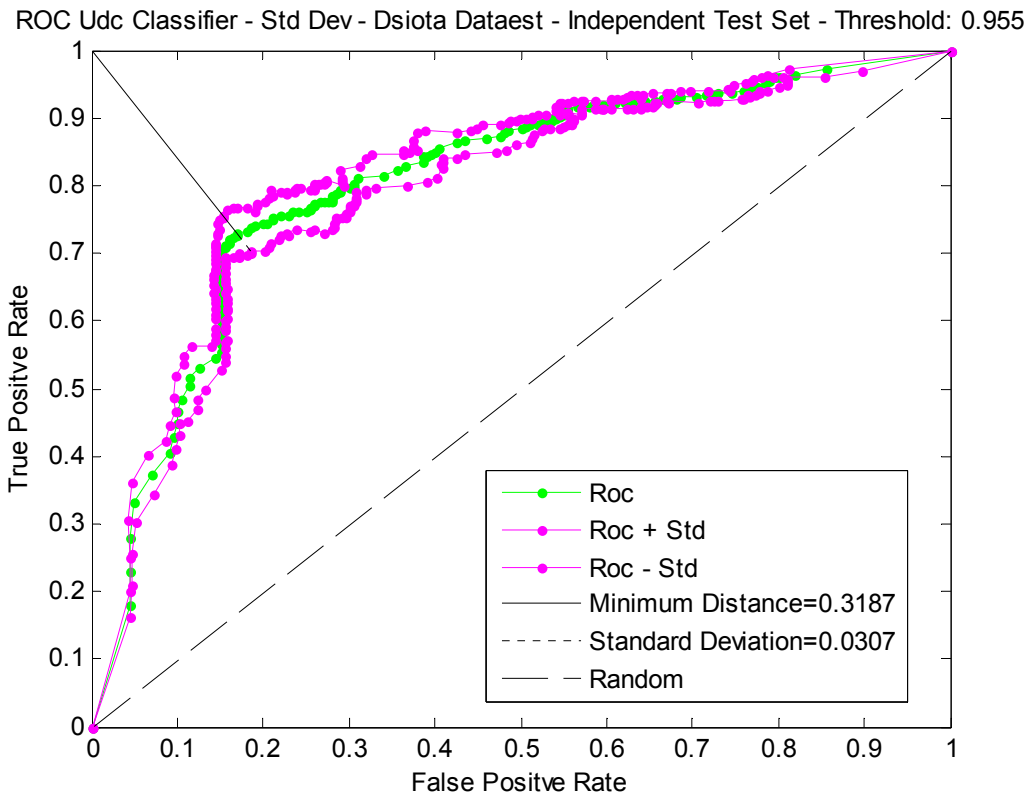
Σε αυτό το σημείο παρουσιάζουμε τα τελικά διαγράμματα της καμπύλης ROC με την τυπική απόκλιση και από αυτά τα διαγράμματα θα πάρουμε την τελική απόφαση για το καλύτερο μοντέλο. Σε κάθε classifier βλέπουμε τη μέση καμπύλη ROC, την

ελάχιστη απόσταση που έχει αυτή από το ιδανικό classification - το σημείο (0,1) - τις καμπύλες της τυπικής απόκλισης της ROC προς το χειρότερο αποτέλεσμα και προς το καλύτερο, καθώς επίσης και την τυπική απόκλιση που έχει η ROC στο σημείο ελάχιστης απόστασης από το ιδανικό classification. Σε κάθε classifier θα δούμε διάγραμμα αυτό για το Evaluation Set αλλά και για το Independent Test Set. Επίσης για κάθε περίπτωση θα δούμε και το κατώφλι βέλτιστης λειτουργίας (Threshold). Επιλέγουμε ενδεικτικά να παρουσιάσουμε τους δύο πρώτους σε απόδοση classifiers από κάθε dataset

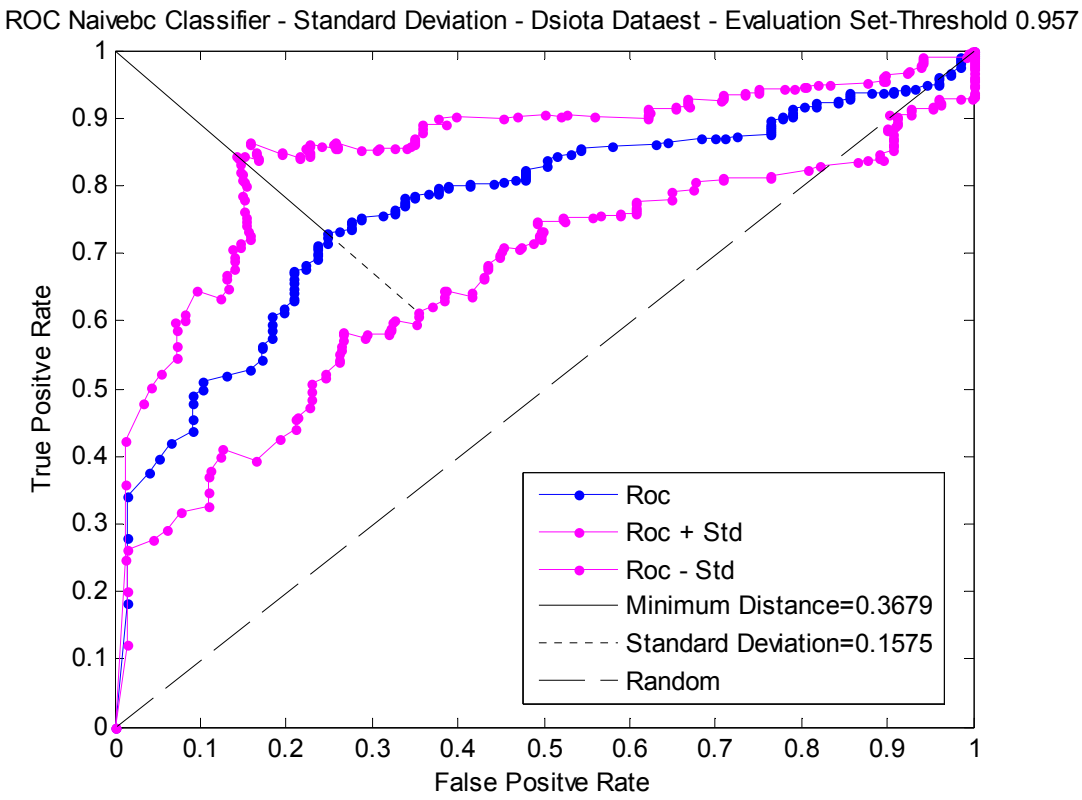
### 4.3.1. Μέση καμπύλη ROC για Ovarian Tumor Dataset με std



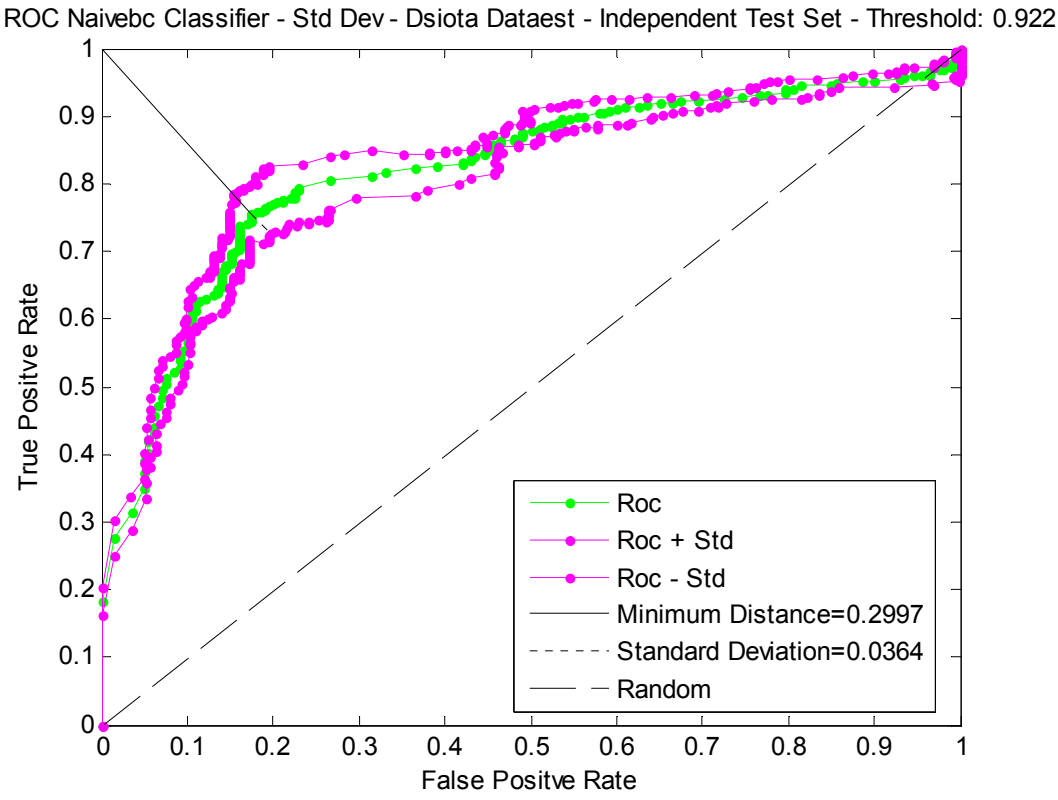

Εικόνα 42 : Ovarian Cancer Dataset – Roc & Standard Deviation – Udc Classifier – Evaluation Set




Εικόνα 43 : Ovarian Cancer Dataset – Roc & Standard Deviation – Udc Classifier – Independent *Set*

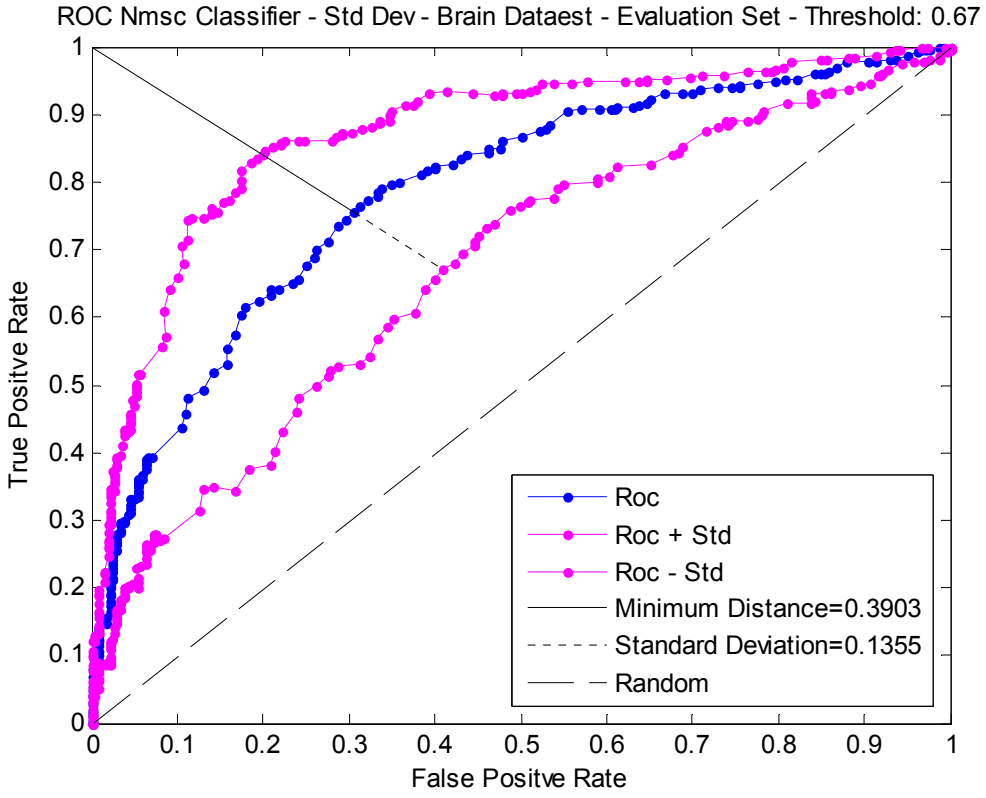



Εικόνα 44 : Ovarian Cancer Dataset – Roc & Standard Deviation – Naivebc Classifier – Evaluation Set

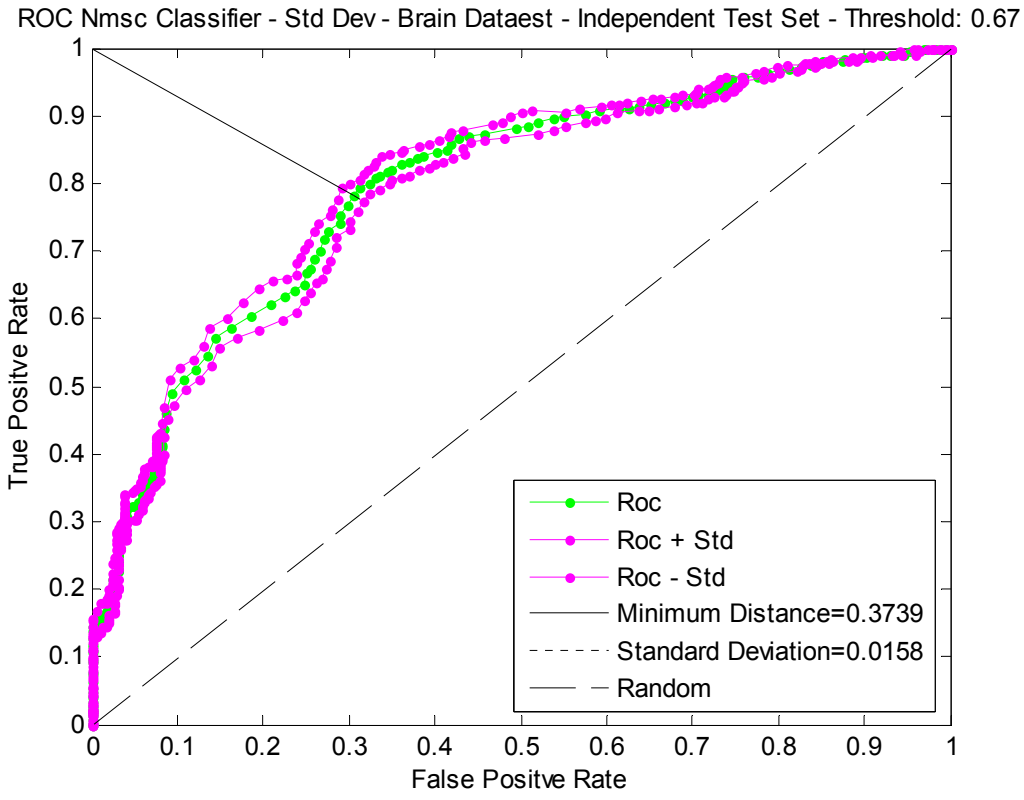



Εικόνα 45 : Ovarian Cancer Dataset – Roc & Standard Deviation – Naivebc Classifier – Independent Set

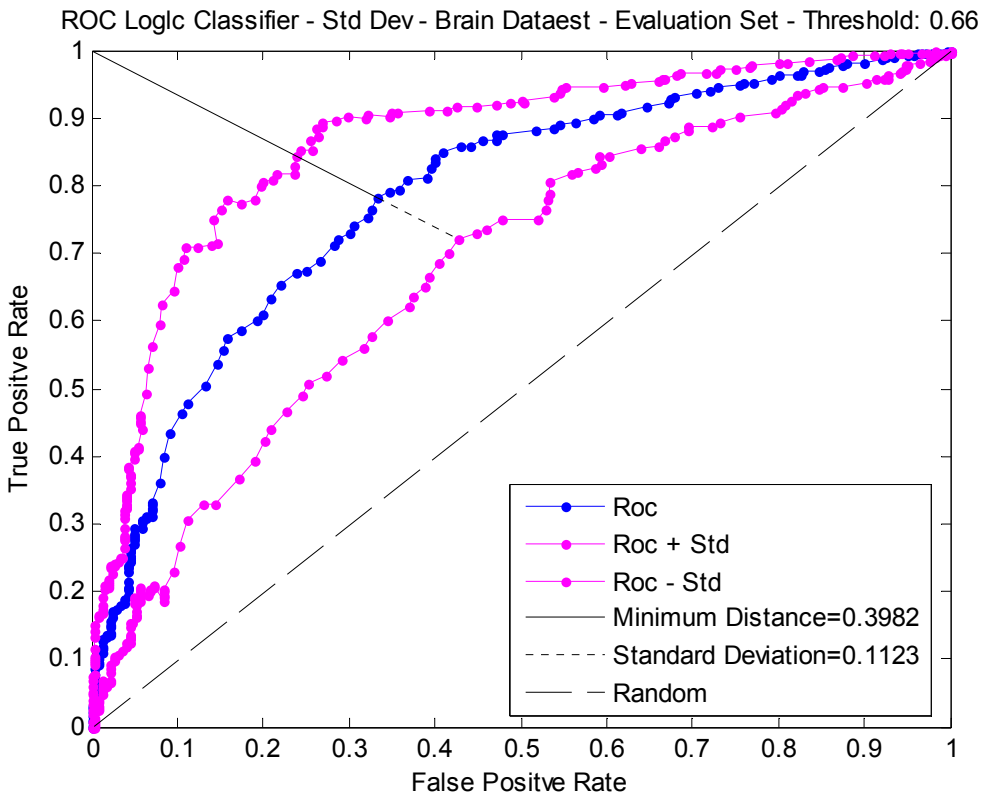
## 4.3.2. Μέση καμπύλη ROC για Brain Tumor Dataset με std



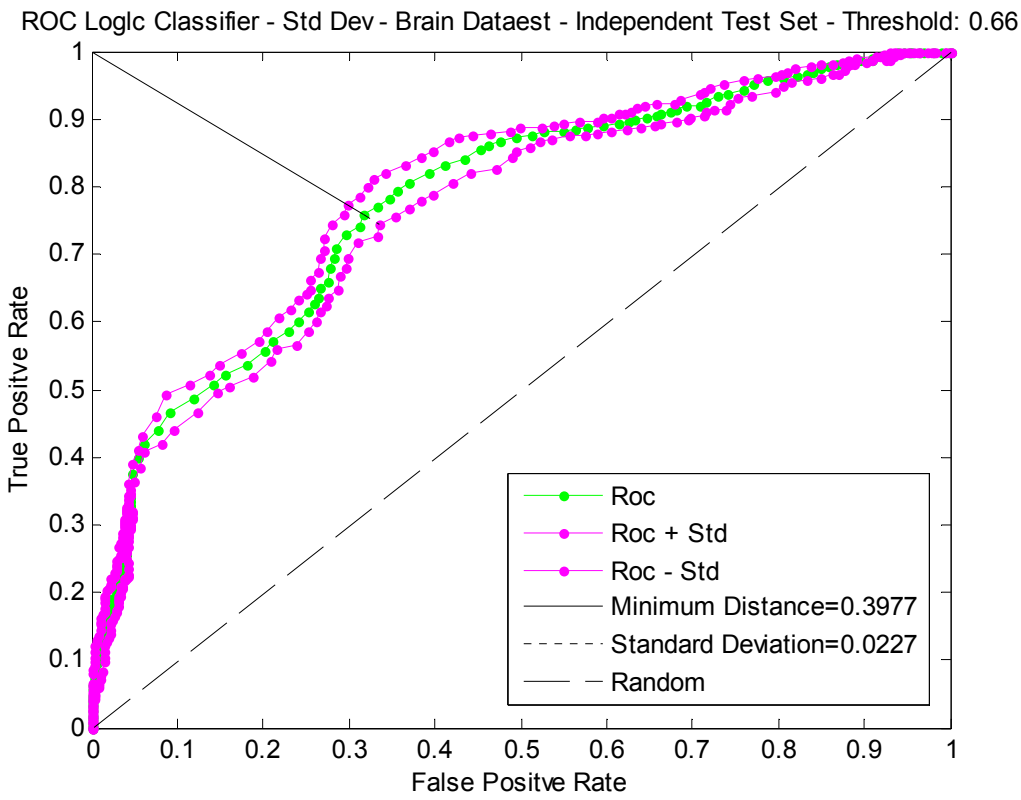

Εικόνα 46 : Brain Cancer Dataset – Roc & Standard Deviation – Nmsc Classifier – Evaluation Set




Εικόνα 47 : Brain Cancer Dataset – Roc & Standard Deviation – Nmsc Classifier – Independent Set

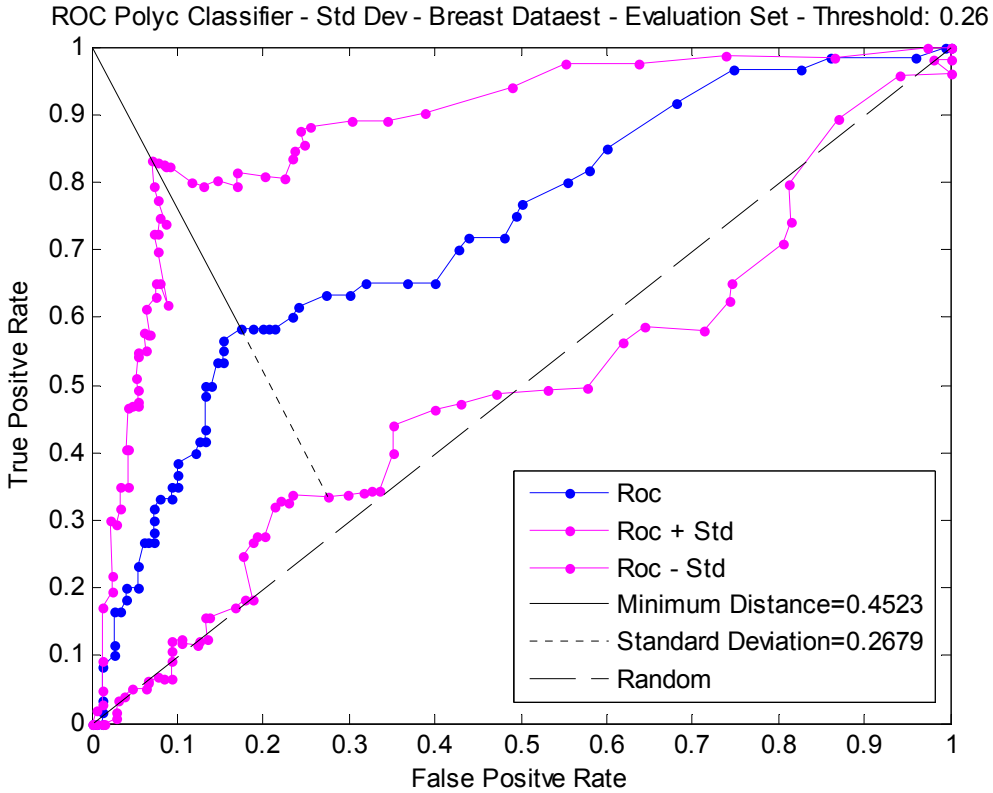



Εικόνα 48 : Brain Cancer Dataset – Roc & Standard Deviation – Loglc Classifier – Evaluation Set

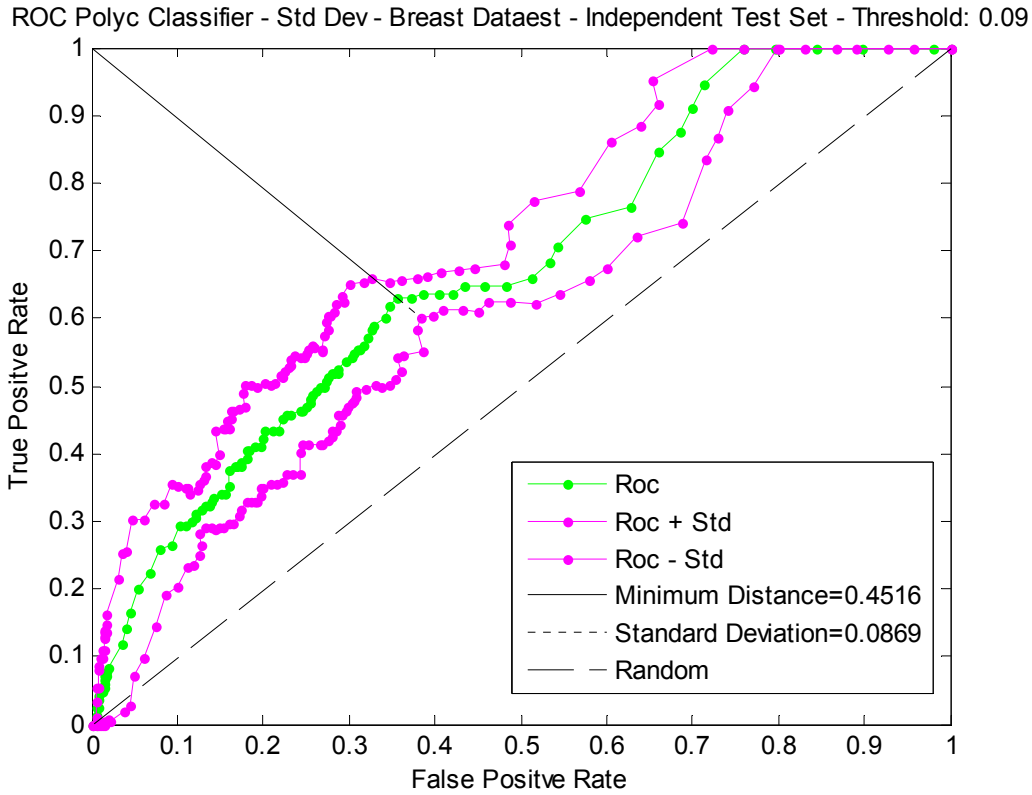



Εικόνα 49 : Brain Cancer Dataset – Roc & Standard Deviation – Loglc Classifier – Independent Set

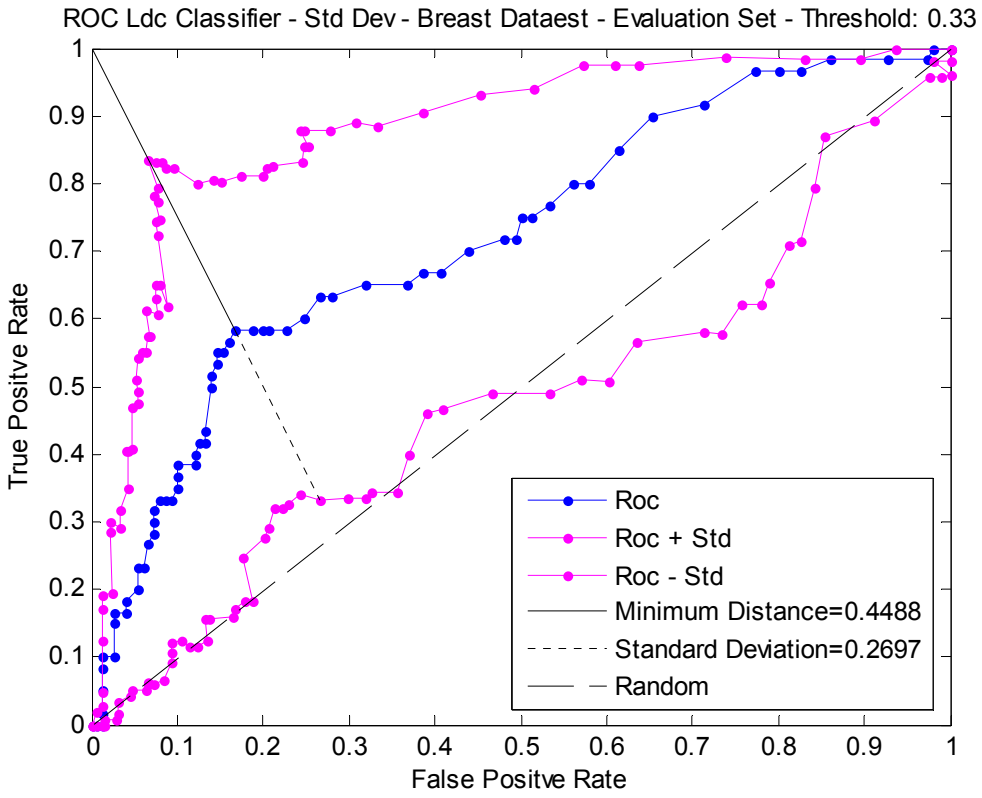
### 4.3.3. Μέση καμπύλη ROC για Breast Cancer Dataset με std



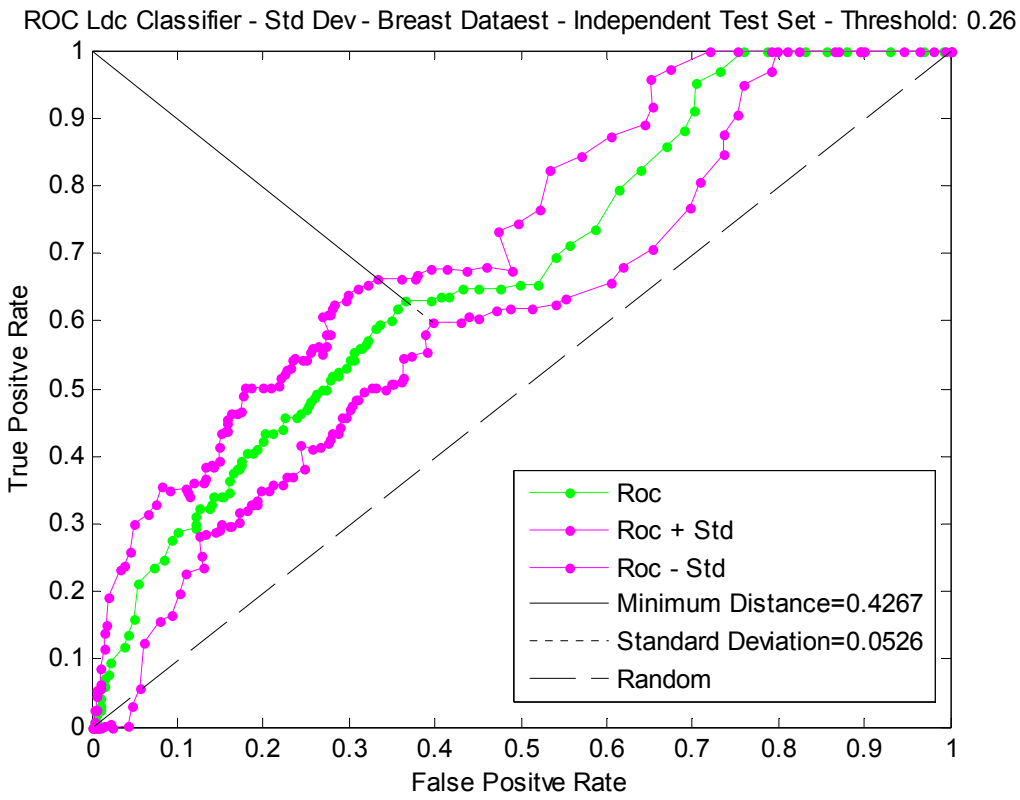

Εικόνα 50 : Breast Cancer Dataset – Roc & Standard Deviation – Polyc Classifier – Evaluation Set



Εικόνα 51 : Breast Cancer Dataset – Roc & Standard Deviation – Polyc Classifier – Independent Set




Εικόνα 52 : Breast Cancer Dataset – Roc & Standard Deviation – Ldc Classifier – Evaluation Set



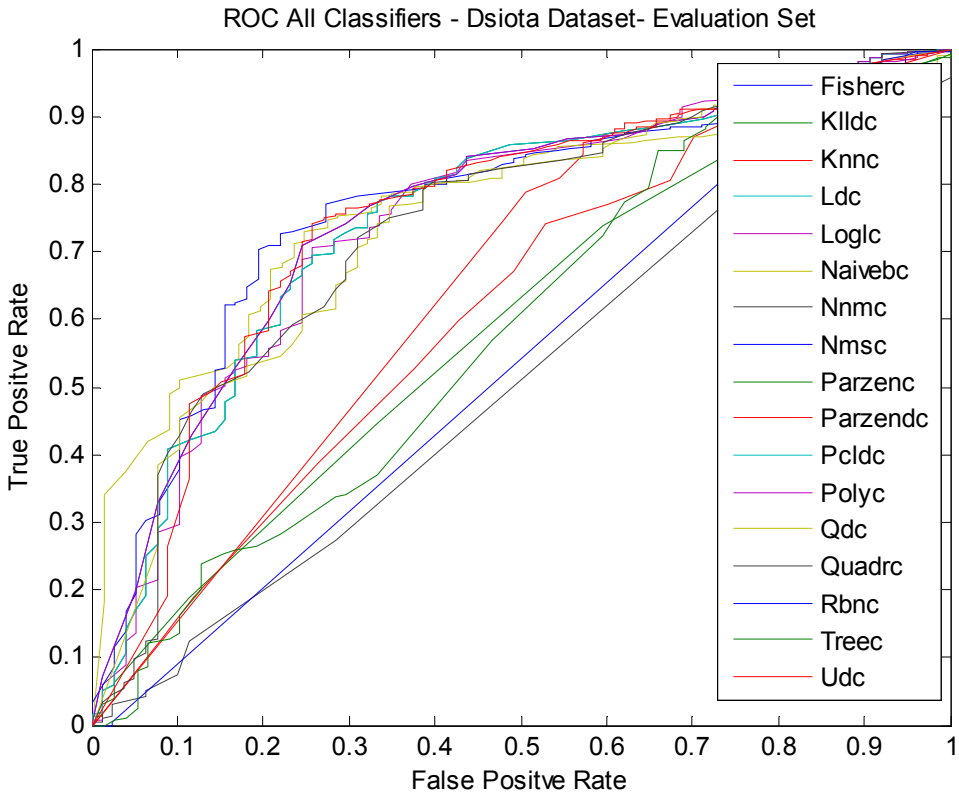

Εικόνα 53 : Breast Cancer Dataset – Roc & Standard Deviation – Ldc Classifier – Independent Set

Παρατηρούμε ότι και στα τρία datasets, καθώς βλέπουμε τη ROC, υπάρχει μια ανομοιομορφία. Κάποιες από τις ROC φαίνεται να έχουν δημιουργηθεί από λιγότερα σημεία, από ότι κάποιες άλλες ή ότι η κατανομή στο χώρο της ROC δεν είναι ομοιόμορφη. Αυτό όμως δεν ισχύει γιατί στις συγκεκριμένες ROC απλά έχουμε μεγάλη επικάλυψη των σημείων κατά τον υπολογισμό της ROC. Δηλαδή σε κάθε κατώφλι, ειδικά στα πολύ γειτονικά, τα σημεία της ROC (True Positive Rate vs. False Positive Rate) είναι τα ίδια.

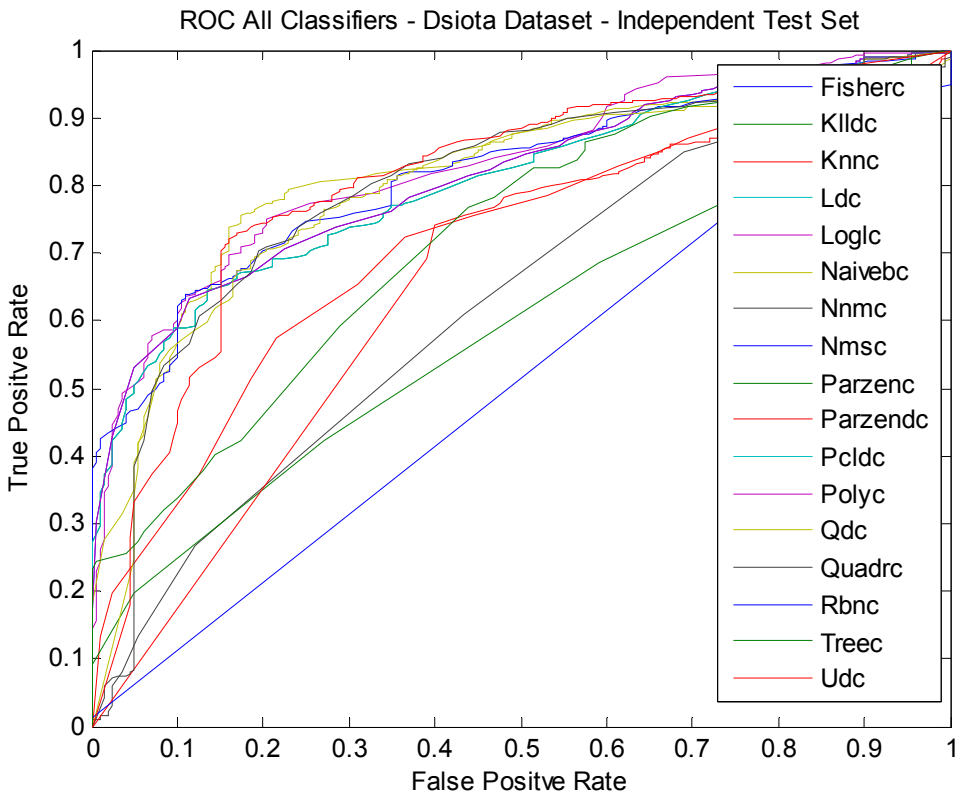
### 4.3.4. Επιλογή του καλύτερου μοντέλου για κάθε dataset

Επιλέγουμε το καλύτερο μοντέλο , τον καλύτερο classifier για κάθε dataset με κριτήριο την μικρότερη απόσταση από το ιδανικό classification , το σημείο (0,1). Το παρακάτω διάγραμμα περιέχει όλες τις ROC καμπύλες για κάθε dataset το Evaluation Set και στο Independent Test Set.

### 4.3.4.1. Ovarian Tumor Dataset Classifier Selection




Εικόνα 54: Ovarian Tumor Dataset – ROC Curve – 17 Classifiers – Evaluation Set

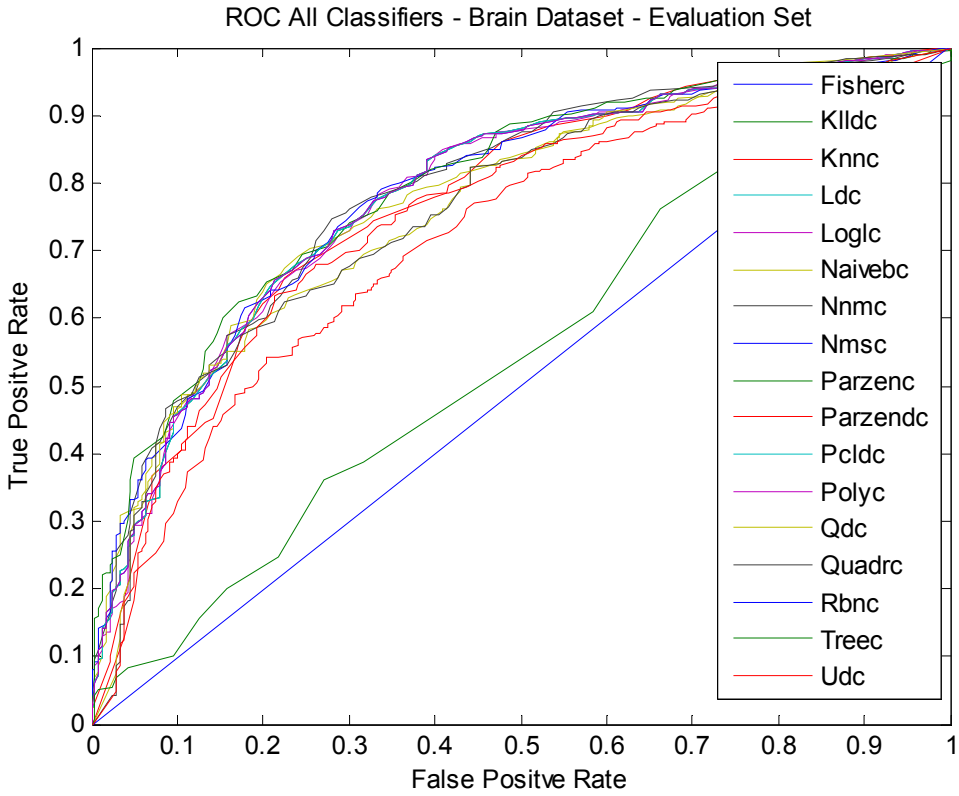



Εικόνα 55 : Ovarian Tumor Dataset – ROC Curve – 17 Classifiers – Independent Set

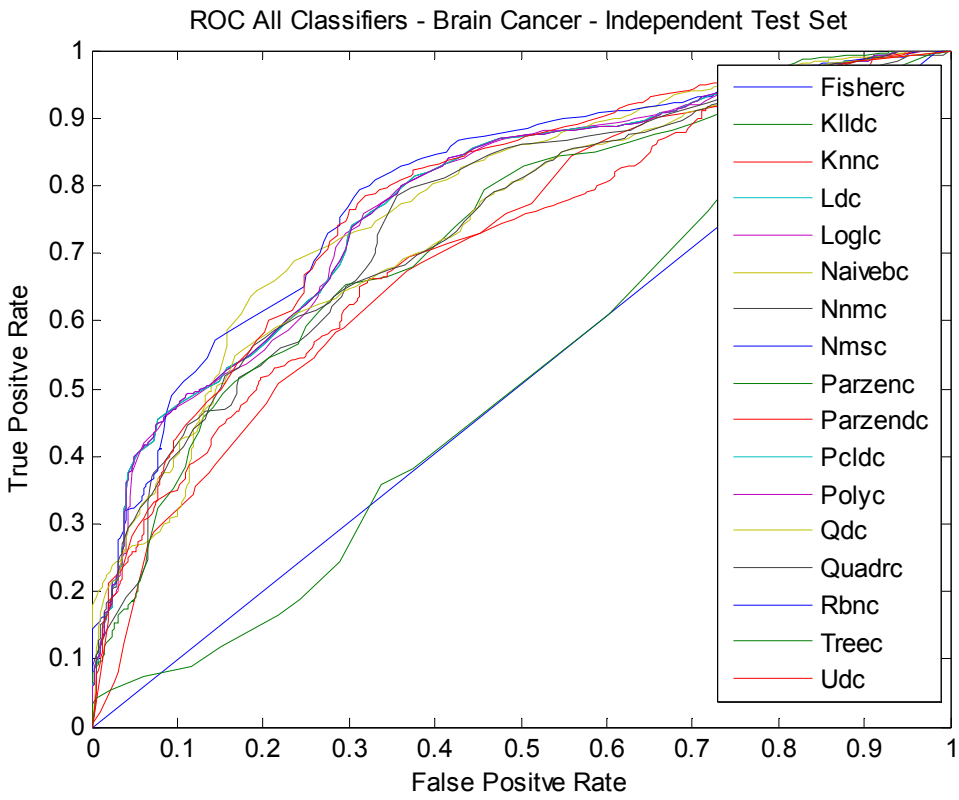
Η ελάχιστη απόσταση όμως από το ιδανικό classification δεν αρκεί. Πρέπει να βρούμε εκείνον το classifiers που όχι μόνο έχει την ελάχιστη απόσταση αλλά έχει και τη μικρότερη δυνατή τυπική απόκλιση.

Για το Ovarian Tumor Dataset μπορούμε να πούμε ότι τα καλύτερα αποτελέσματα τα έχουμε για τον Naivebc classifier με 0.2997 ελάχιστη απόσταση από το ιδανικό classification, std 0.0364 και operating point ~0.922.

### 4.3.4.2. Brain Tumor Dataset Classifier Selection



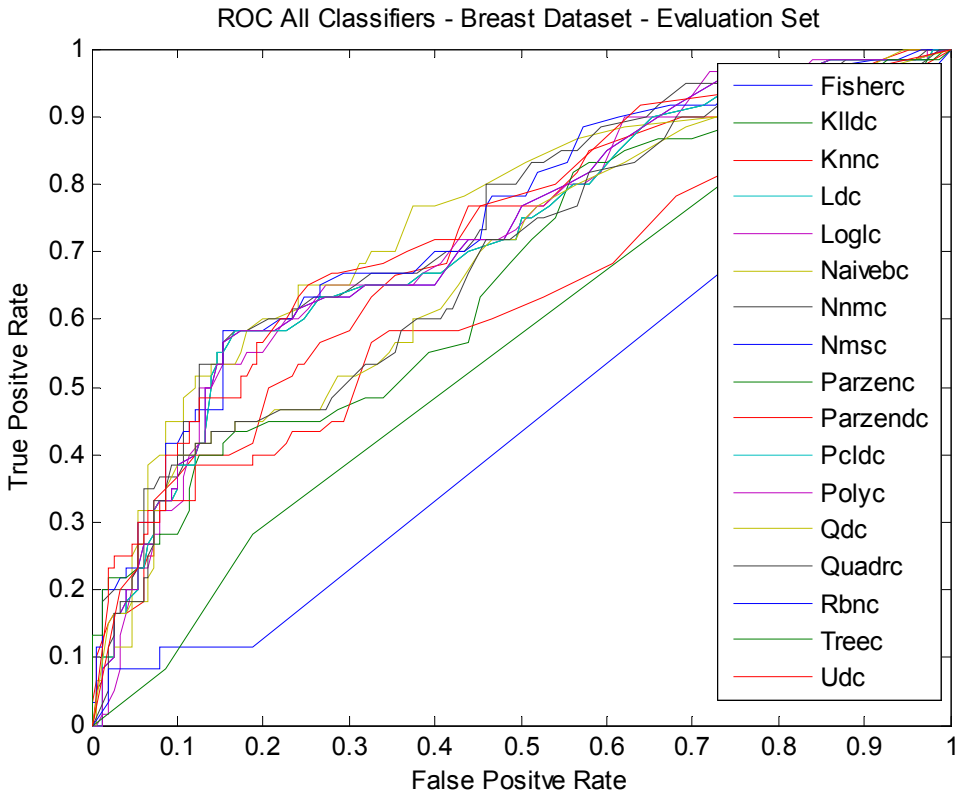

Εικόνα 56 : Brain Tumor Dataset – ROC Curve – 17 Classifiers – Evaluation Set



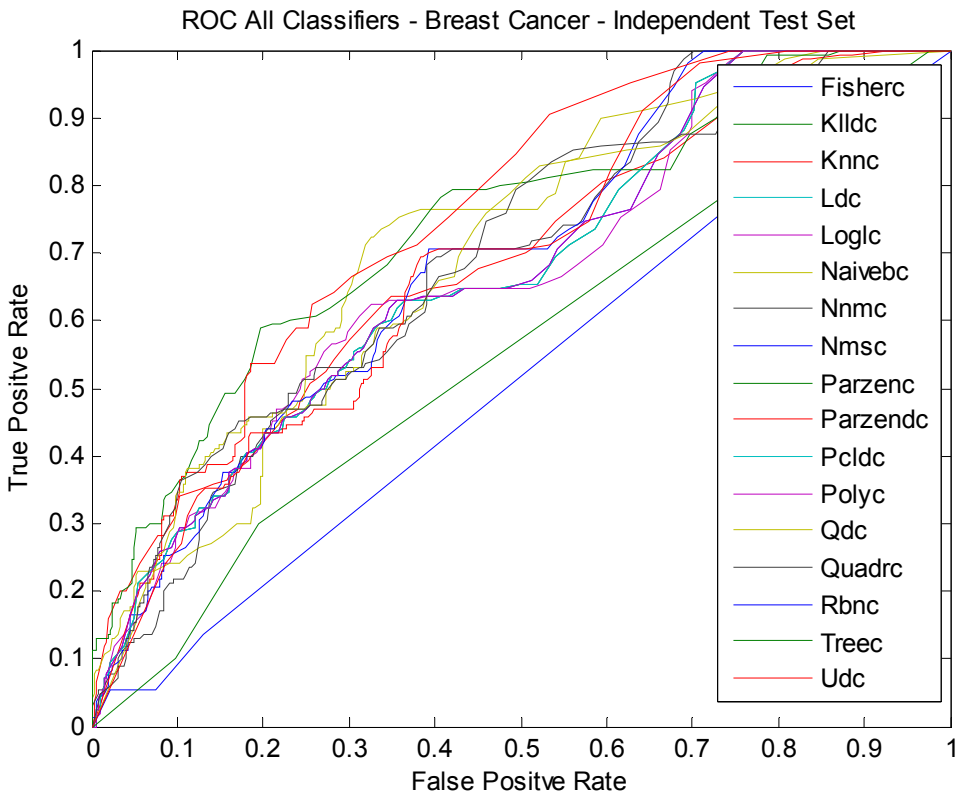

Εικόνα 57 : Brain Tumor Dataset – ROC Curve – 17 Classifiers – Independent Set

Για το Brain Cancer Dataset μπορούμε να πούμε ότι τα καλύτερα αποτελέσματα τα έχουμε για τον Nmsc classifier με 0.3739 ελάχιστη απόσταση από το ιδανικό classification, std 0.0158 και operating point ~0.67.

### 4.3.4.3. Breast Tumor Dataset Classifier Selection




Εικόνα 58 : Breast Tumor Dataset – ROC Curve – 17 Classifiers – Evaluation Set




Εικόνα 59 : Breast Tumor Dataset – ROC Curve – 17 Classifiers – Independent Set

Για το Breast Cancer Dataset μπορούμε να πούμε ότι τα καλύτερα αποτελέσματα τα έχουμε για τον Ldc classifier με 0.4266 ελάχιστη απόσταση από το ιδανικό classification, std 0.0526 και operating point ~0.26.

# 5. Συμπεράσματα

Διαπιστώνουμε τελικά ότι το Error Rate από μόνο του δεν μπορεί να αποτελέσει αξιόπιστη μετρική για το performance ενός classifier. Αυτό γιατί με βάση το Error Rate είχαμε επιλέξει στην αρχή για παράδειγμα για το Ovarian Tumor Dataset τον parzendc classifier ως καλύτερο μοντέλο ταξινόμησης , που είχε το μικρότερο Error Rate, αλλά στην καμπύλη Roc διαπιστώσαμε ένα πάρα πολύ κακό performance με ελάχιστη ευκλείδεια απόσταση από ιδανικό classification 0,4758 ενώ ο καλύτερος classifier για το dataset - ο Naivebc- αυτό είχε 0,2997. Αυτό συνέβη , όπως αναφέρθηκε και παραπάνω, γιατί ο Parzendc παρόλο που είχε πολύ χαμηλό Error Rate είχε επίσης και αποτυχία αναγνώρισης των δειγμάτων της κλάσης #2 100%, δηλαδή δεν αναγνώρισε κανένα δείγμα από την κλάση αυτή. Επομένως σε περιπτώσεις όπως αυτή του Ovarian Tumor Dataset που τα δείγματα που ανήκουν σε μια συγκεκριμένα κλάση διατηρούν μεγάλη πλειοψηφία σε σχέση με αυτά της άλλης κλάσης το Error Rate δεν δίνει απαντήσεις και δε μπορεί να υπολογιστεί σαν μετρική performance του classifier.

Στο Brain Tumor Dataset τα αποτελέσματα Error Rate και καμπύλης ROC φαίνεται να συμφωνούν στην επιλογή βέλτιστου μοντέλου ταξινόμησης, αφού εμφανίζουν και οι δύο μέθοδοι τον Nmsc Classifier σαν καλύτερο. Αυτό, γιατί το συγκεκριμένο dataset έχει σαφώς πιο ομοιόμορφα κατανεμημένα τα δείγματά του μεταξύ των δύο κλάσεων.

Το Breast Cancer Dataset παρόλο που έχει και αυτό ομοιόμορφα κατανεμημένα τα δείγματά του μεταξύ των δύο κλάσεων τα αποτελέσματα δείχνουν ότι τα διαστήματα εμπιστοσύνης δεν είναι ενδεικτικά καθώς οι τυπικές αποκλίσεις είναι μεγαλύτερες από ότι στα άλλα datasets. Αυτό εξηγείται από τα γεγονός ότι τα δείγματα του dataset αυτού και κατά συνέπεια το σετ εκπαίδευσης ήταν πολύ μικρότερο σε σχέση με τα άλλα dataset.

Οι Parzendc, Parzenc, Treec, Knnc, Rbnc, Quadrc, είναι οι classifiers οι οποίοι αποδείχτηκαν οι πιο αναξιόπιστοι σαν μοντέλα ταξινόμησης για τα συγκεκριμένα datasets. Οι καμπύλες ROC τους και στα τρία datasets εφάπτονταν της ευθείας y=x που αντιπροσωπεύει την τυχαία ταξινόμηση. Επομένως γενικά οι οικογένειες

ταξινομητών K-Nearest Neighbor, Quadratic, Neural Network, Decision Trees, Parzen βασίζονται σε αλγόριθμους υπολογισμών που δεν έχουν ικανοποιητικά αποτελέσματα στα συγκεκριμένα δεδομένα.

Από την άλλη μεριά οι classifiers οι οποίοι μας έδωσαν τα καλύτερα αποτελέσματα και στα τρία datasets ήταν οι γραμμικοί, Normal Density, Logistic Regreation, Naïve Bayes, Polynomial. Συγκεκριμένα οι classifiers που είχαν ένα σταθερά ικανοποιητικό αποτέλεσμα και στα τρία datasets ήταν οι Naivebc, Nmsc, Klldc, Udc, Ldc, Polyc. Αν τώρα αναζητήσουμε έναν classifier ο οποίος να έχει πολύ ικανοποιητικά αποτελέσματα και στα τρία datasets και να μπορεί να χρησιμοποιηθεί με αξιοπιστία αυτός θα είναι ο Naivebc, ο οποίος διατηρεί σταθερά και σταθερά μικρή απόσταση από το ιδανικό classification και στα τρία datasets.

# 6. Μελλοντική Εργασία

Το Classifier Fusion, μπορεί να αποτελέσει μας δώσει ακόμα καλύτερα αποτελέσματα στις επιδόσεις και στην τελική αξιολόγηση. Εφόσον το είδαμε ότι έχουμε συγκεκριμένες κατηγορίες ταξινομητών οι οποίες έχουν καλό performance, με το Classifier Fusion θα μπορούμε  να πετύχουμε ίσως ακόμα καλύτερα πιο καλό performance και πιο μικρή απόκλιση στα αποτελέσματα.

Το Feature Selection για κάθε dataset επίσης μπορεί να βελτιώσει ακόμα περισσότερο το performance. Αφαιρώντας δηλαδή τα χαρακτηριστικά που επηρεάζουν αρνητικά και οδηγούν τον classifier σε λανθασμένες αποφάσεις, θα έχουμε ακόμα καλύτερα αποτελέσματα.

# 7. Αναφορές

[1] Barnard, C.M. van der Walt and E. "Data characteristics that determine classifier performance," SAIEE Africa Research Journal, Vol 98 (3), pp 87-93, September 2007.

[2] C.M. van der Walt, "Data measures that characterise classification problems," Master's dissertation, Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa, February 2008.

[3] S. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica Journal 31 (2007) 249-268 (http://www.informatica.si/PDF/31-3/11_Kotsiantis%20-%20Supervised%20Machine%20Learning%20-%20A%20Review%20of...pdf).

[4] Bast RC, Brewer M, Zou C, et al (2007). "Prevention and early detection of ovarian cancer: mission impossible?". Recent Results Cancer Res. 174: 91–100. PMID 17302189.

[5] Kosary, Carol L. (2007), "Chapter 16: Cancers of the Ovary", in Ries, LAG; Young, JL; Keel, GE et al., SEER Survival Monograph: Cancer Survival Among Adults: US SEER Program, 1988-2001, Patient and Tumor Characteristics, SEER Program, NIH Pub. No. 07-6215, Bethesda, MD: National Cancer Institute, pp. 133-144

[6] www.braintumor.org

[7] "Breast Cancer: Statistics on Incidence, Survival, and Screening". Imaginis Corporation. 2006. http://imaginis.com/breasthealth/statistics.asp. Retrieved on 2006-10-09.

[8] R.O. Duda, P.E. Hart, D.G. Stork, "Pattern Classification", Wiley, (2001). ISBN 0-471-05669-3

[9] Belur V. Dasarathy, editor (1991) Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques, ISBN 0-8186-8930-7

[10]Nearest-Neighbor Methods in Learning and Vision, edited by Shakhnarovish, Darrell, and Indyk, The MIT Press, 2005, ISBN 0-262-19547-X

[11] Agresti, Alan. (2002). Categorical Data Analysis. New York: Wiley-Interscience. ISBN 0-471-36093-7.

[12] Rish, Irina. (2001). "An empirical study of the naive Bayes classifier". IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. (available online: PDF, PostScript)

[13] Duda, R. and Hart, P. (1973). Pattern Classification and Scene Analysis. John Wiley & Sons. ISBN 0-471-22361-1.

[14] Smith, Murray (1993) Neural Networks for Statistical Modeling, Van Nostrand Reinhold, ISBN 0-442-01310-8

[15] Y. Yuan and M.J. Shaw, Induction of fuzzy decision trees. Fuzzy Sets and Systems 69 (1995), pp. 125–139

[16] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D.M.J. Tax, “A Matlab Toolbox for Pattern Recognition” , February 2004.

[17] Corina Cortes, Mehryar Mohri, “Confidence Intervals for the Area under the ROC Curve”

[18] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27, 861-874.

[19] Hanley, JA; BJ McNeil (1983-09-01). "A method of comparing the areas under receiver operating characteristic curves derived from the same cases". Radiology 148 (3): 839-843. http://radiology.rsnajnls.org/cgi/content/abstract/148/3/839. Retrieved on 2008-12-03.

[20] Dodd, Lori E.; Margaret S. Pepe (2003). "Partial AUC Estimation and Regression". Biometrics 59 (3): 614-623. doi:10.1111/1541-0420.00071. http://www.blackwell-synergy.com/doi/abs/10.1111/1541-0420.00071. Retrieved on 2007-12-18.

[21] Altman DG, Bland JM (1994). "Diagnostic tests. 1: Sensitivity and specificity". BMJ 308 (6943): 1552. PMID 8019315. http://www.bmj.com/cgi/content/full/308/6943/1552.

[22] Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L; International Ovarian Tumor Analysis Group. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. J Clin Oncol 2005; 23: 8794-801

[23] Blake, C.L., & Merz, C.J. (1998). UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. http://www.ics.uci.edu/~mlearn/MLRepository.html

[24] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143. http://citeseer.ist.psu.edu/kohavi95study.html.(Morgan Kaufmann, San Mateo)

[25] Devijver, P. A., and J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, London, 1982

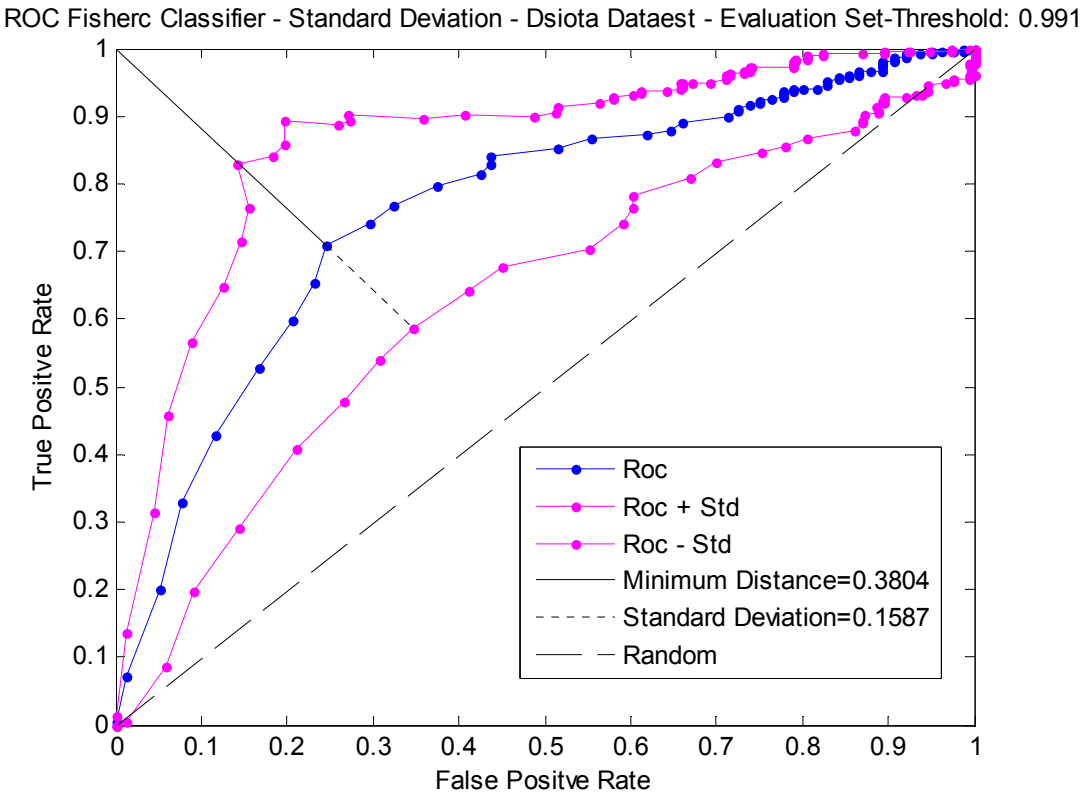
[26] UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/)

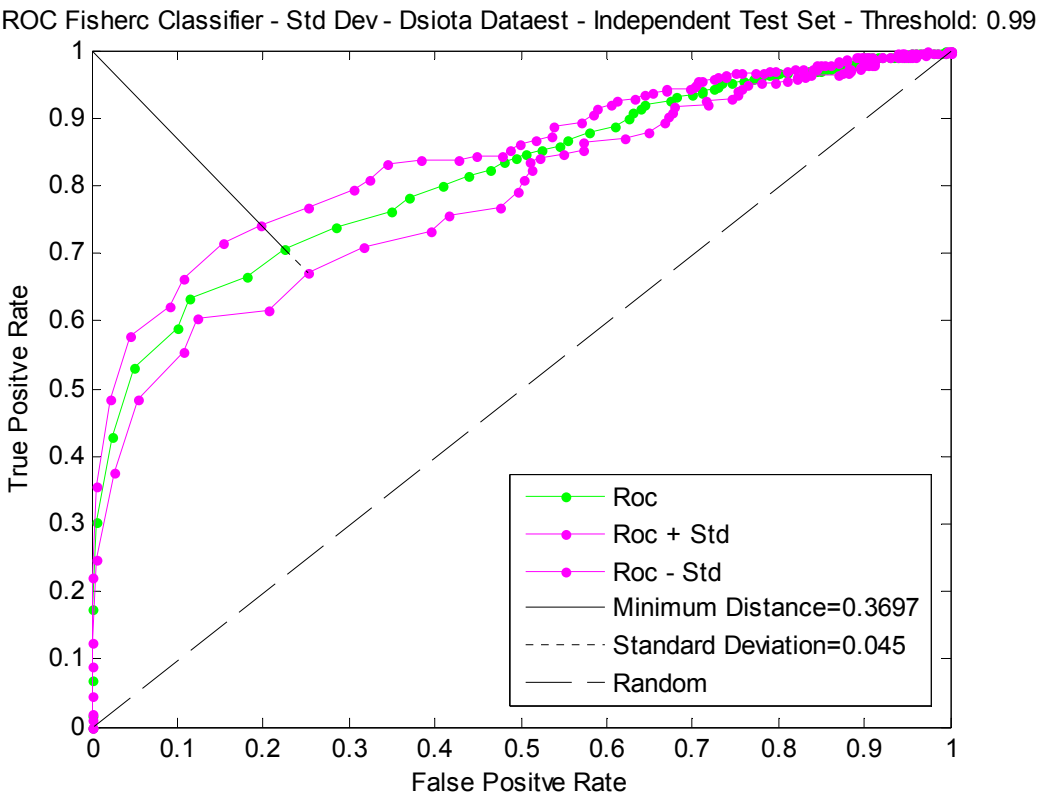
# 8. Παράρτημα

## 8.1. Καμπύλες ROC Με std

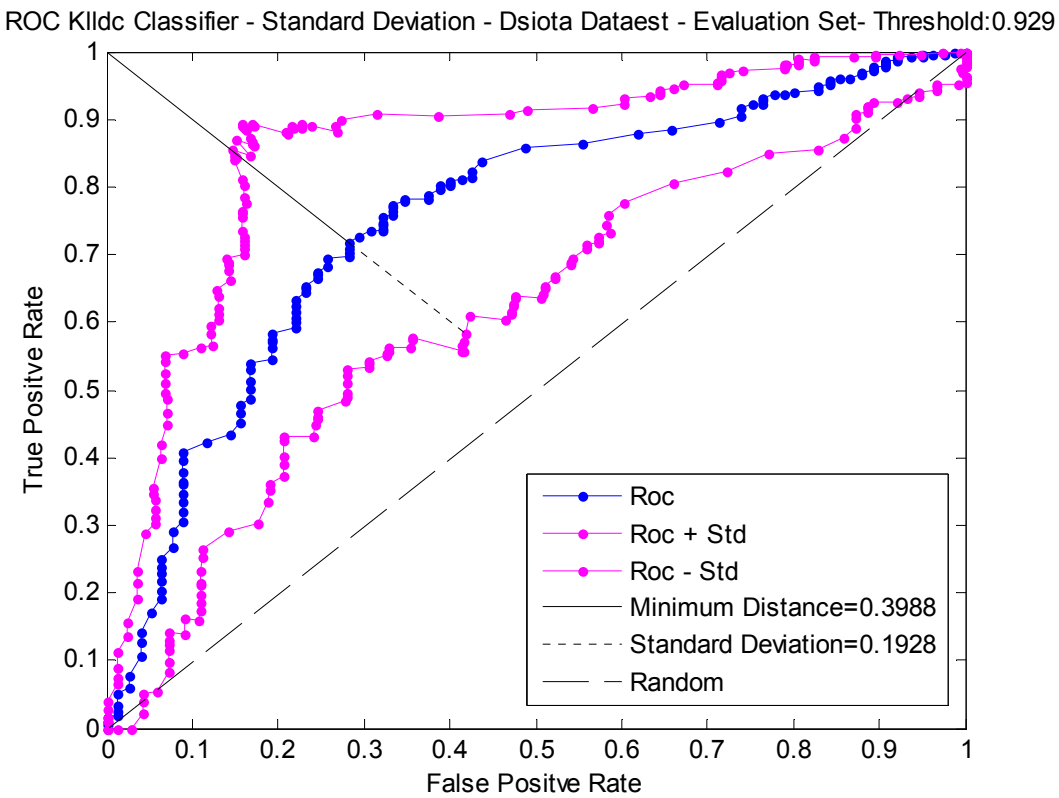
### 8.1.1. Ovarian Tumor Dataset

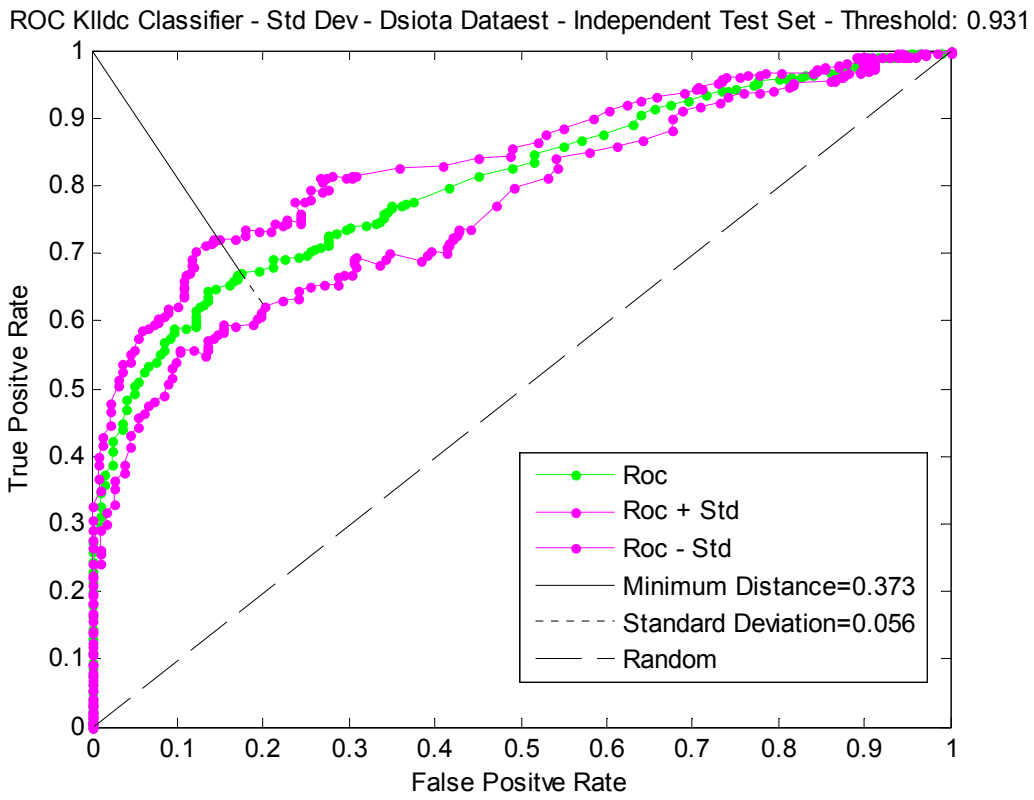
Fisherc



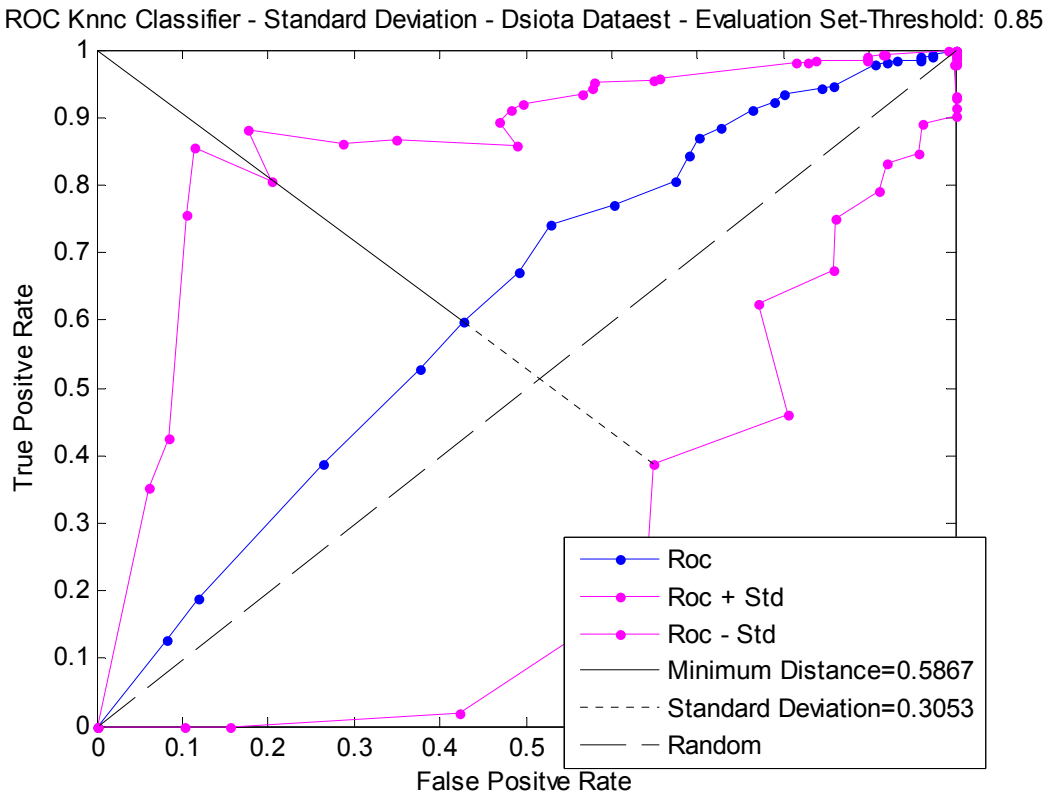

ROC Fisherc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.99
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3697
Standard Deviation=0.045
Random

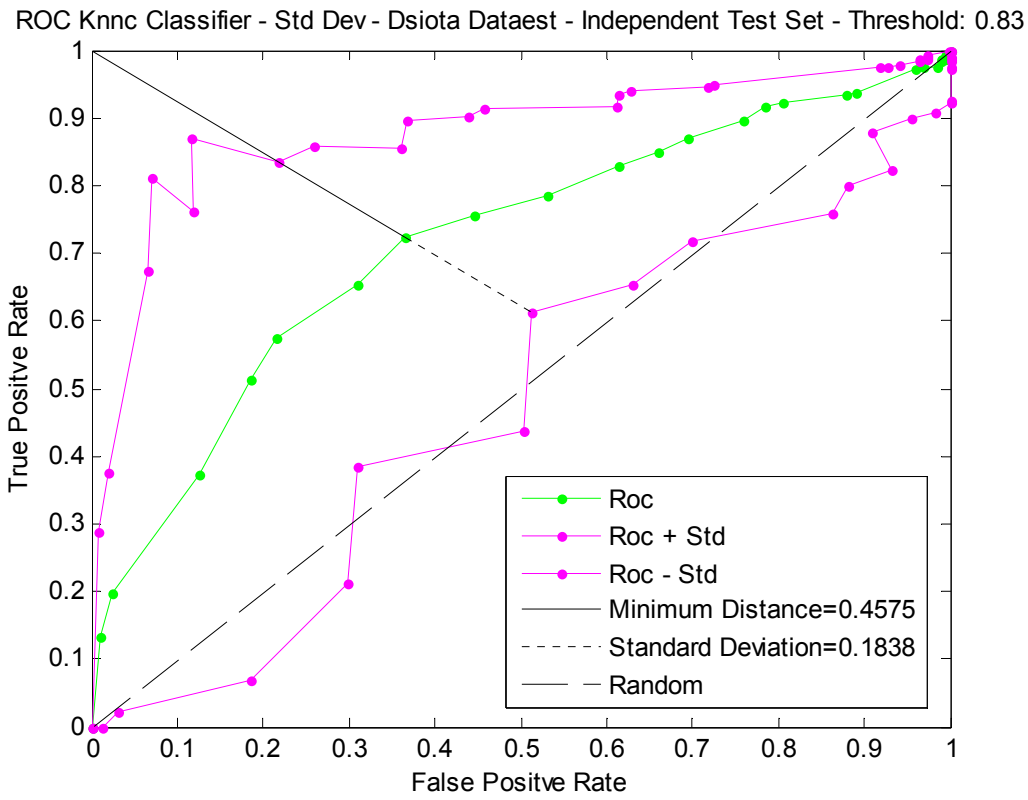
Klldc


ROC Klldc Classifier - Standard Deviation - Dsiota Dataest - Evaluation Set- Threshold:0.929
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3988
Standard Deviation=0.1928
Random

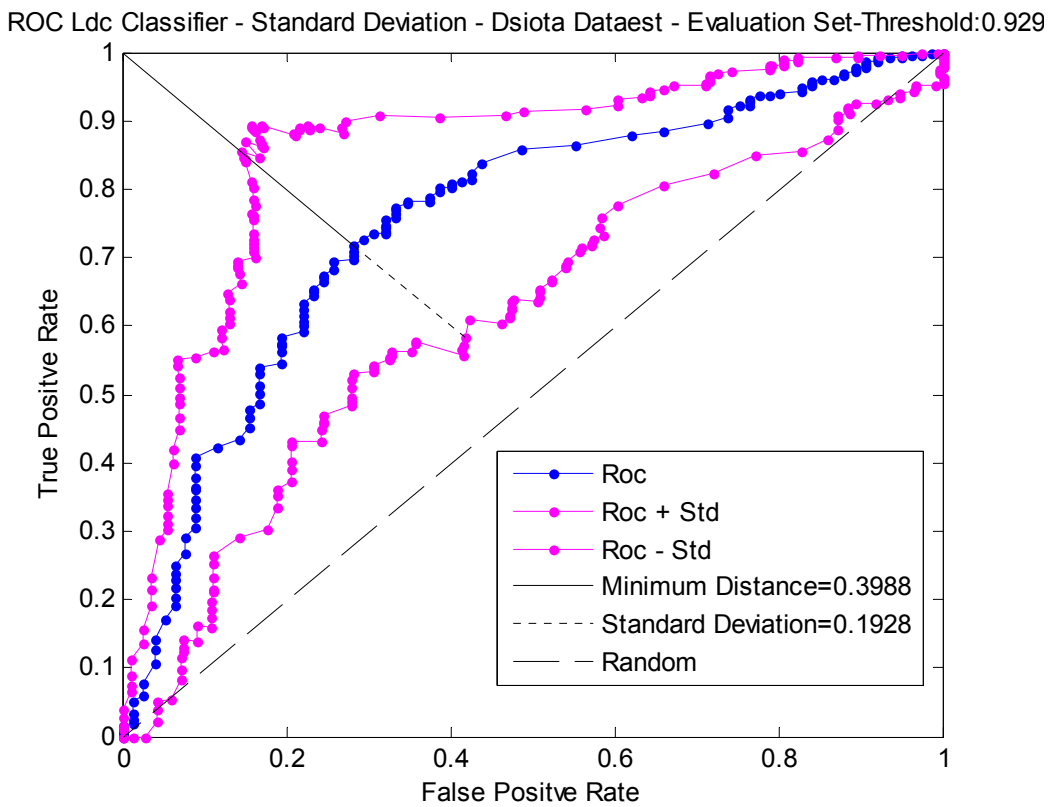

ROC Klldc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.931
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.373
Standard Deviation=0.056
Random

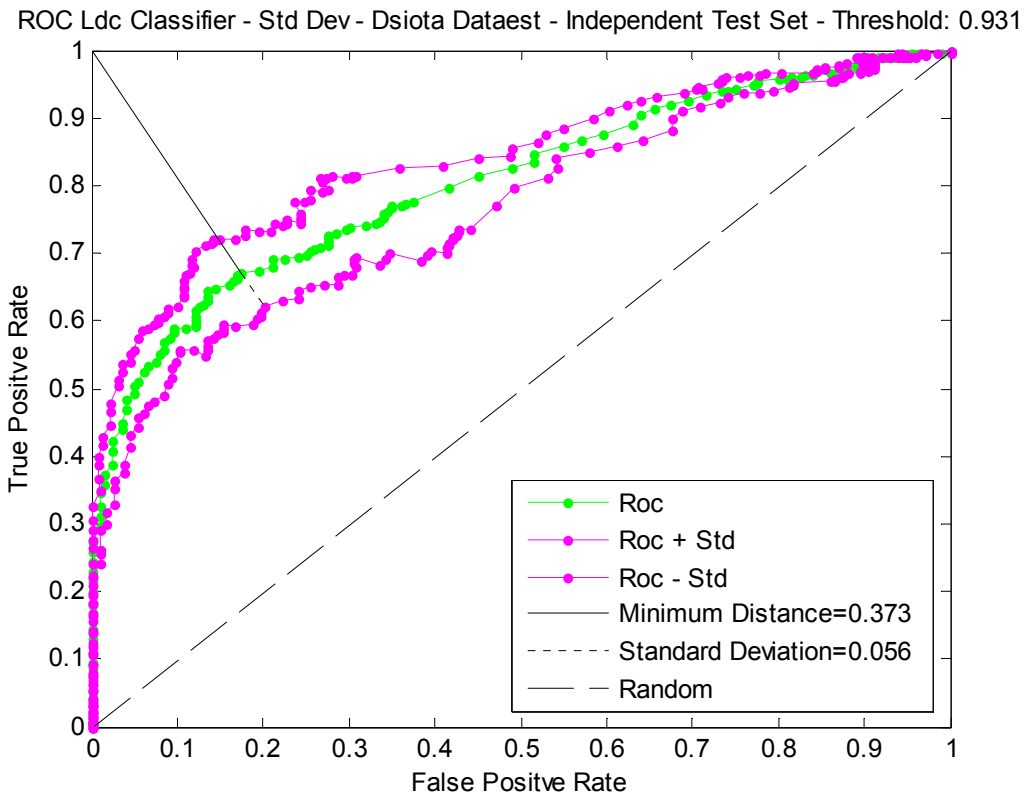
Knnc


ROC Knnc Classifier - Standard Deviation - Dsiota Dataest - Evaluation Set-Threshold: 0.85
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5867
Standard Deviation=0.3053
Random

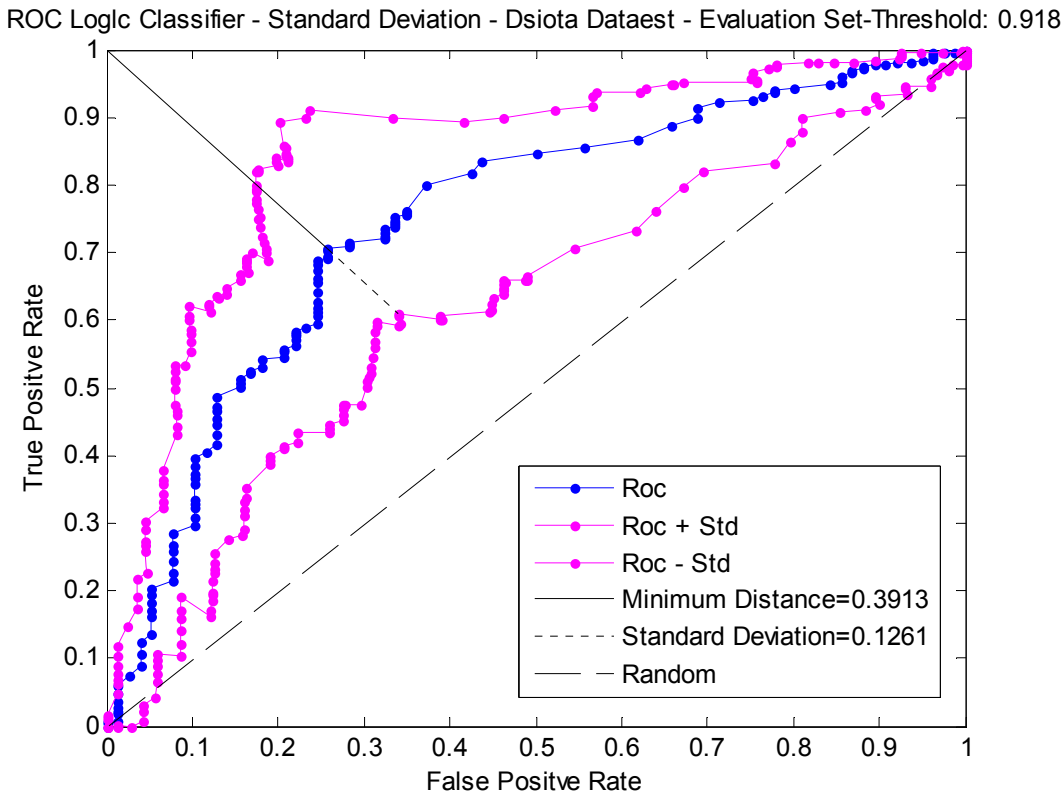

ROC Knnc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.83
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4575
Standard Deviation=0.1838
Random

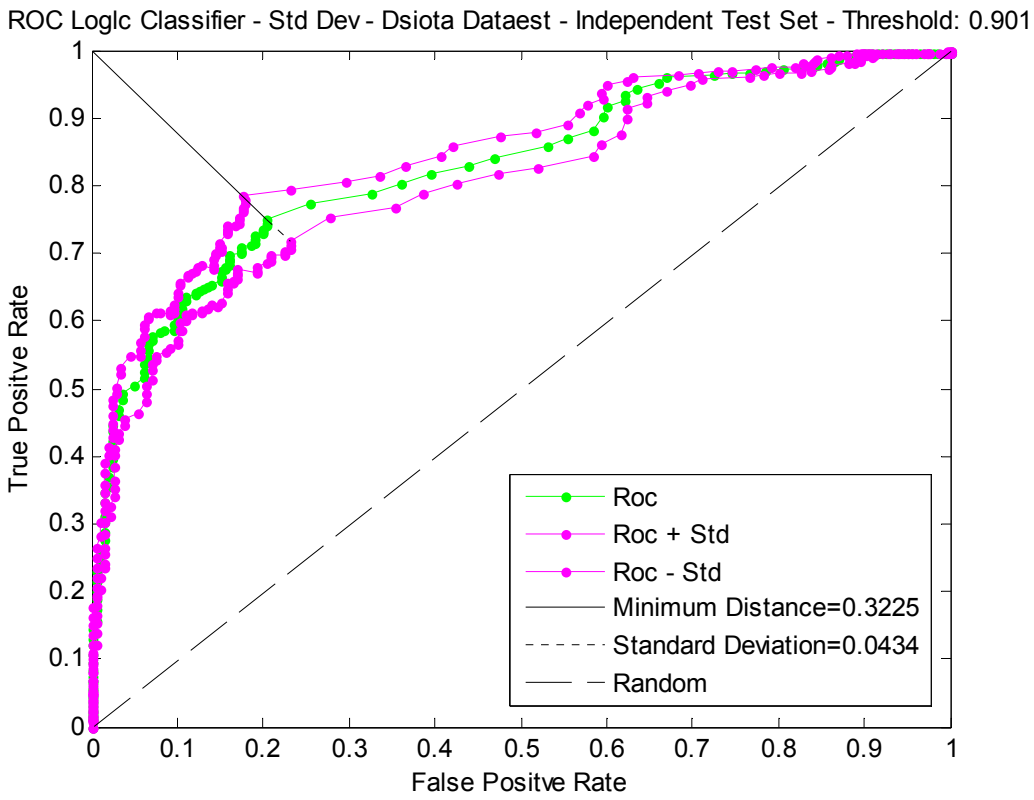
Ldc


ROC Ldc Classifier - Standard Deviation - Dsiota Dataest - Evaluation Set-Threshold:0.929
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3988
Standard Deviation=0.1928
Random

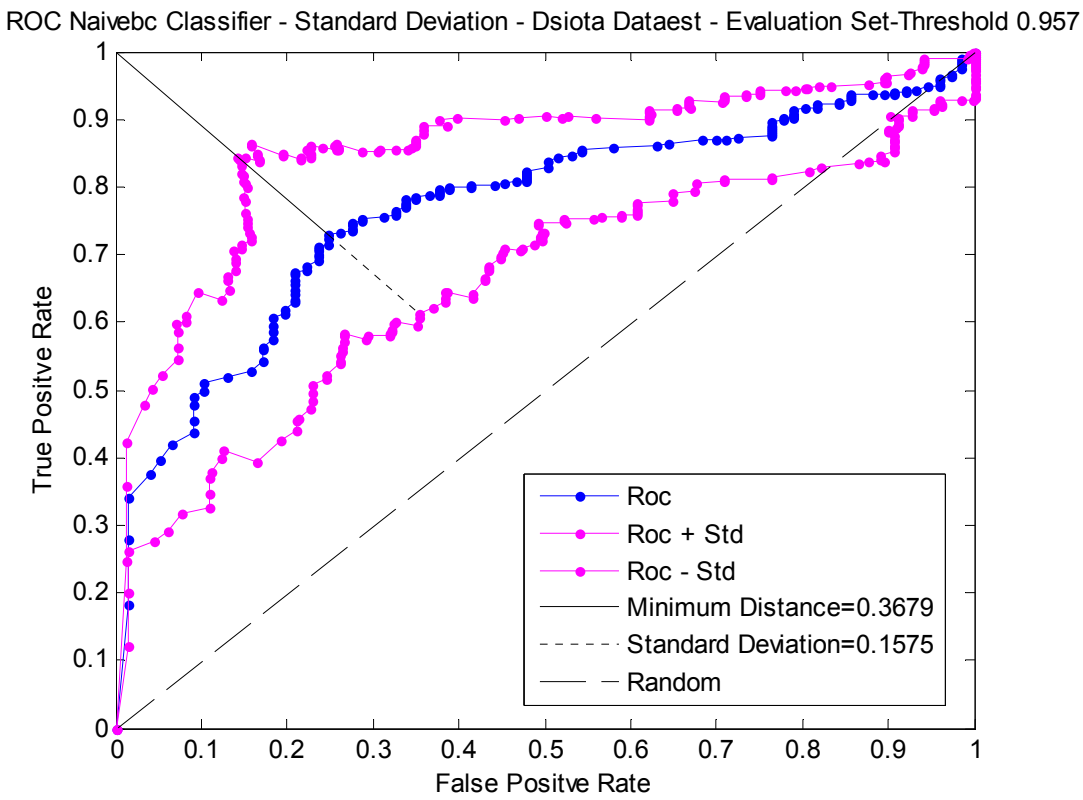

ROC Ldc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.931
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.373
Standard Deviation=0.056
Random

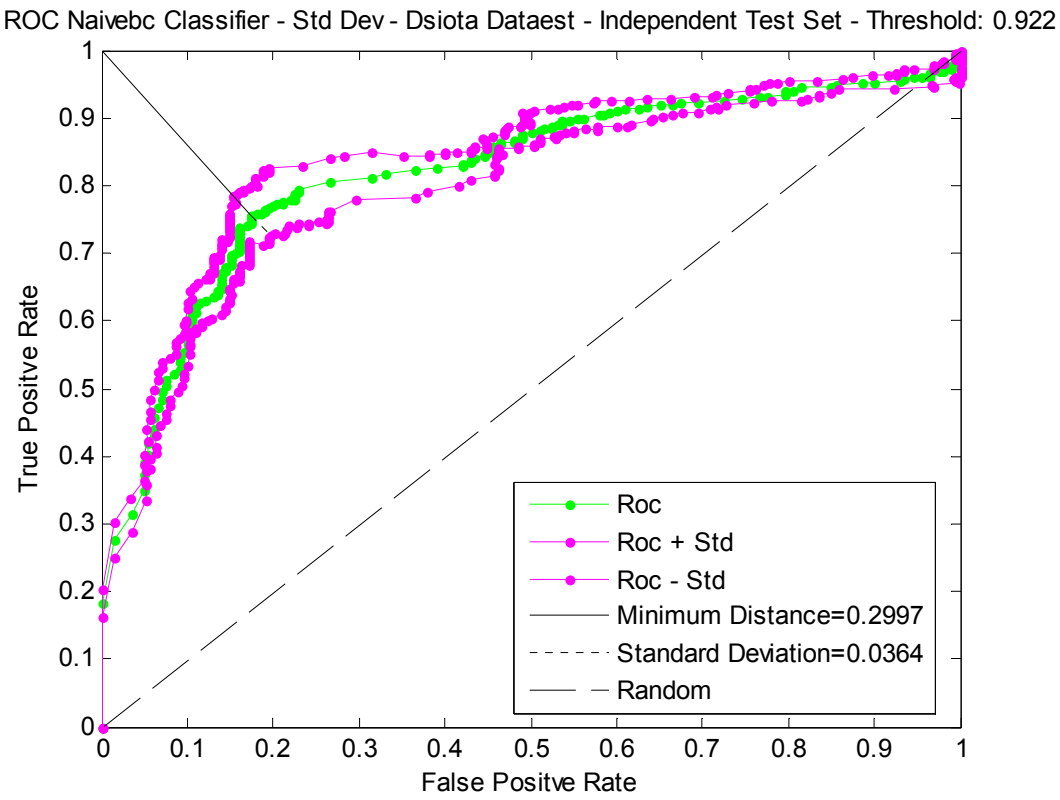
Loglc


ROC Loglc Classifier - Standard Deviation - Dsiota Dataest - Evaluation Set-Threshold: 0.918
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3913
Standard Deviation=0.1261
Random

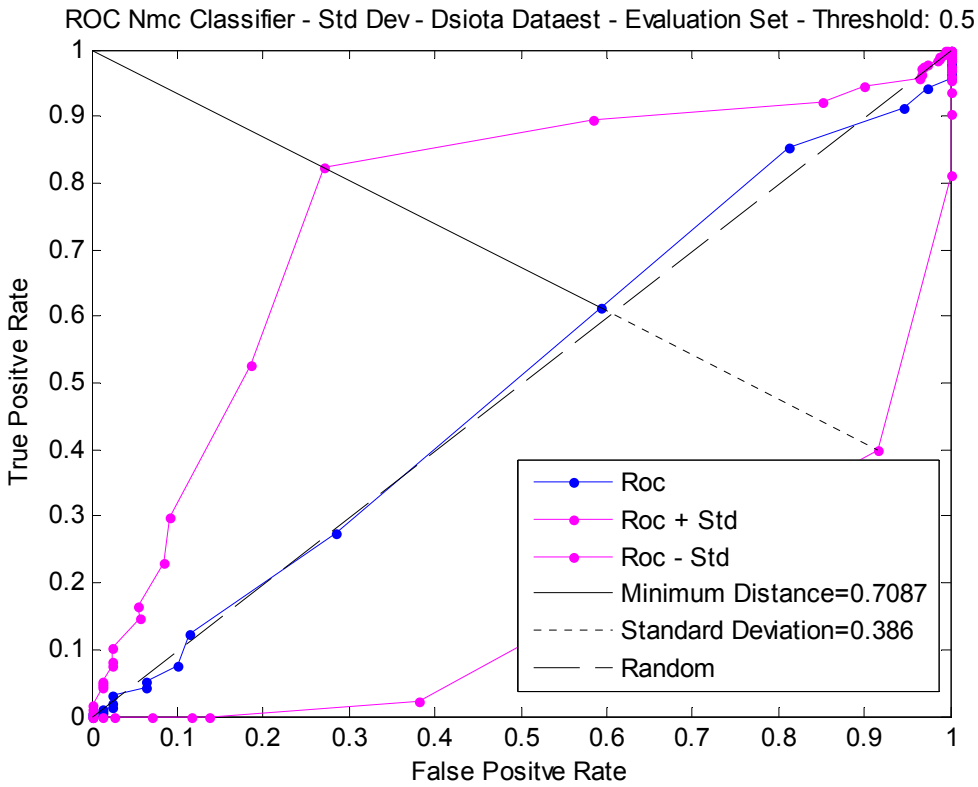

ROC Loglc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.901
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3225
Standard Deviation=0.0434
Random

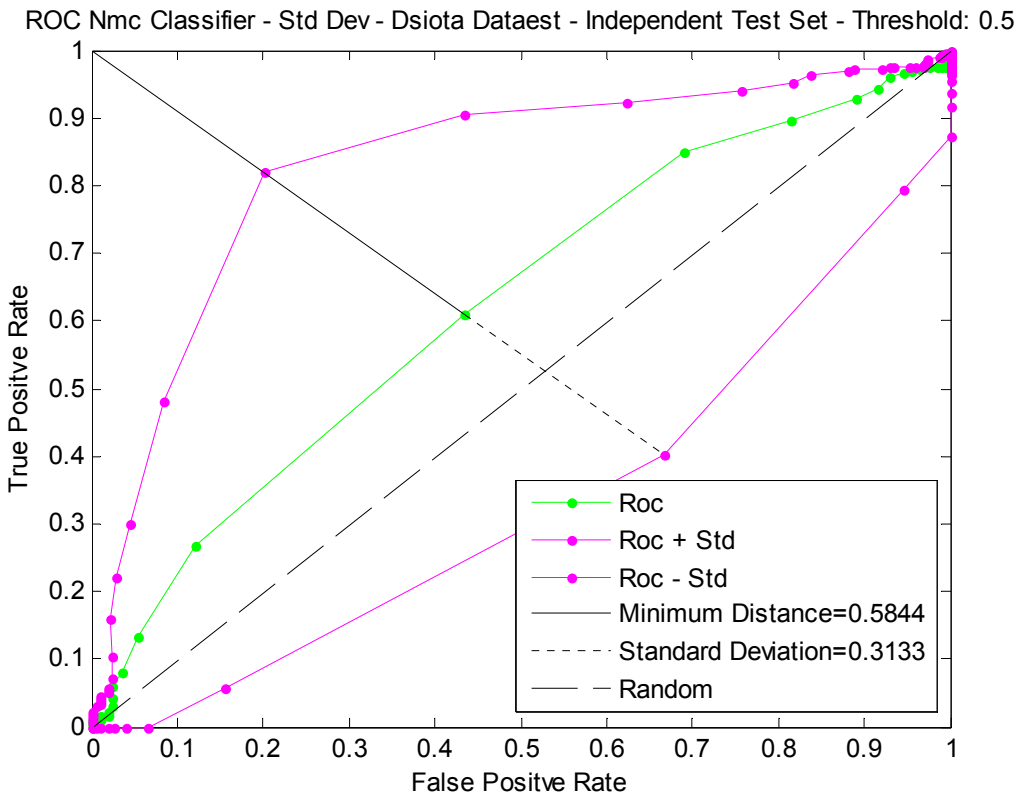
Naivebc


ROC Naivebc Classifier - Standard Deviation - Dsiota Dataest - Evaluation Set-Threshold 0.957
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3679
Standard Deviation=0.1575
Random

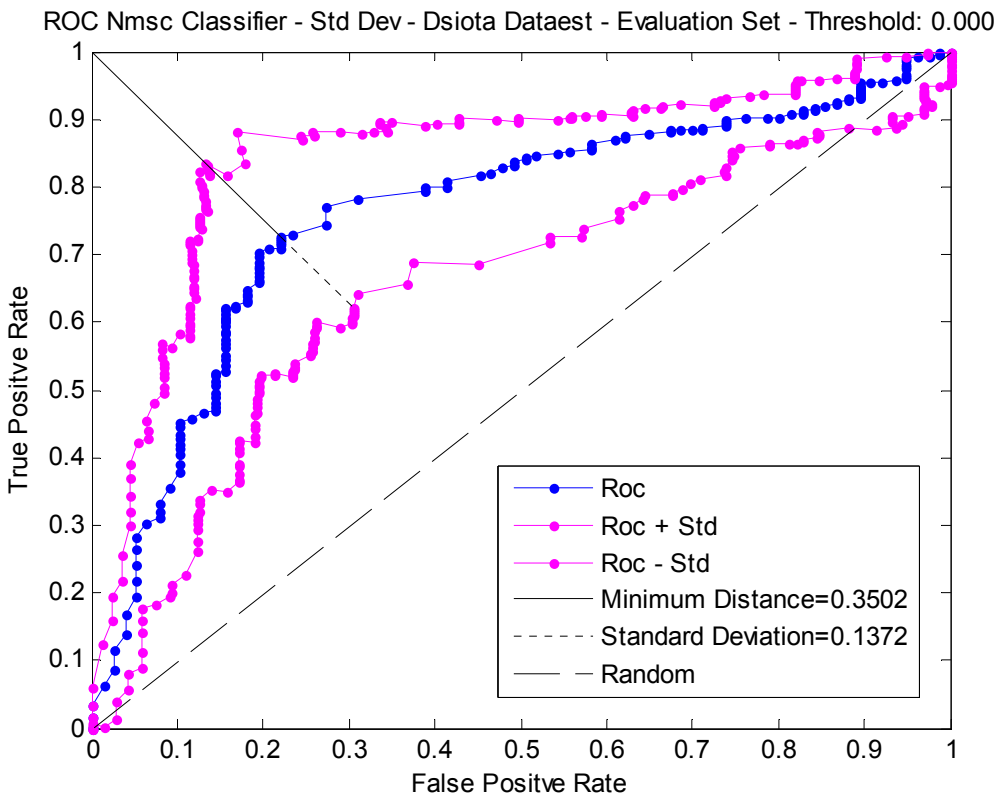

ROC Naivebc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.922
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.2997
Standard Deviation=0.0364
Random

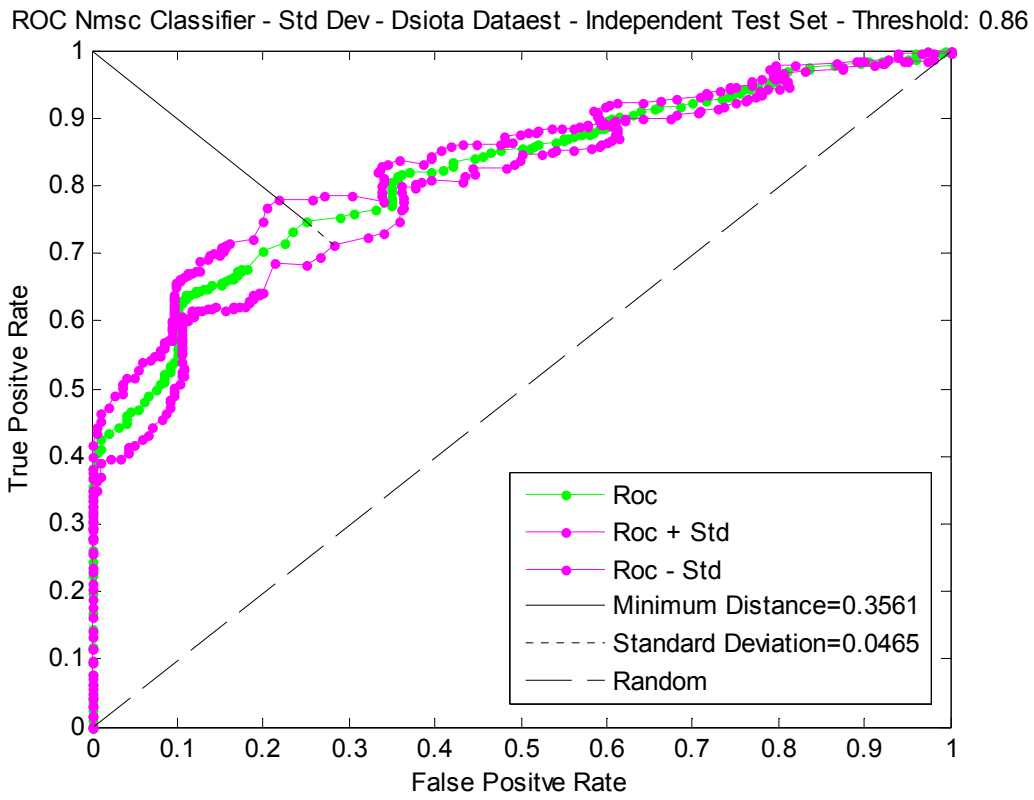
Nmc


ROC Nmc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.5
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.7087
Standard Deviation=0.386
Random

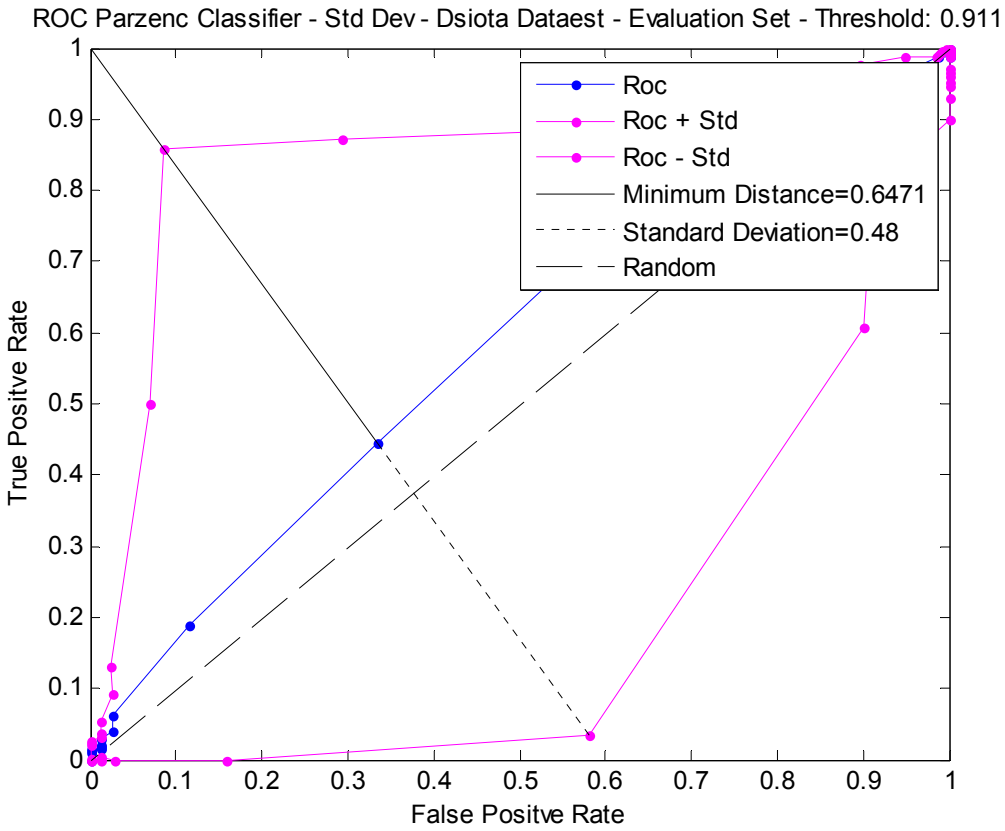

ROC Nmc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.5
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5844
Standard Deviation=0.3133
Random

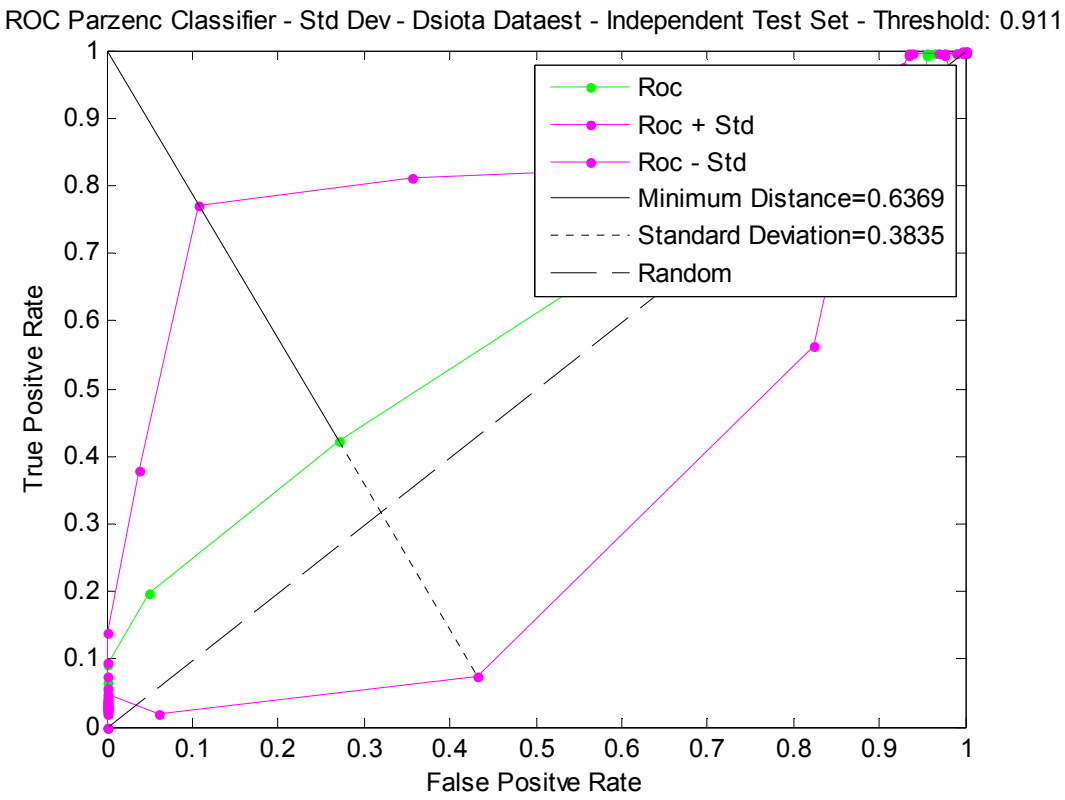
Nmsc


ROC Nmsc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.000
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3502
Standard Deviation=0.1372
Random

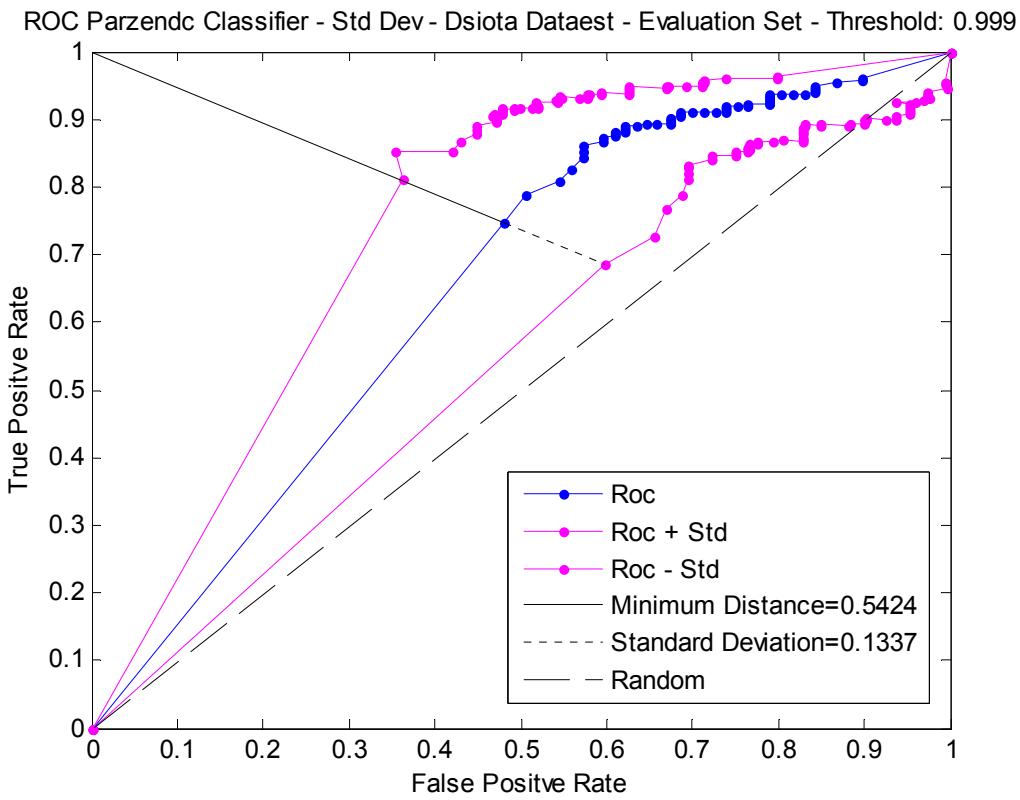

ROC Nmsc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.86
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3561
Standard Deviation=0.0465
Random

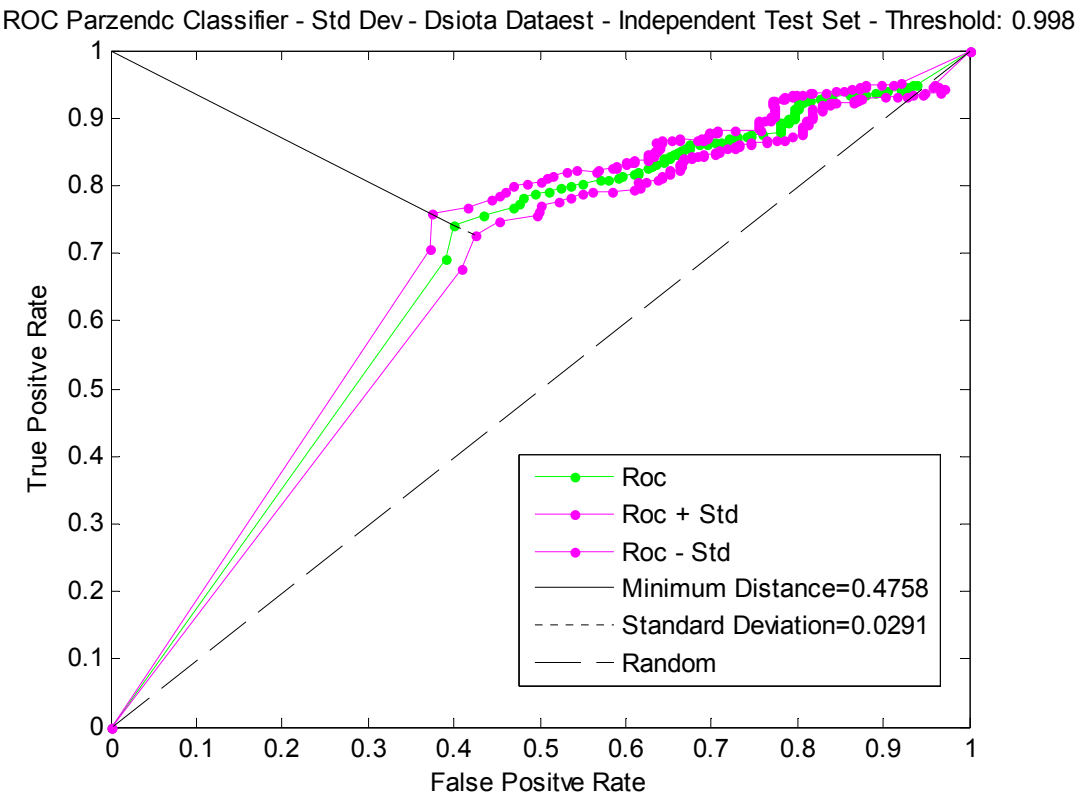
Parzenc


ROC Parzenc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.911
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.6471
Standard Deviation=0.48
Random

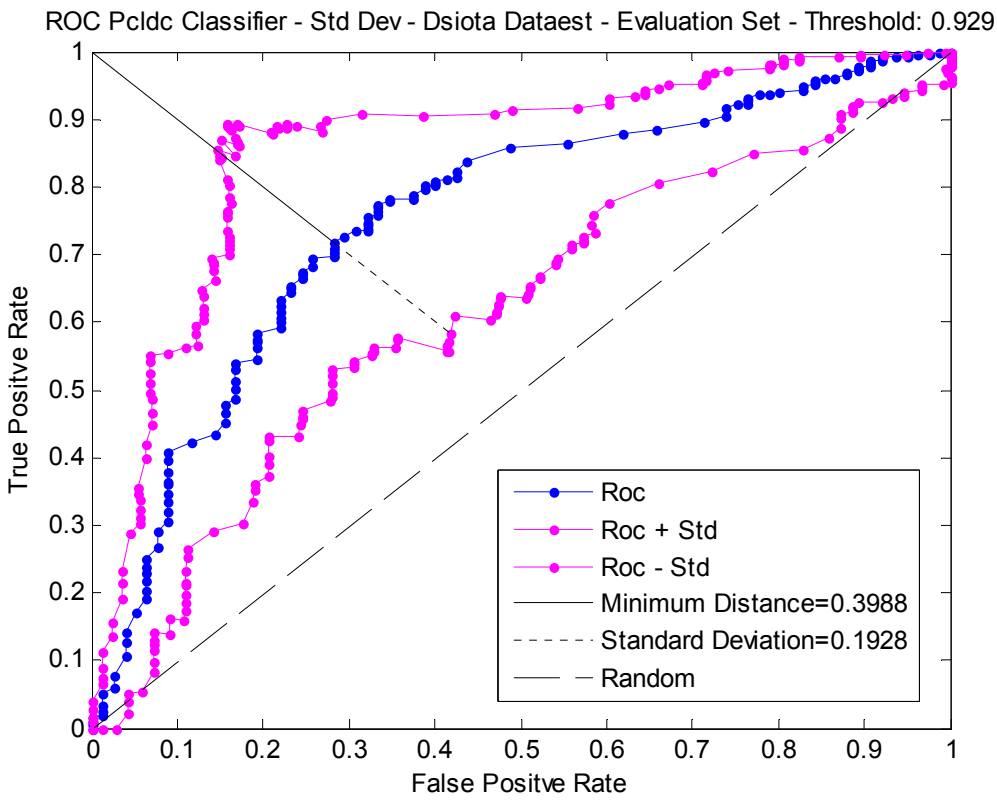

ROC Parzenc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.911
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.6369
Standard Deviation=0.3835
Random

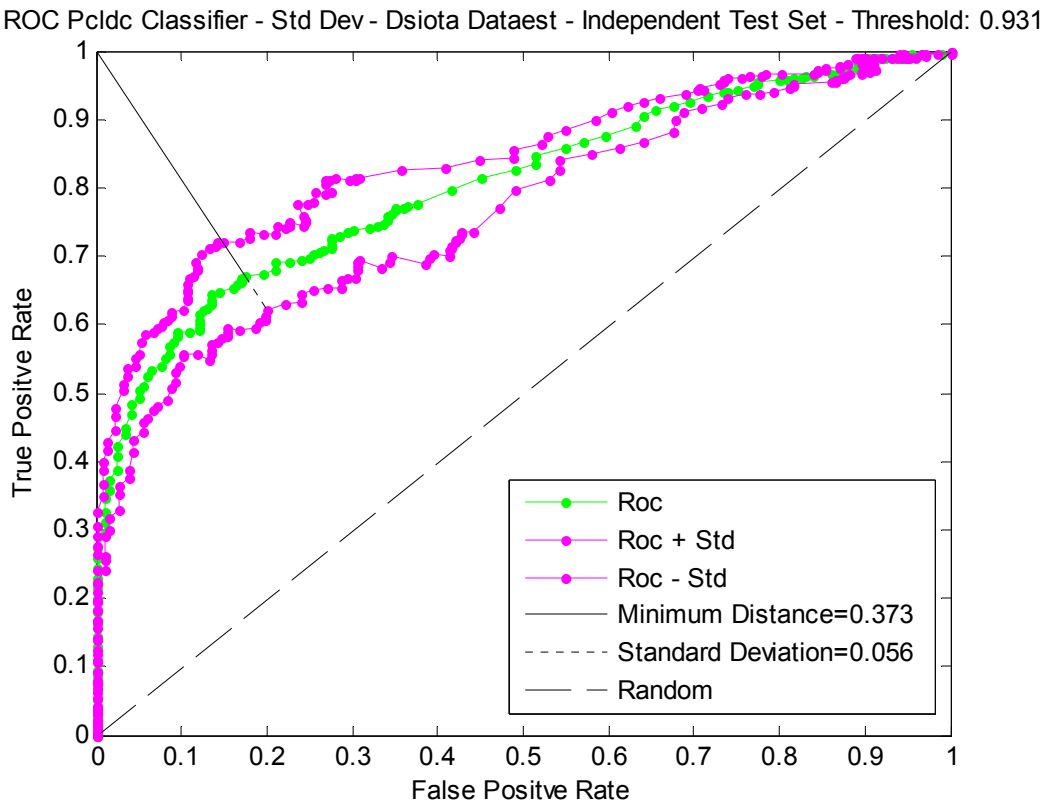
Parzendc


ROC Parzendc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.999
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5424
Standard Deviation=0.1337
Random

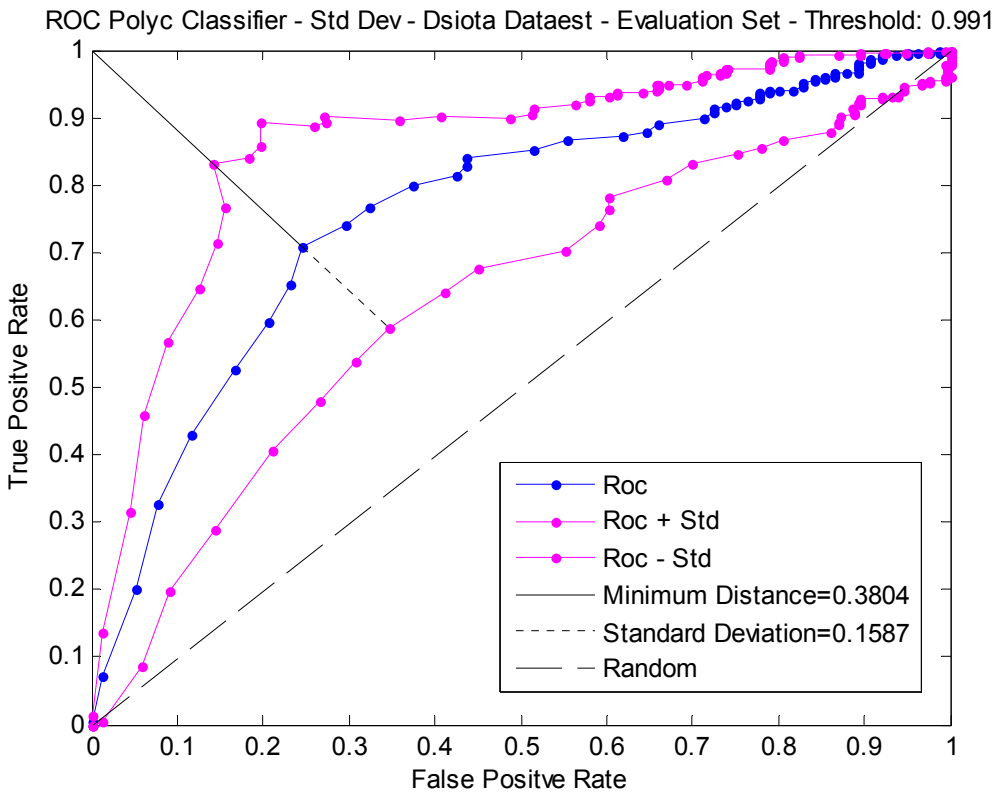

ROC Parzendc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.998
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4758
Standard Deviation=0.0291
Random

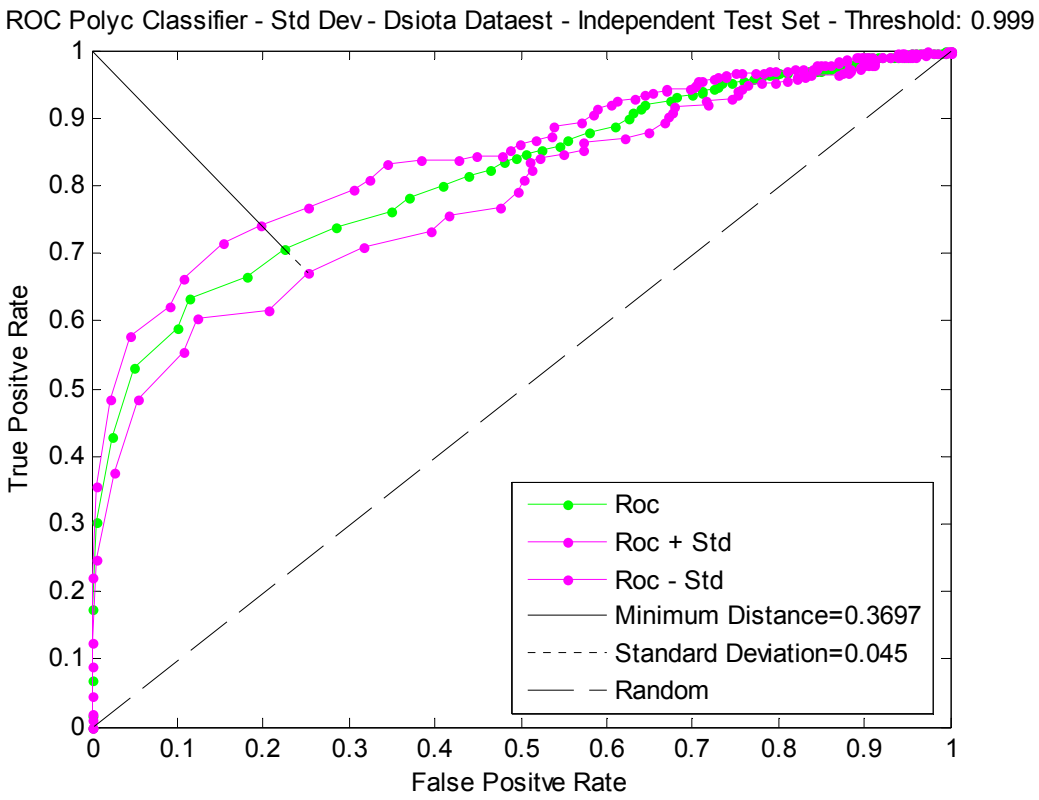
Pcldc


ROC Pcldc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.929
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3988
Standard Deviation=0.1928
Random

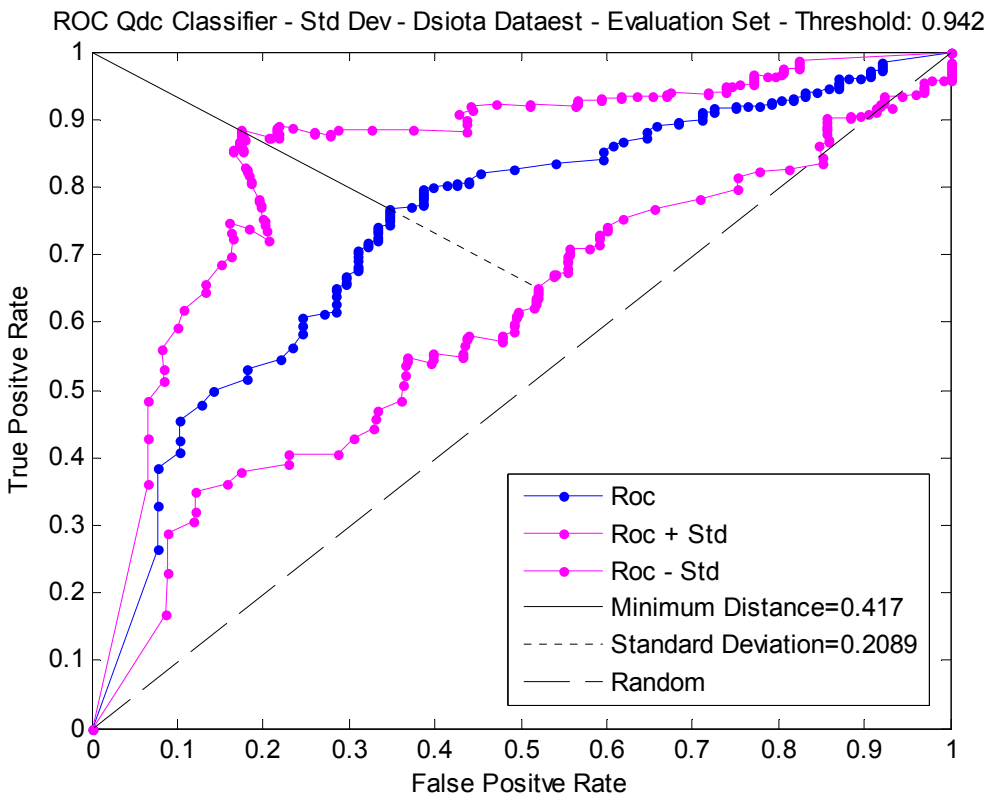

ROC Pcldc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.931
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.373
Standard Deviation=0.056
Random

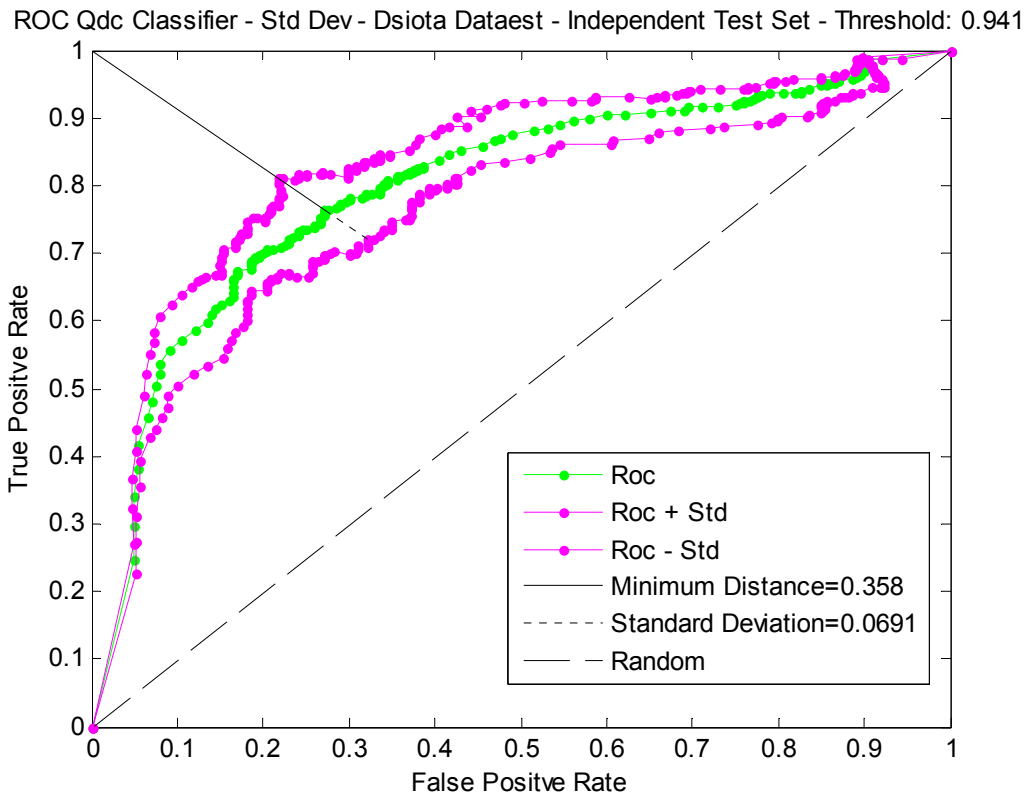
Polyc


ROC Polyc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.991
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3804
Standard Deviation=0.1587
Random

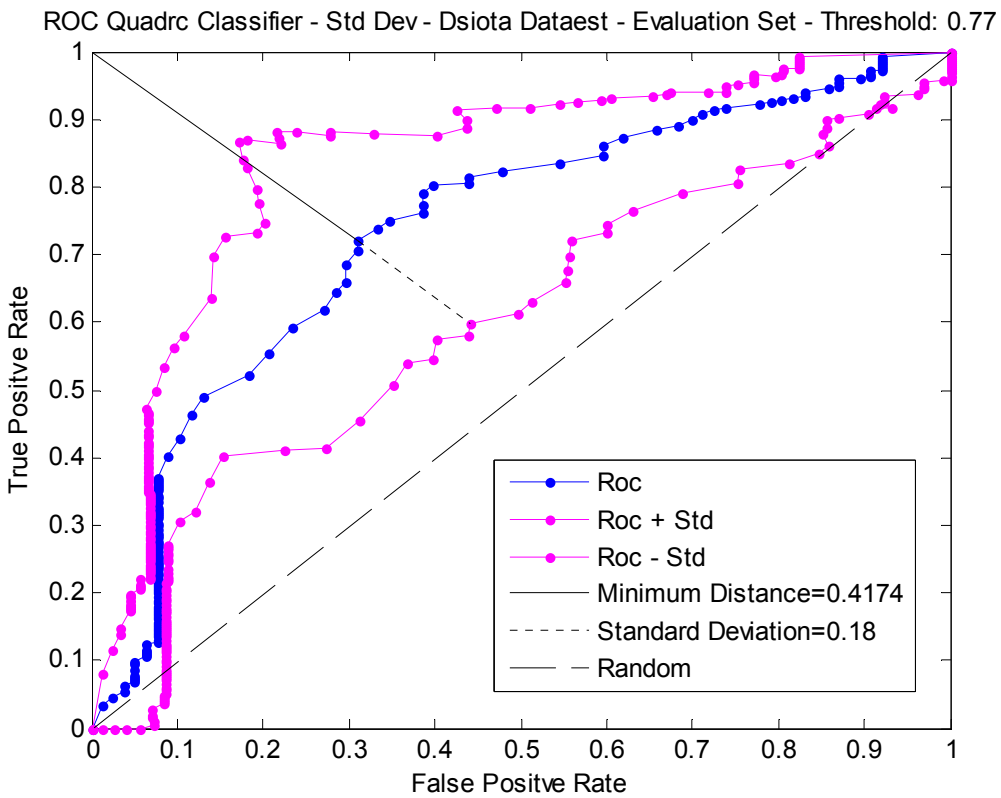

ROC Polyc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.999
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3697
Standard Deviation=0.045
Random

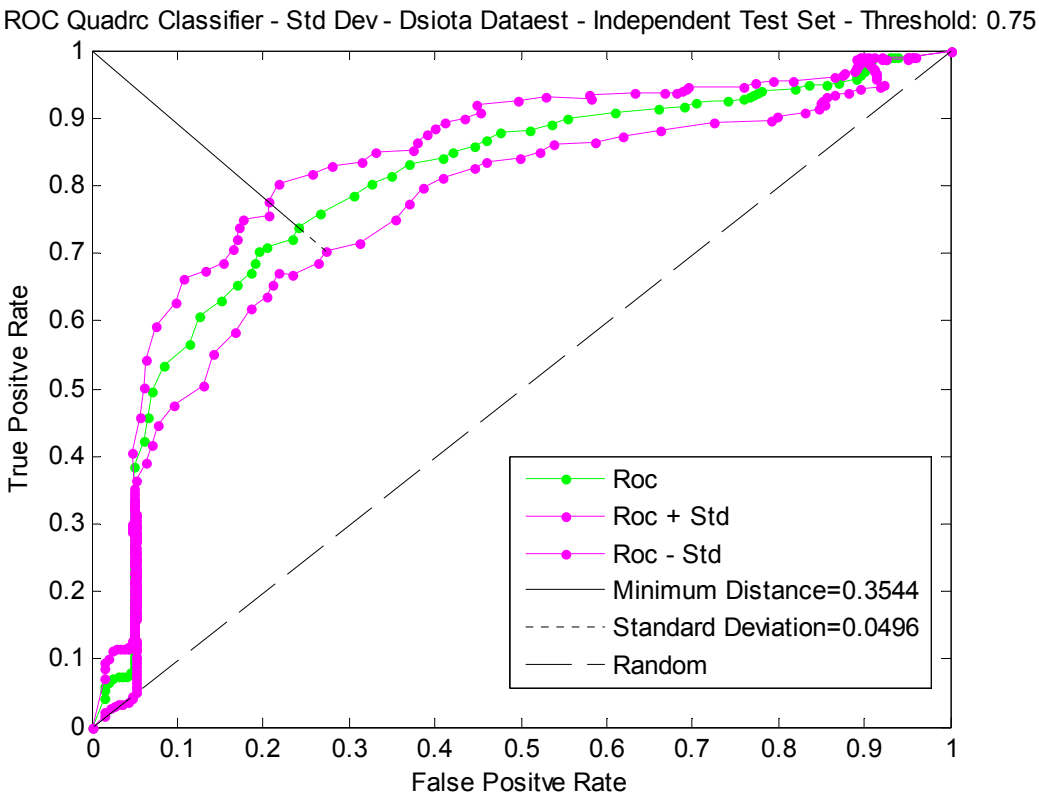
Qdc


ROC Qdc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.942
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.417
Standard Deviation=0.2089
Random

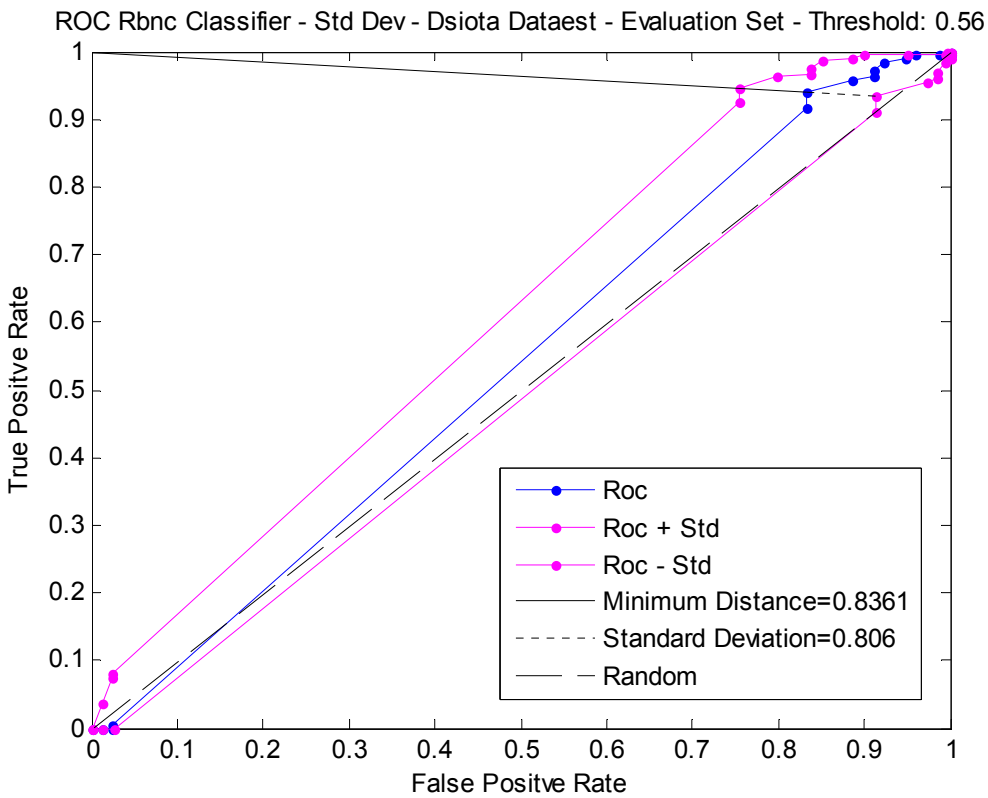

ROC Qdc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.941
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.358
Standard Deviation=0.0691
Random

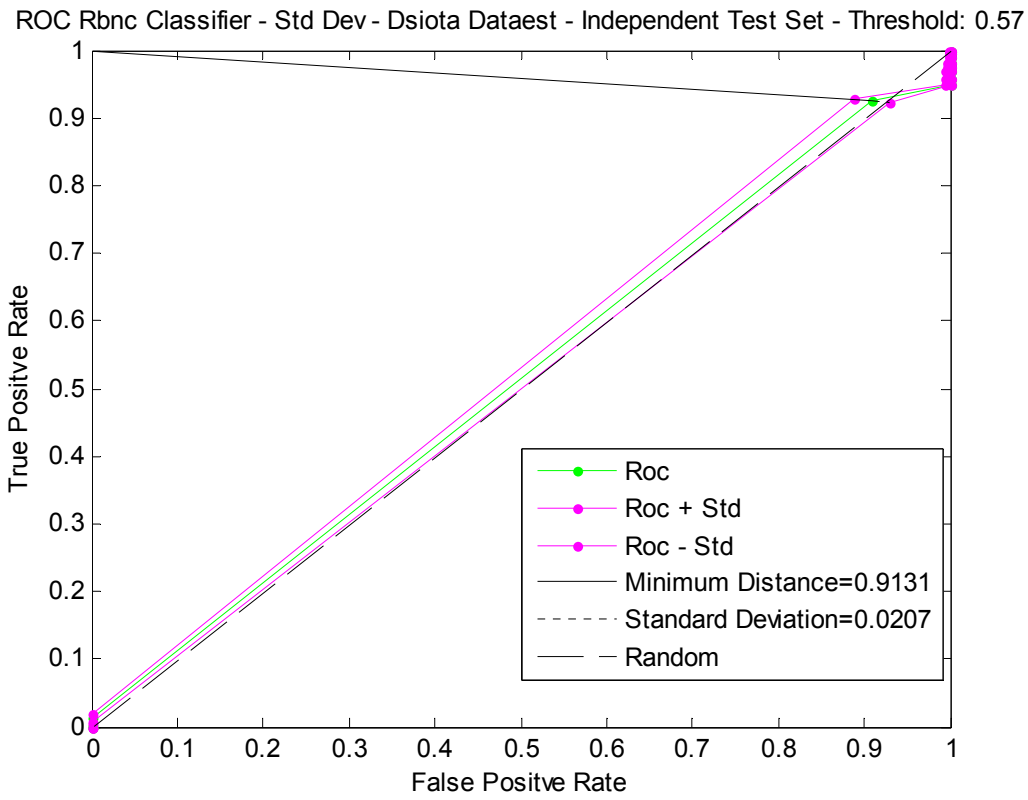
Quadrc


ROC Quadrc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.77
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4174
Standard Deviation=0.18
Random

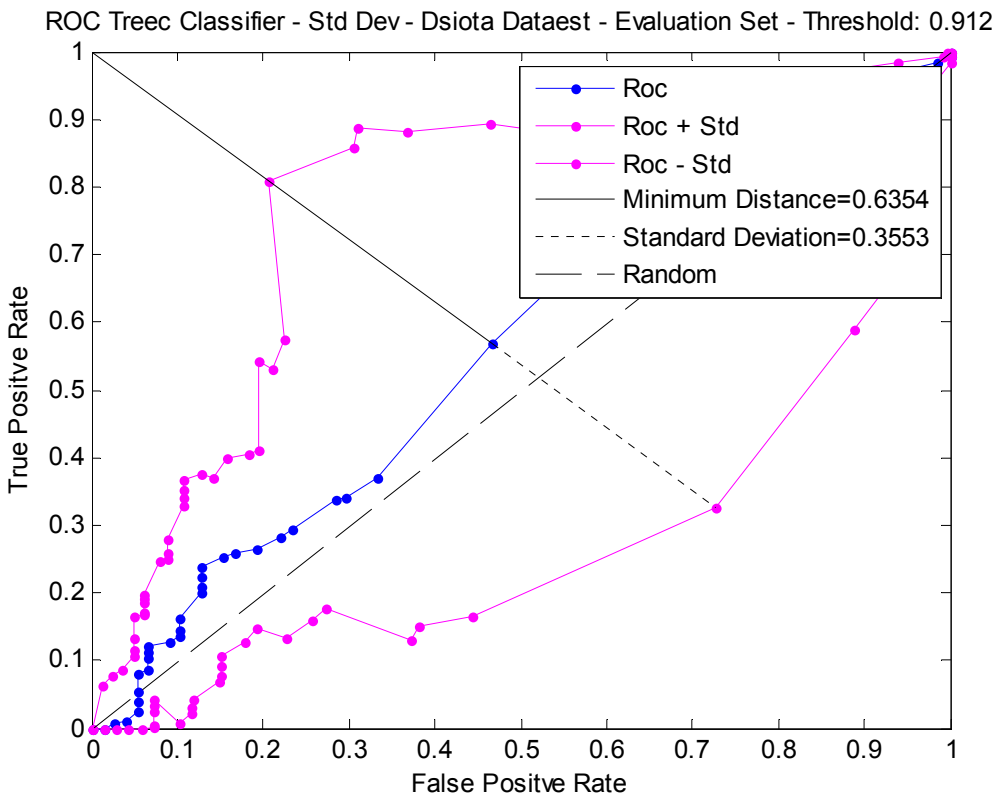

ROC Quadrc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.75
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3544
Standard Deviation=0.0496
Random

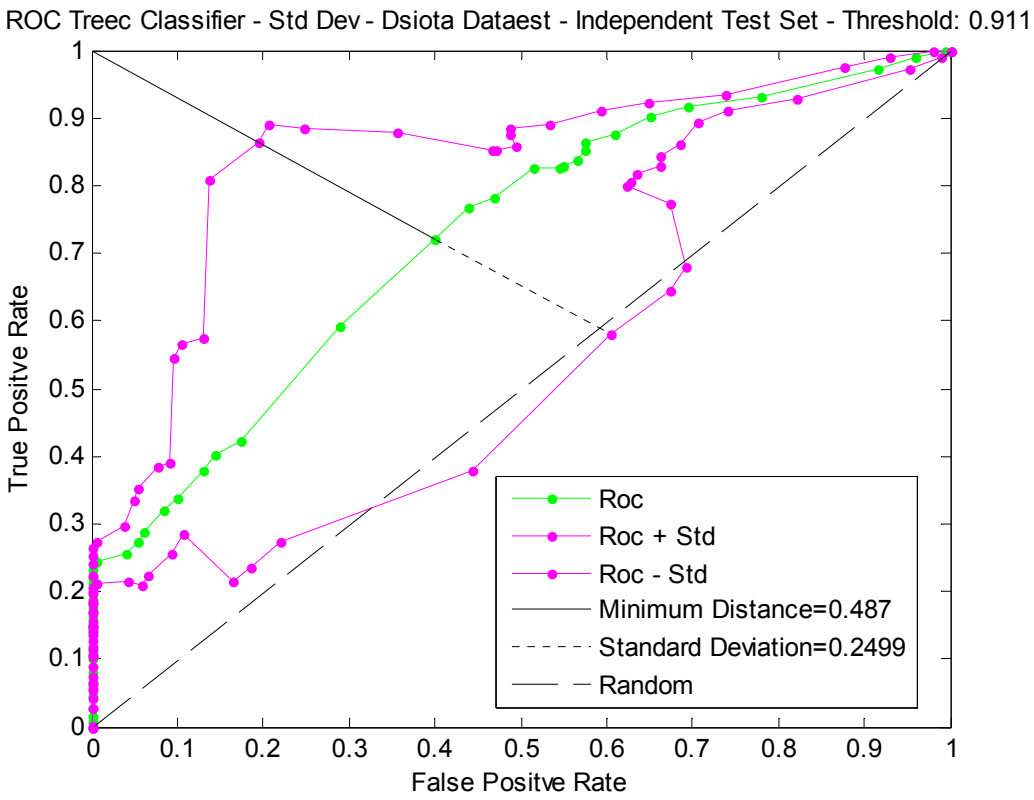
Rbnc


ROC Rbnc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.56
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.8361
Standard Deviation=0.806
Random

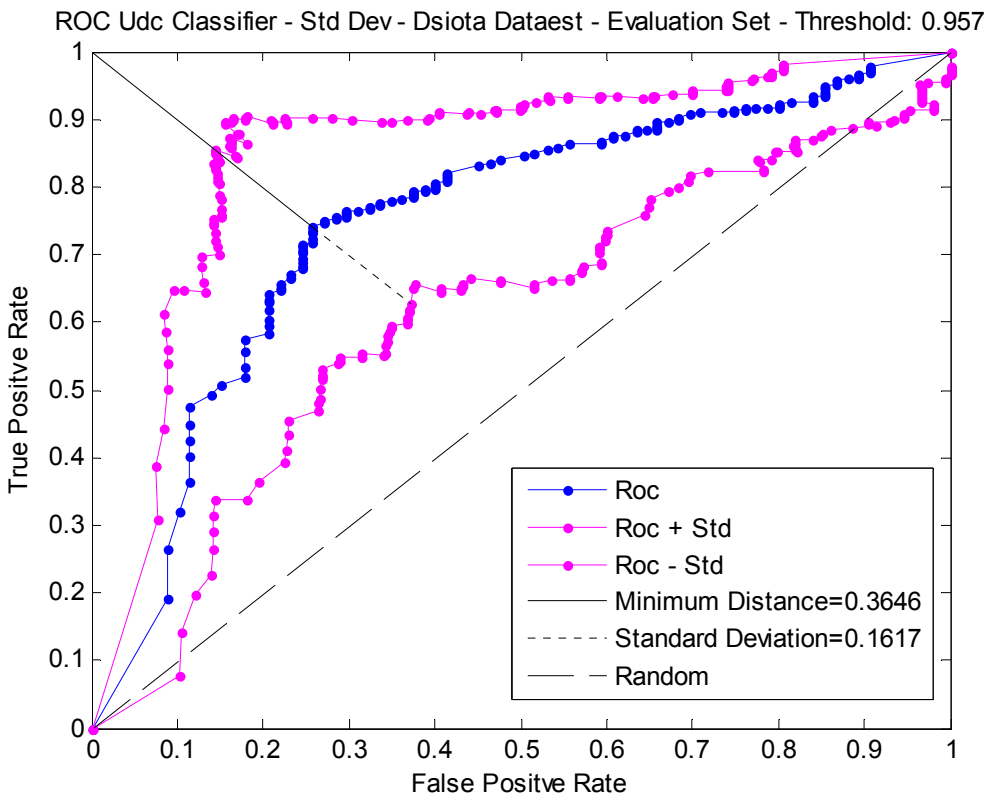

ROC Rbnc Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.57
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.9131
Standard Deviation=0.0207
Random

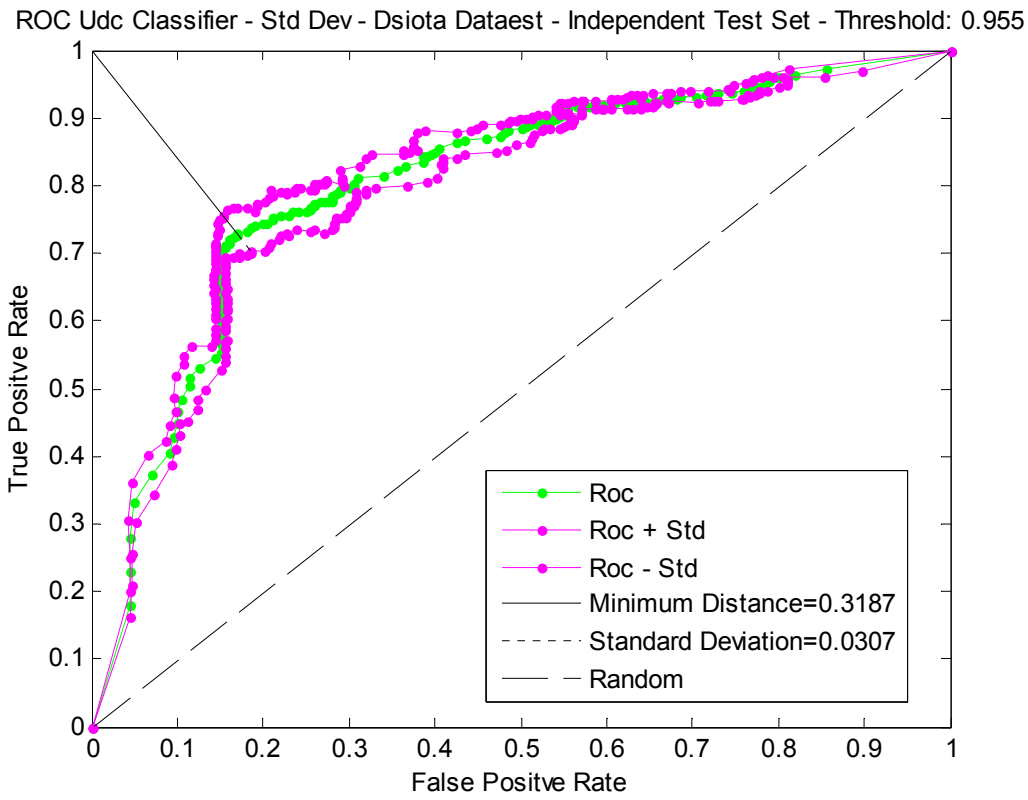
Treec


ROC Treec Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.912
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.6354
Standard Deviation=0.3553
Random


ROC Treec Classifier - Std Dev - Dsiota Dataest - Independent Test Set - Threshold: 0.911
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.487
Standard Deviation=0.2499
Random

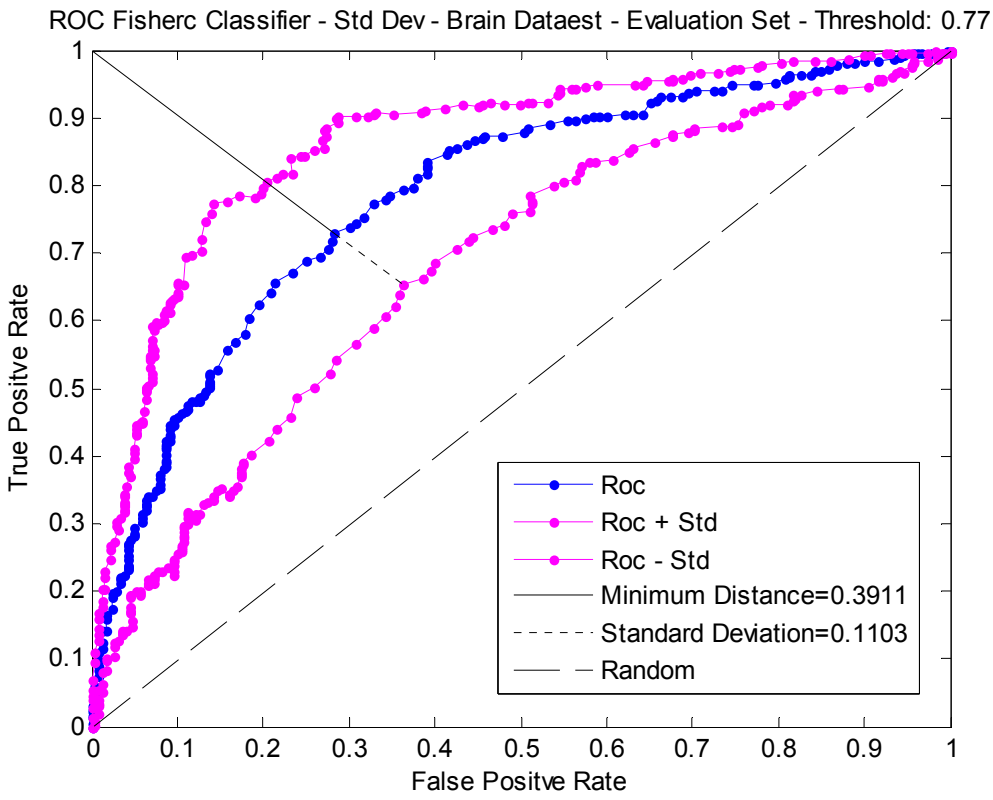
Udc

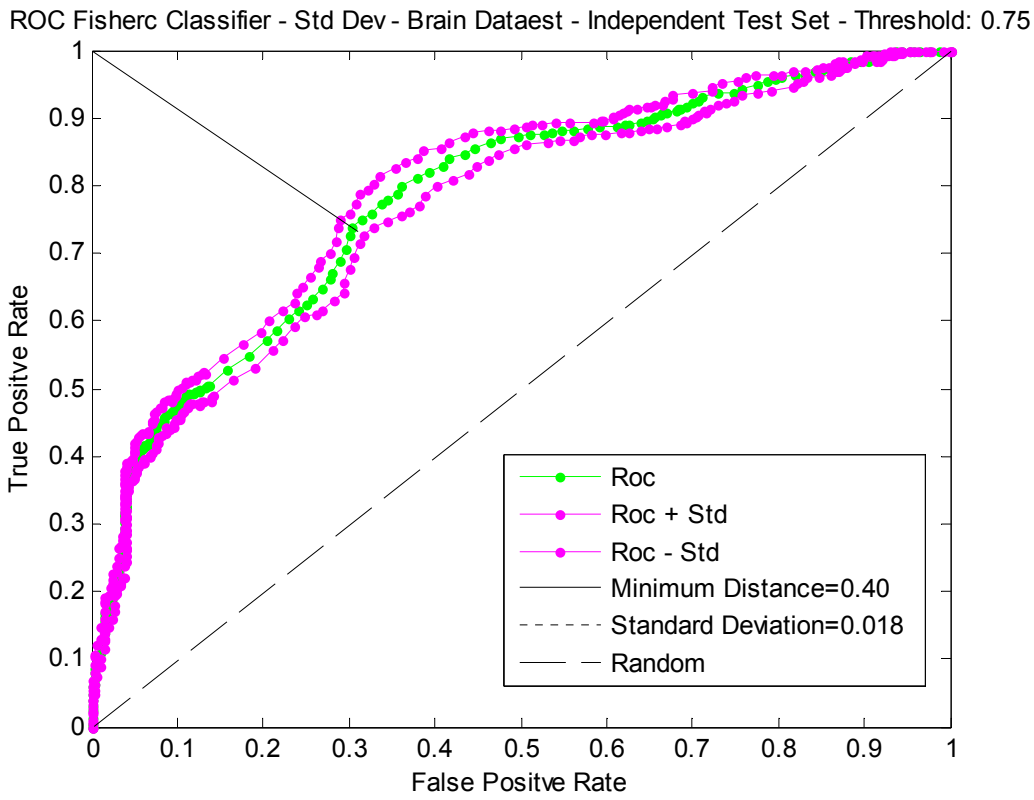

ROC Udc Classifier - Std Dev - Dsiota Dataest - Evaluation Set - Threshold: 0.957
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3646
Standard Deviation=0.1617
Random



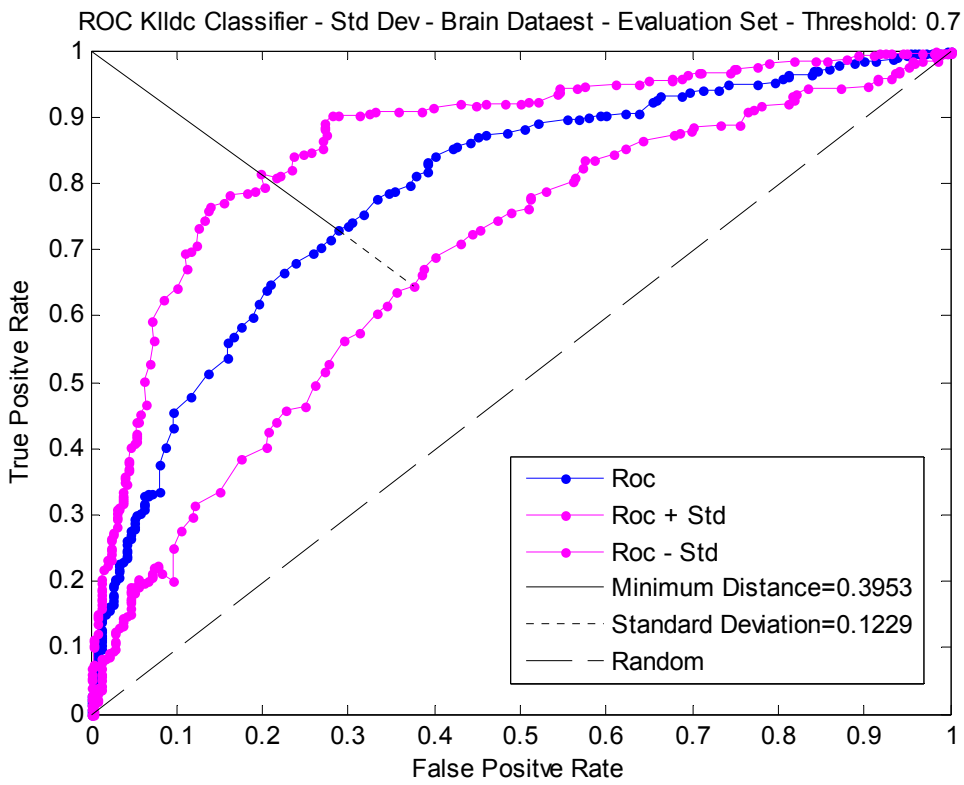

### 8.1.2. Brain Cancer Dataset

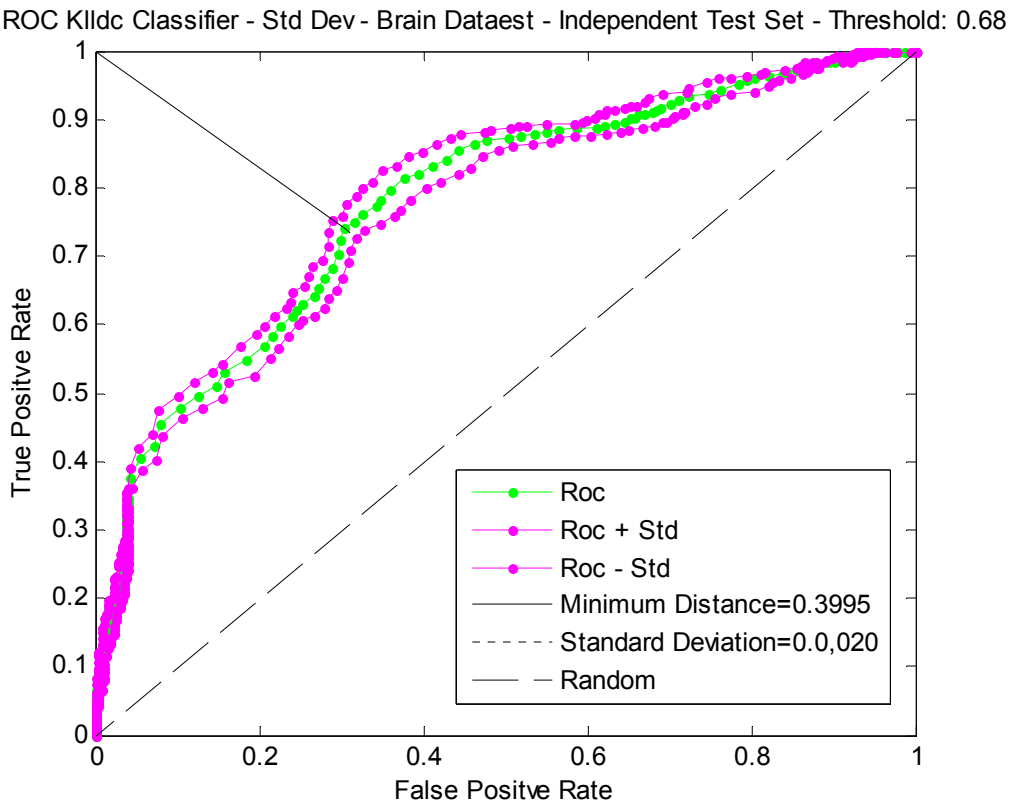
Fisherc



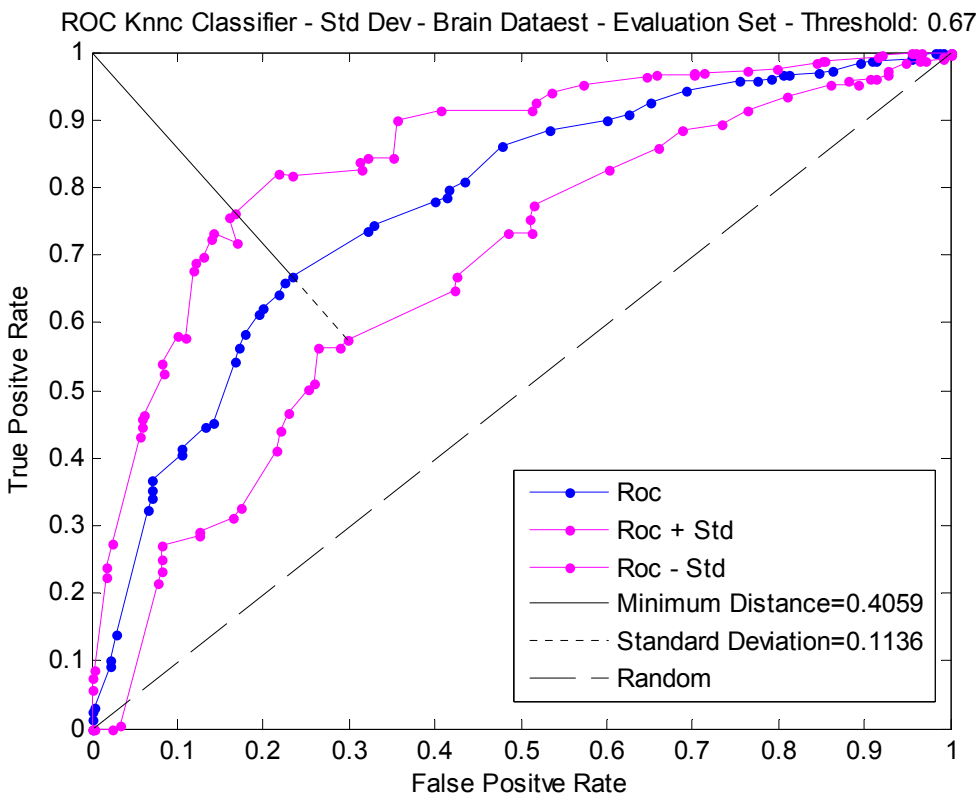

ROC Fisherc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.75
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.40
Standard Deviation=0.018
Random

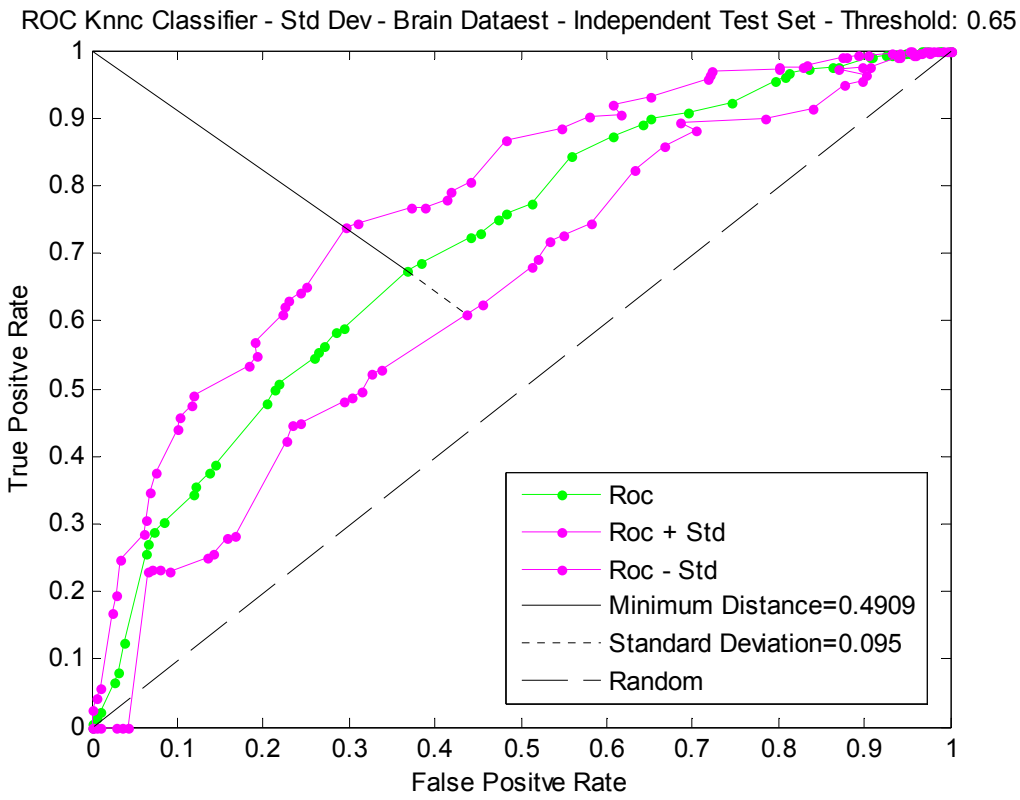
Klldc


ROC Klldc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.7
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3953
Standard Deviation=0.1229
Random

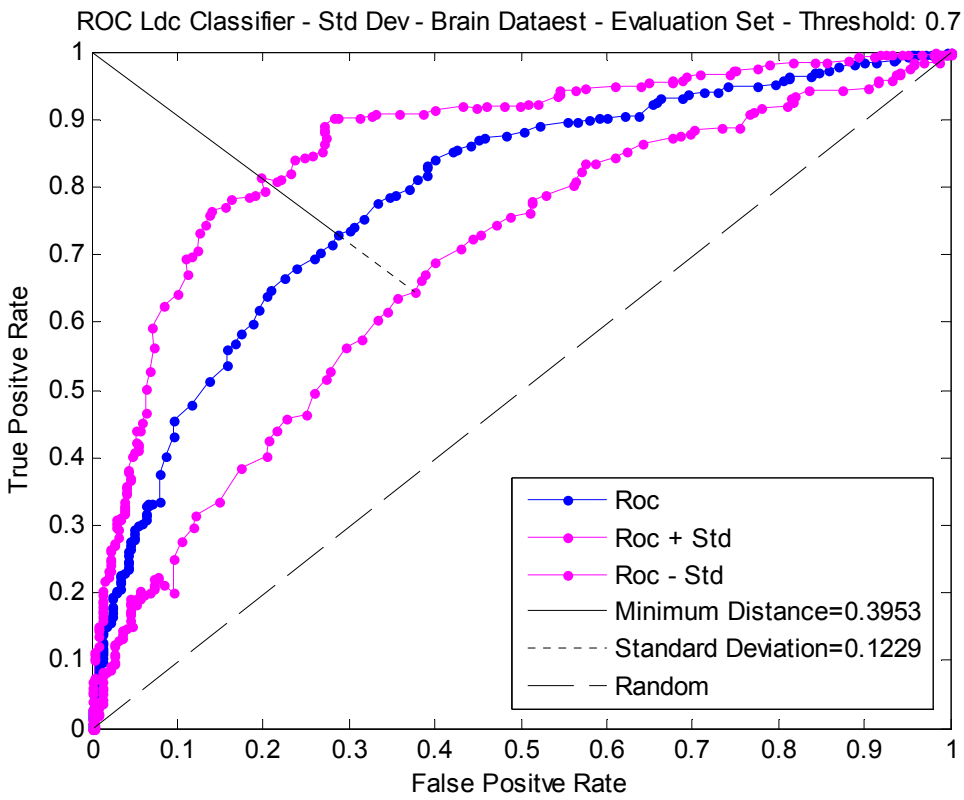

ROC Klldc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.68
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3995
Standard Deviation=0.0,020
Random

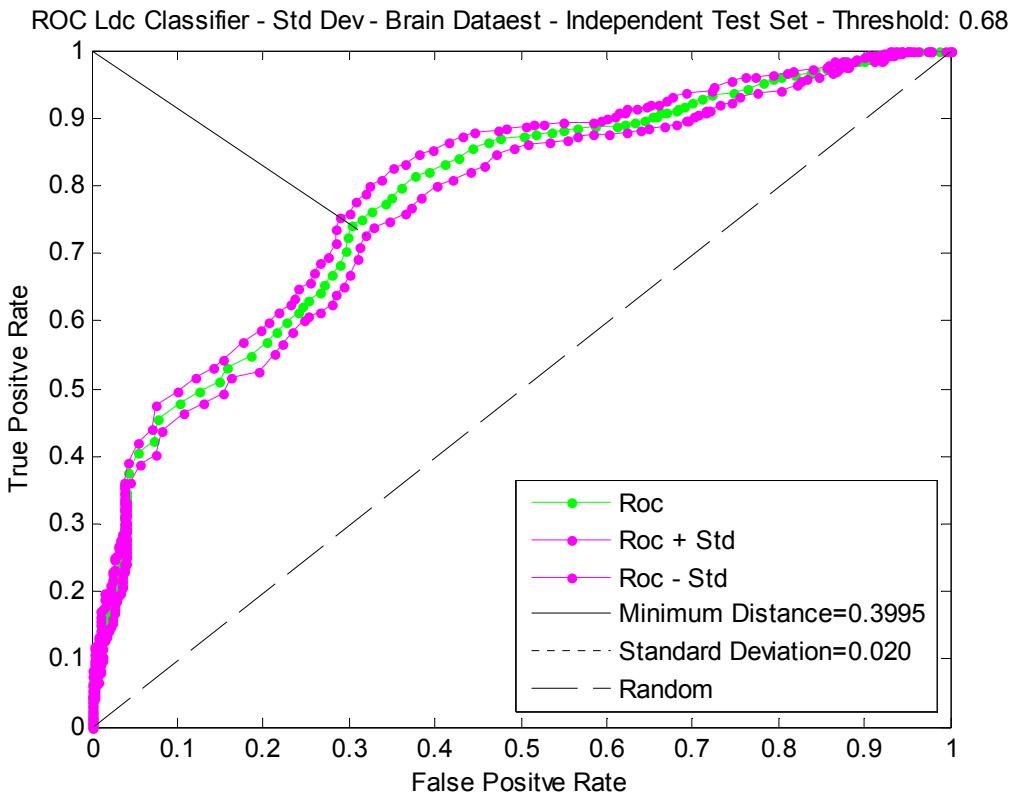
Knnc


ROC Knnc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.67
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4059
Standard Deviation=0.1136
Random

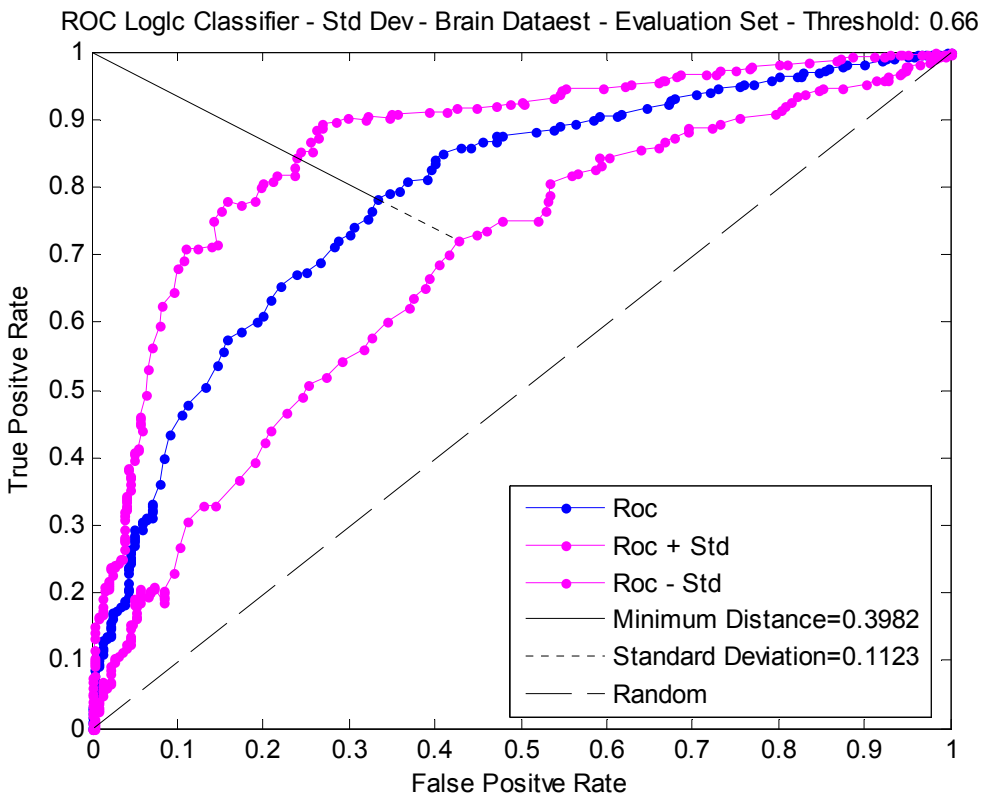

ROC Knnc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.65
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4909
Standard Deviation=0.095
Random

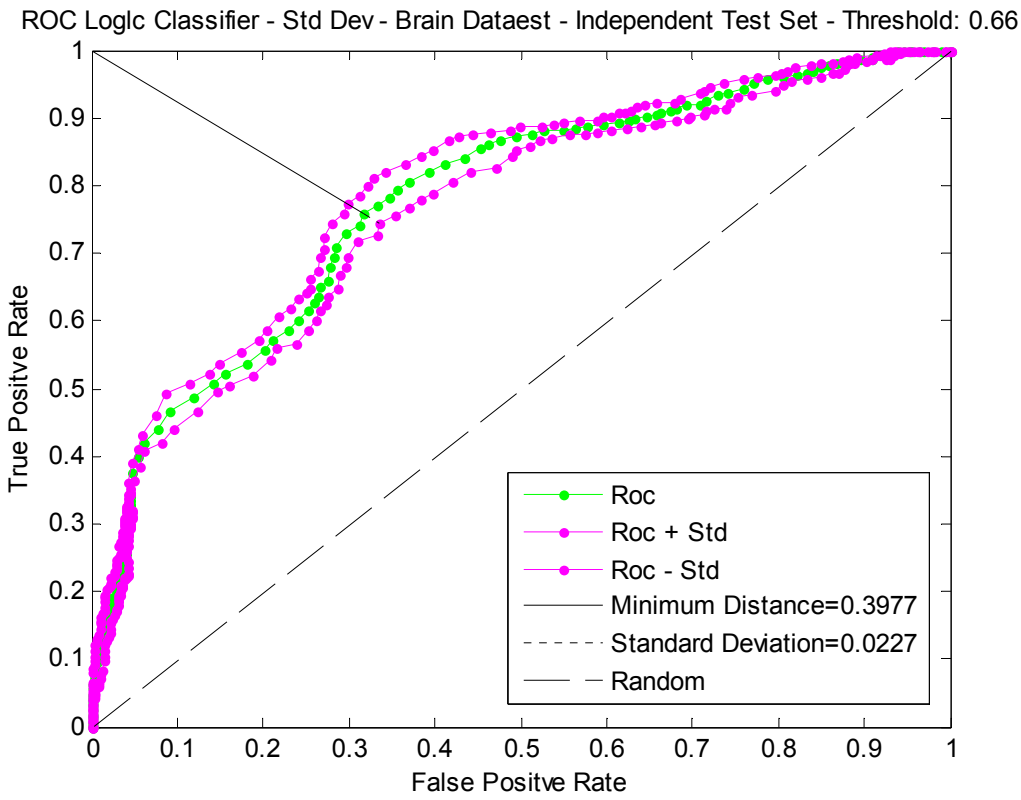
Ldc


ROC Ldc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.7
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3953
Standard Deviation=0.1229
Random

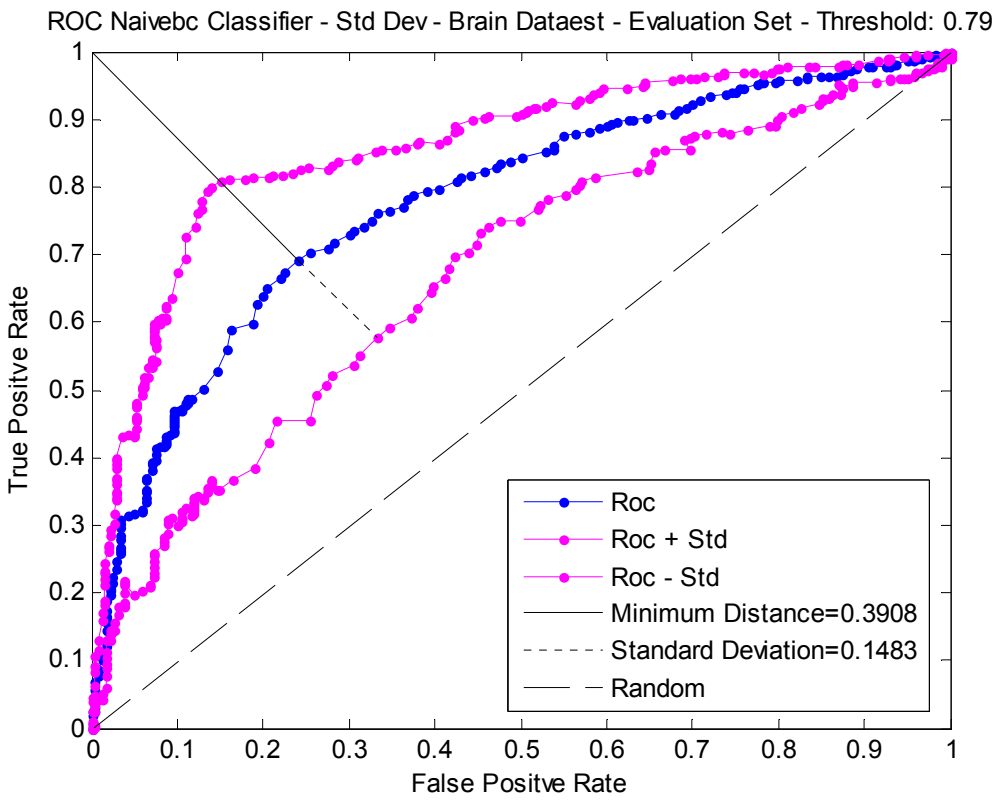

ROC Ldc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.68
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3995
Standard Deviation=0.020
Random

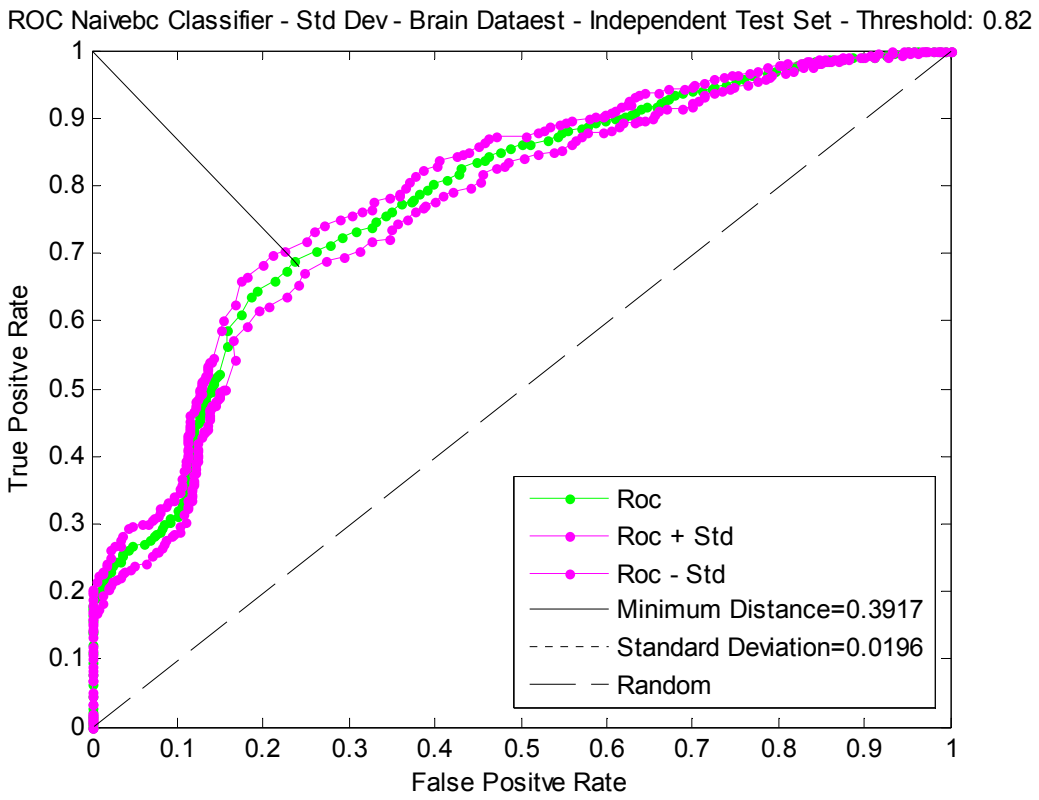
Loglc


ROC Loglc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.66
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3982
Standard Deviation=0.1123
Random

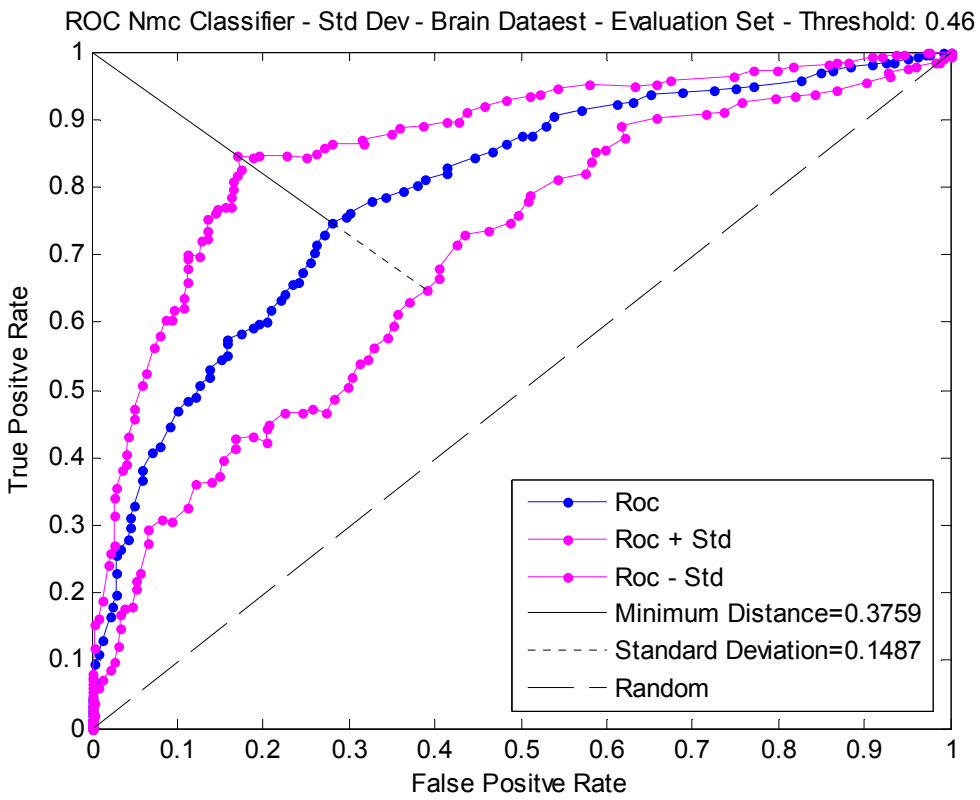

ROC Loglc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.66
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3977
Standard Deviation=0.0227
Random

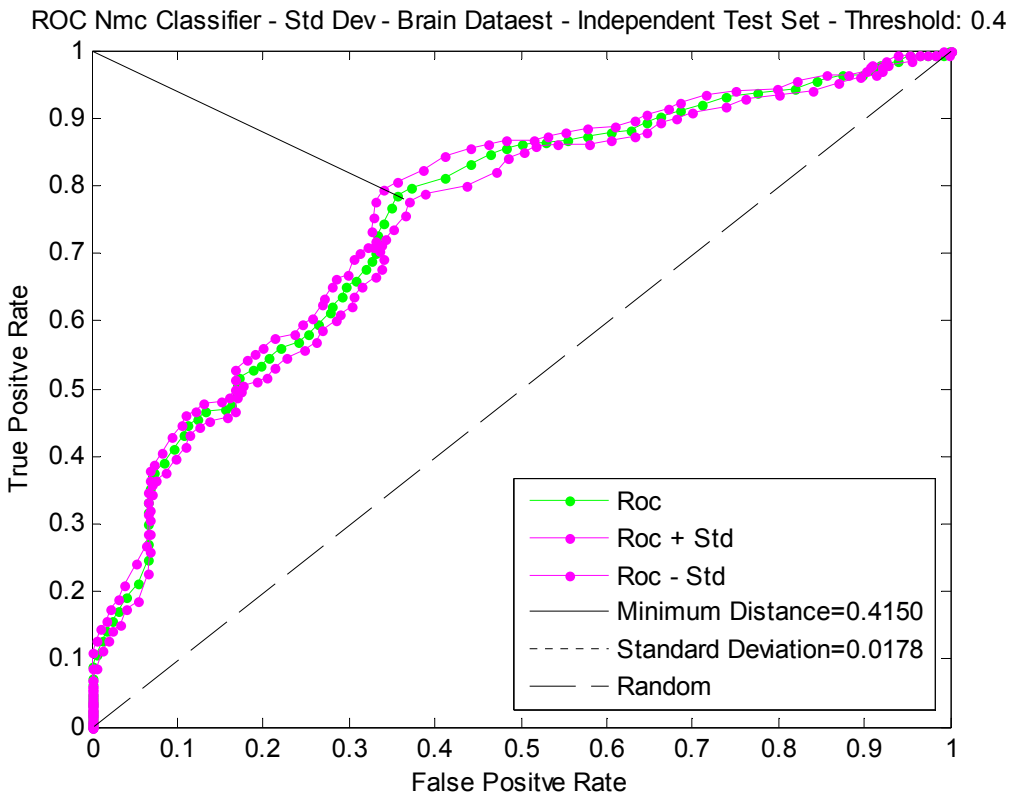
Naivebc


ROC Naivebc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.79
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3908
Standard Deviation=0.1483
Random

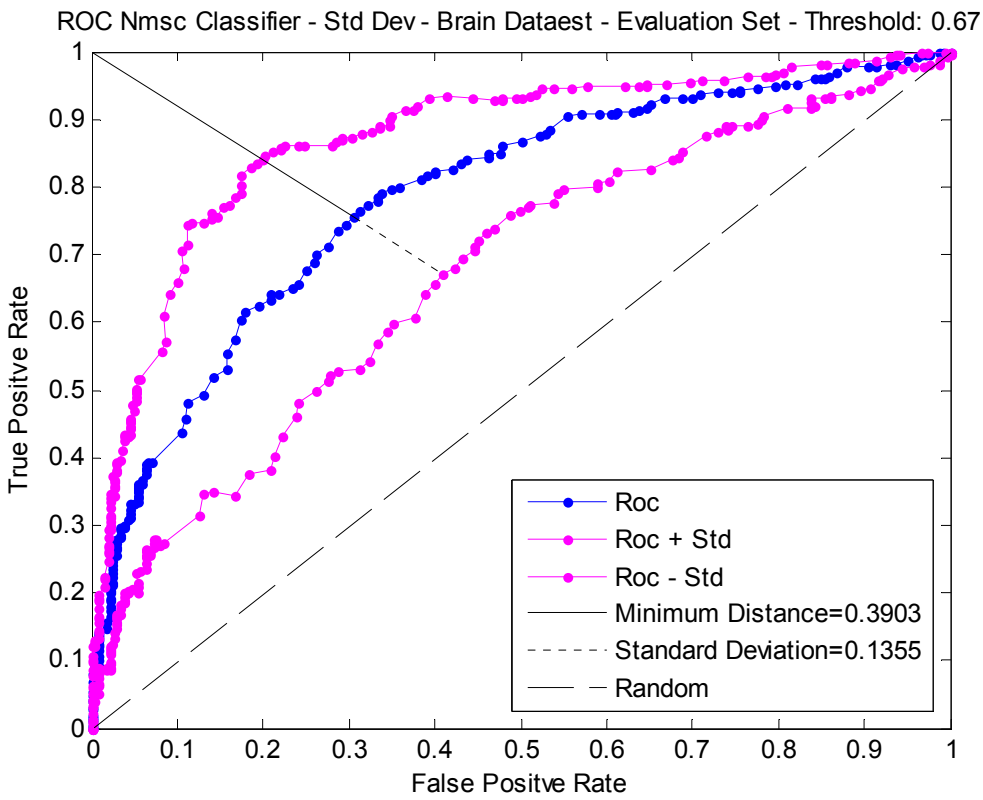

ROC Naivebc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.82
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3917
Standard Deviation=0.0196
Random

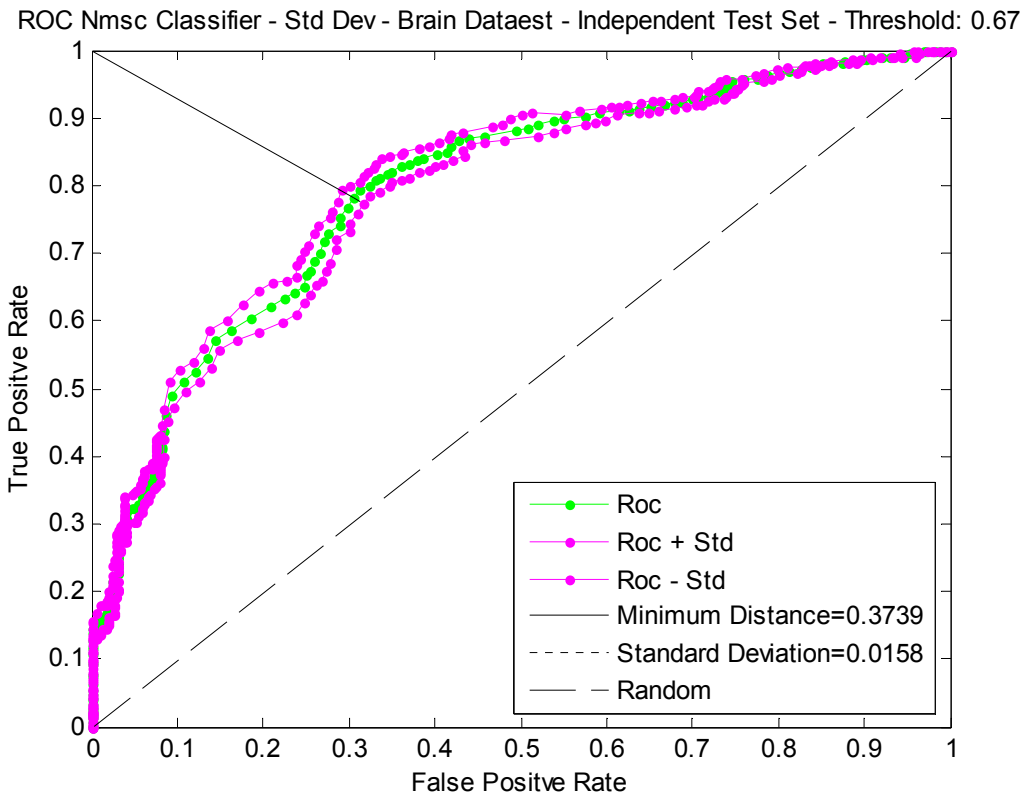
Nmc


ROC Nmc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.46
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3759
Standard Deviation=0.1487
Random

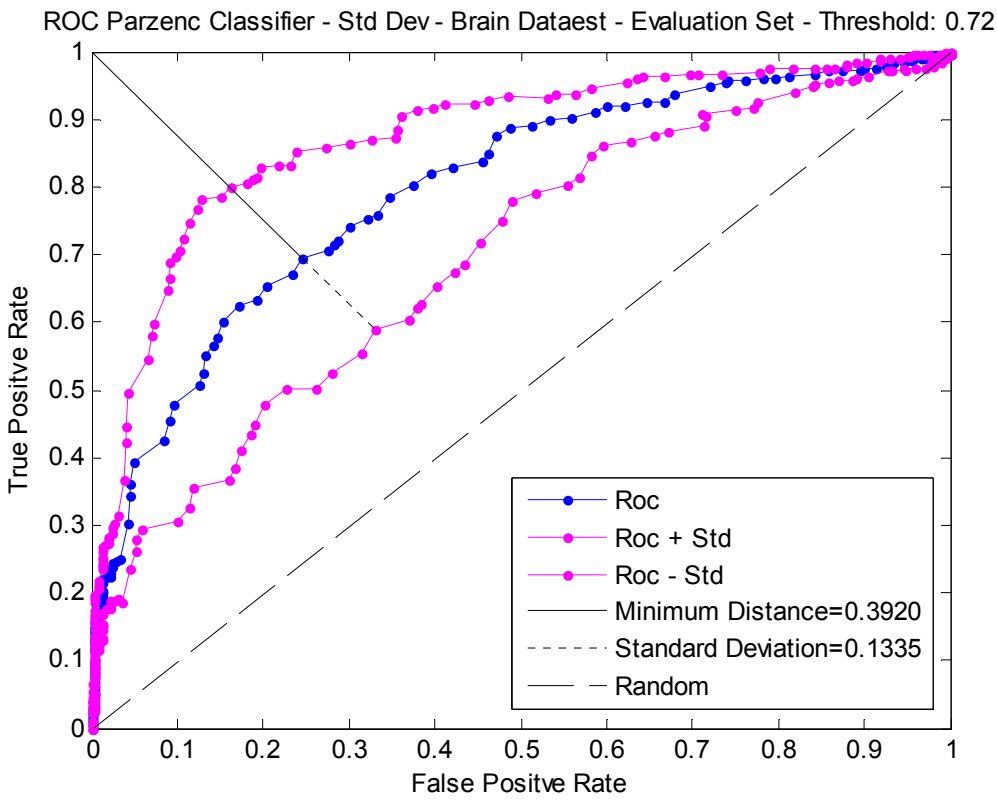

ROC Nmc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.4
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4150
Standard Deviation=0.0178
Random

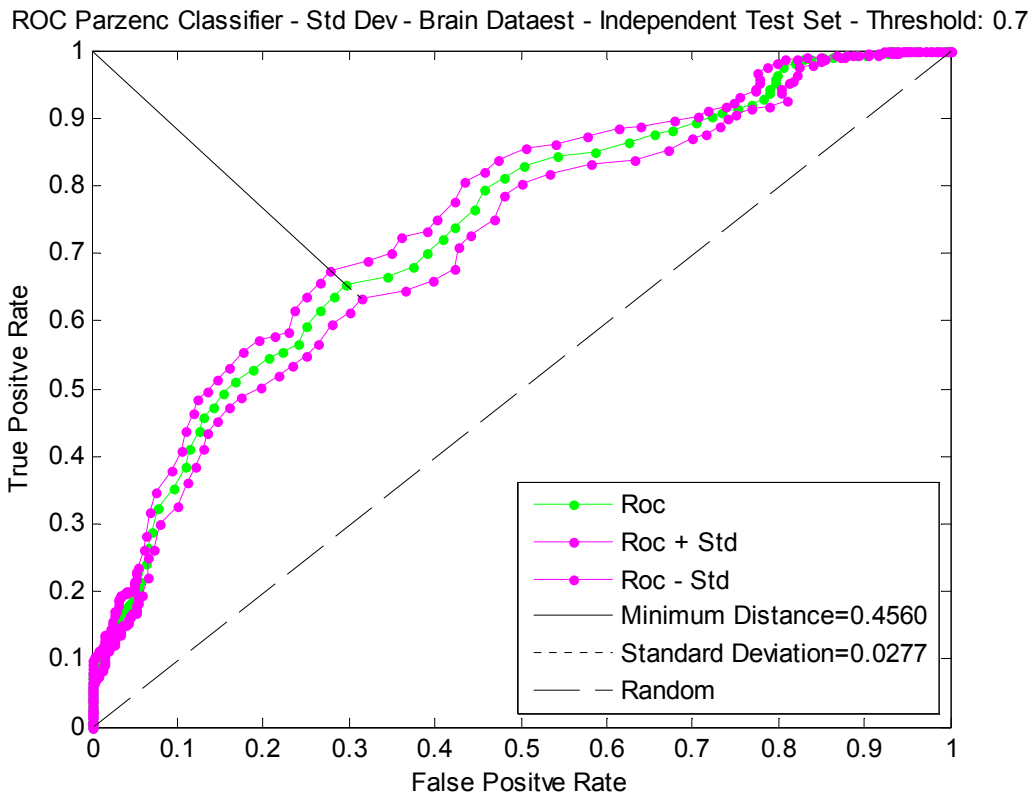
Nmsc


ROC Nmsc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.67
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3903
Standard Deviation=0.1355
Random

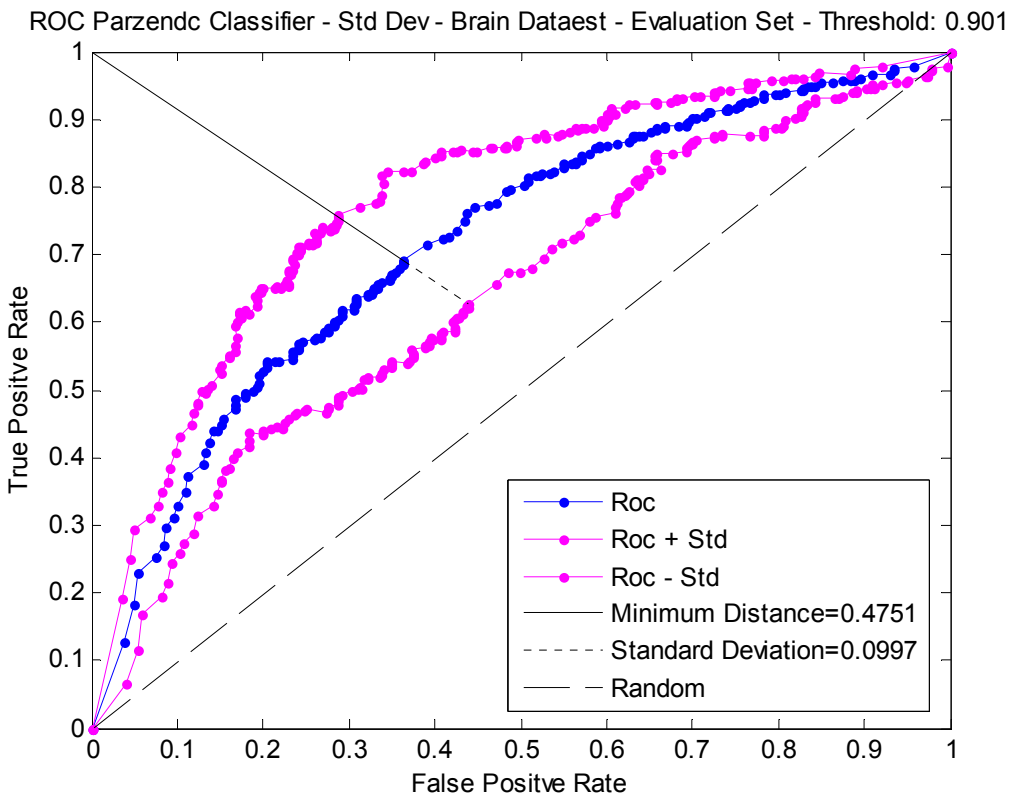

ROC Nmsc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.67
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3739
Standard Deviation=0.0158
Random

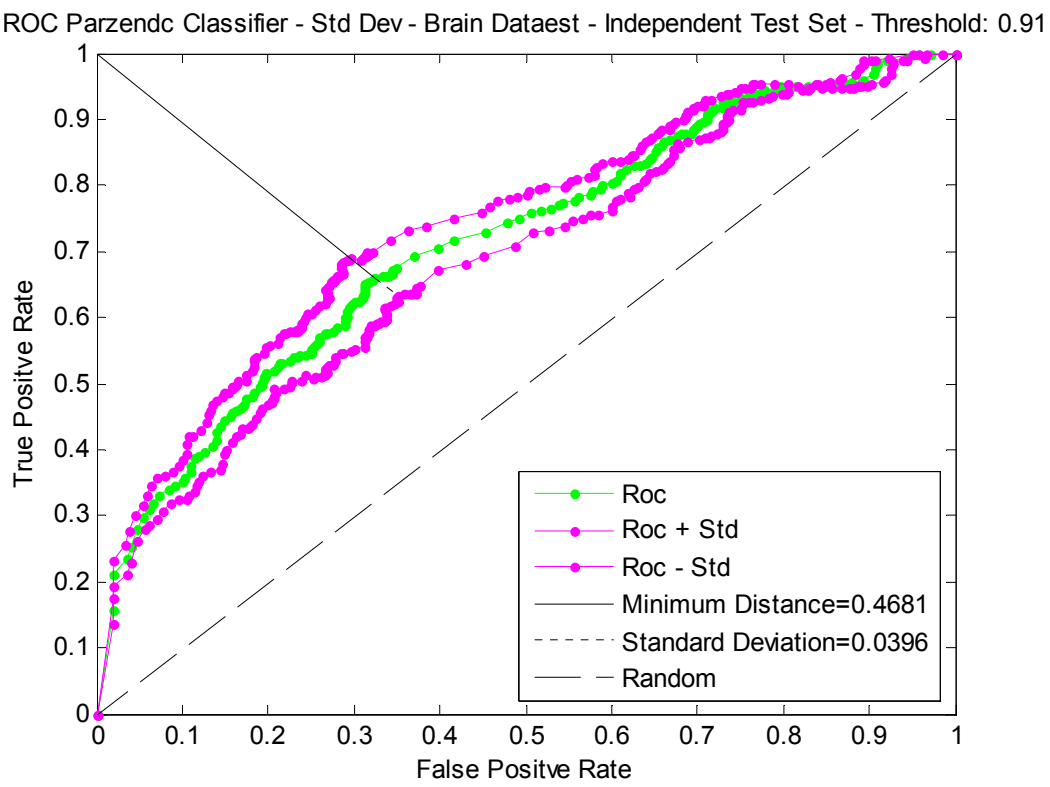
Parzenc


ROC Parzenc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.72
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3920
Standard Deviation=0.1335
Random

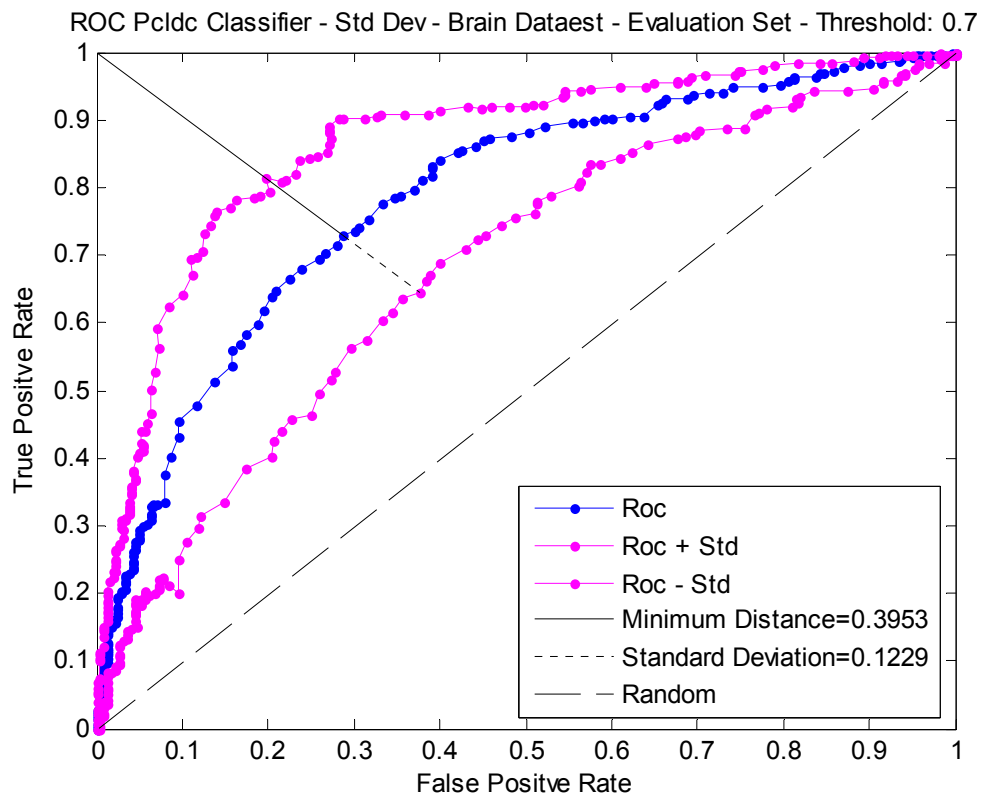

ROC Parzenc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.7
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4560
Standard Deviation=0.0277
Random

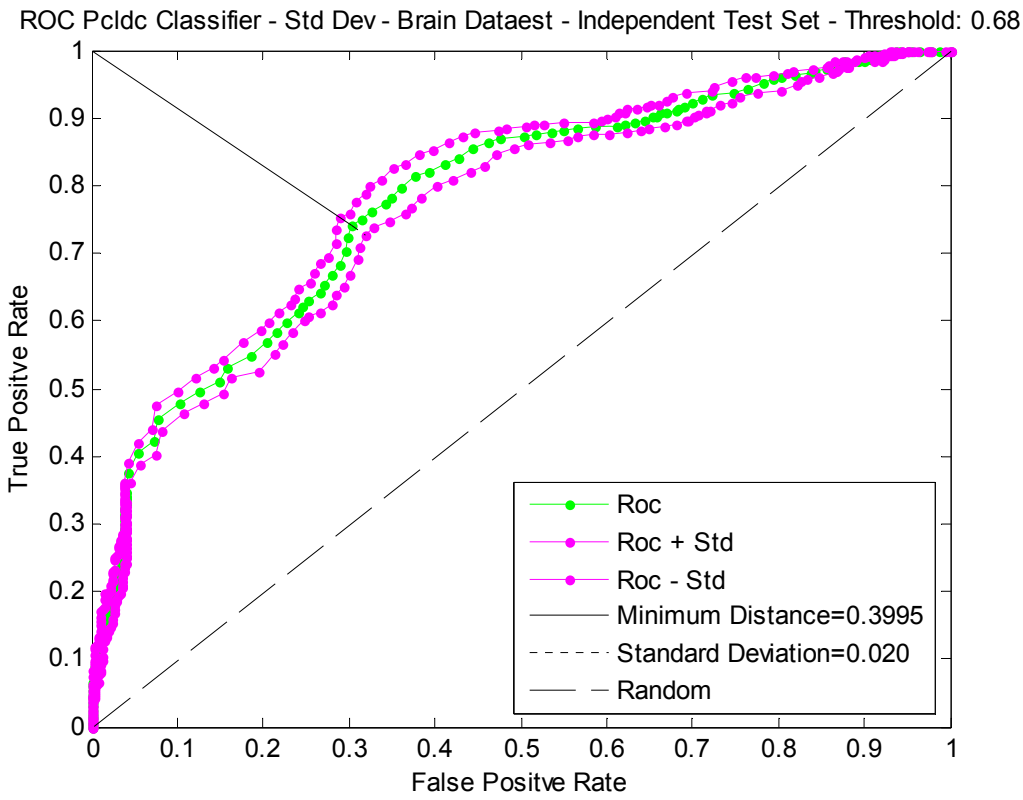
Parzendc


ROC Parzendc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.901
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4751
Standard Deviation=0.0997
Random

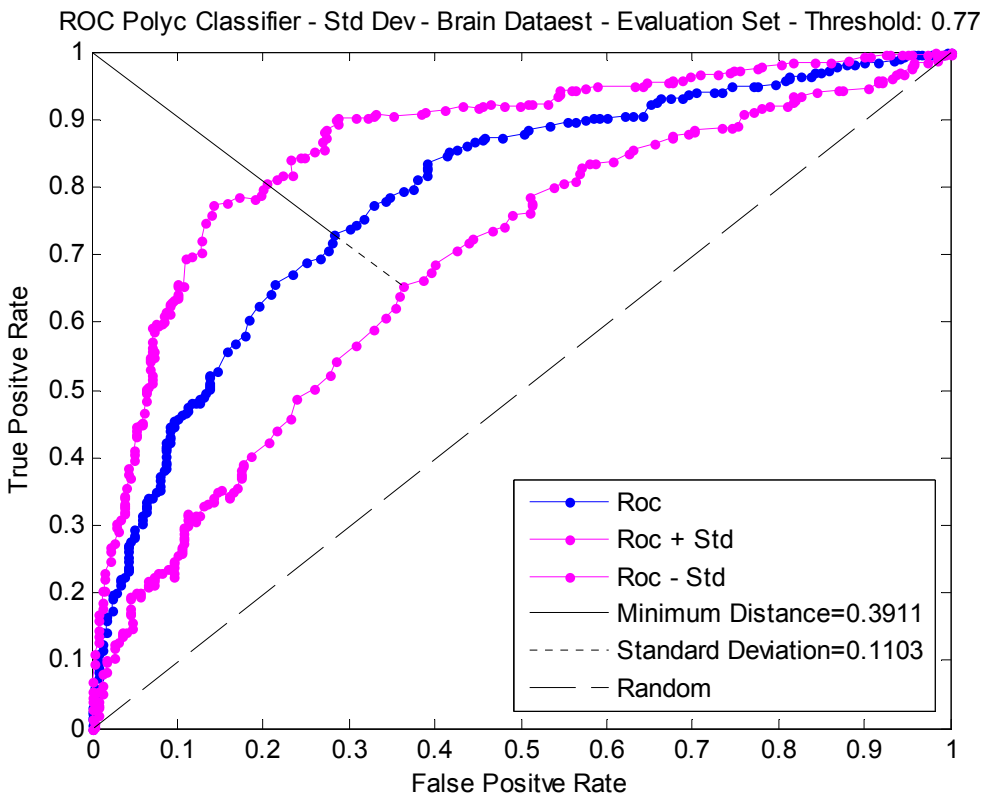

ROC Parzendc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.91
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4681
Standard Deviation=0.0396
Random

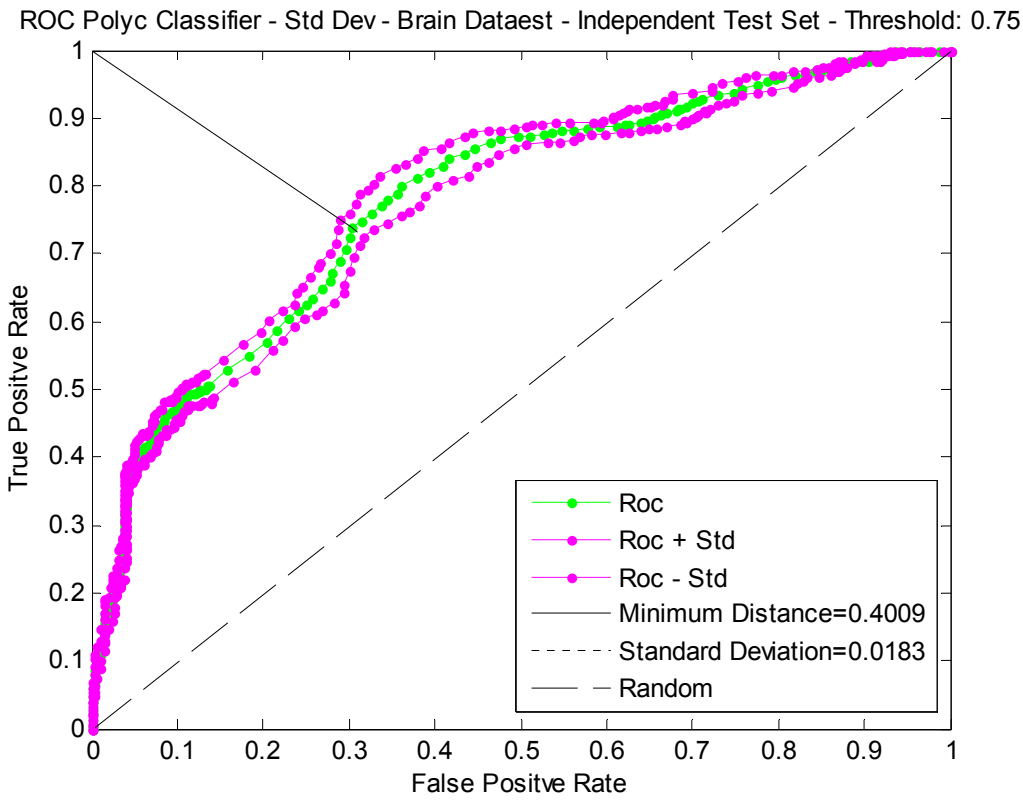
Pcldc


ROC Pcldc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.7
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3953
Standard Deviation=0.1229
Random

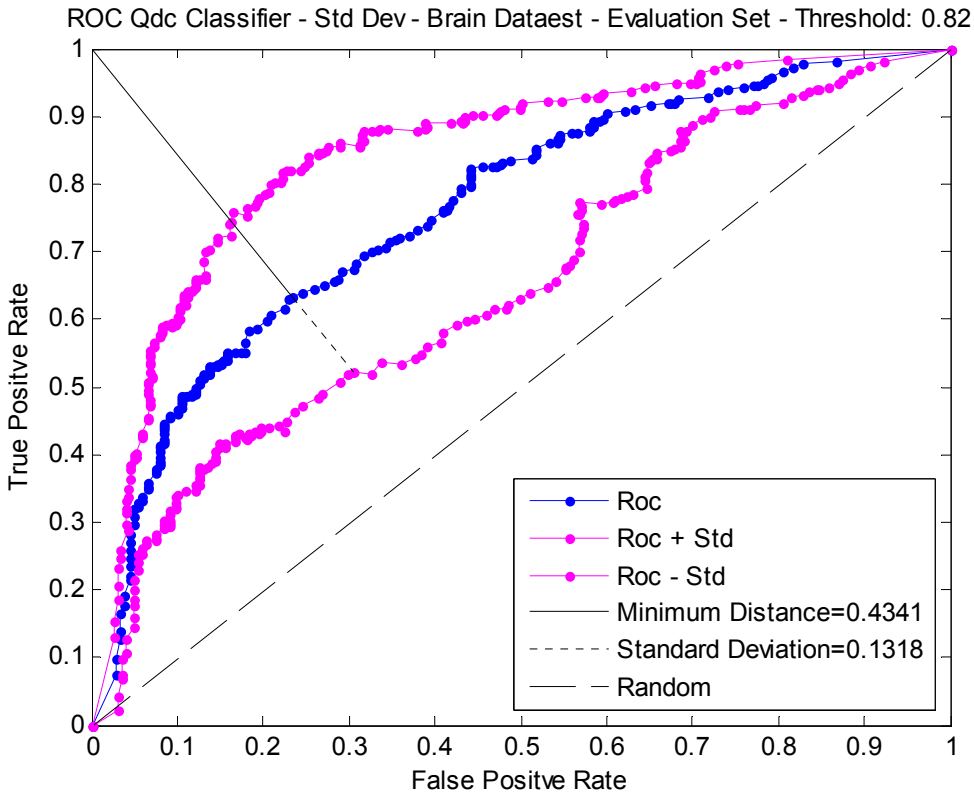

ROC Pcldc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.68
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3995
Standard Deviation=0.020
Random

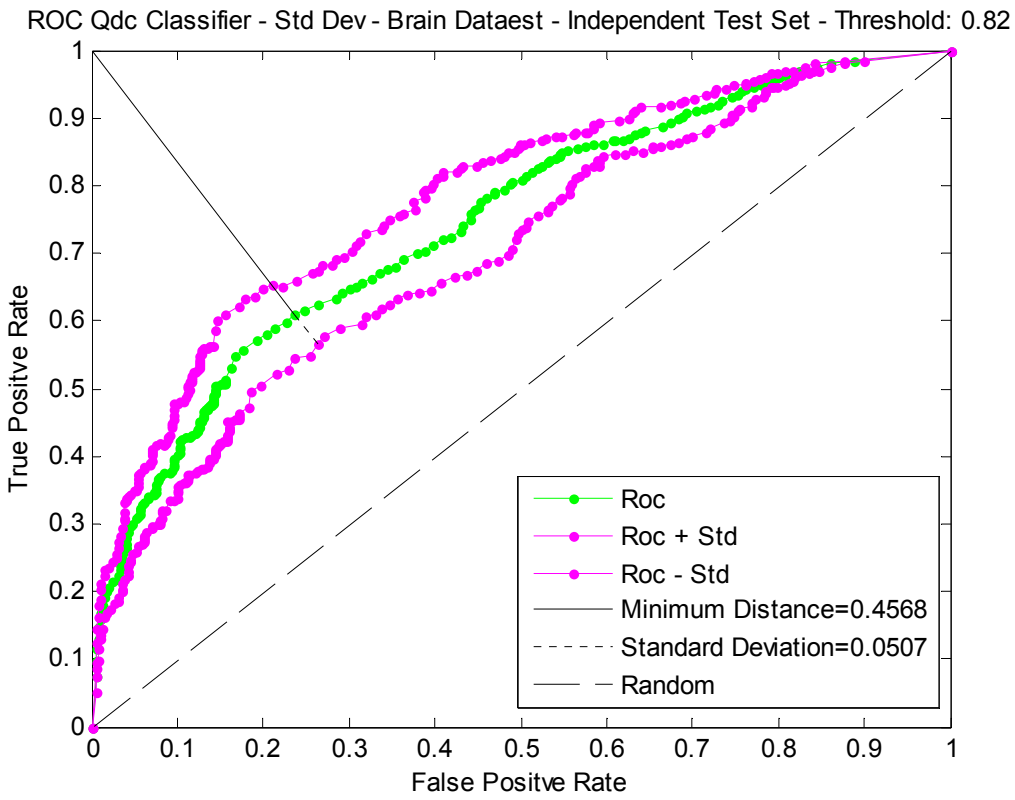
Polyc


ROC Polyc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.77
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.3911
Standard Deviation=0.1103
Random

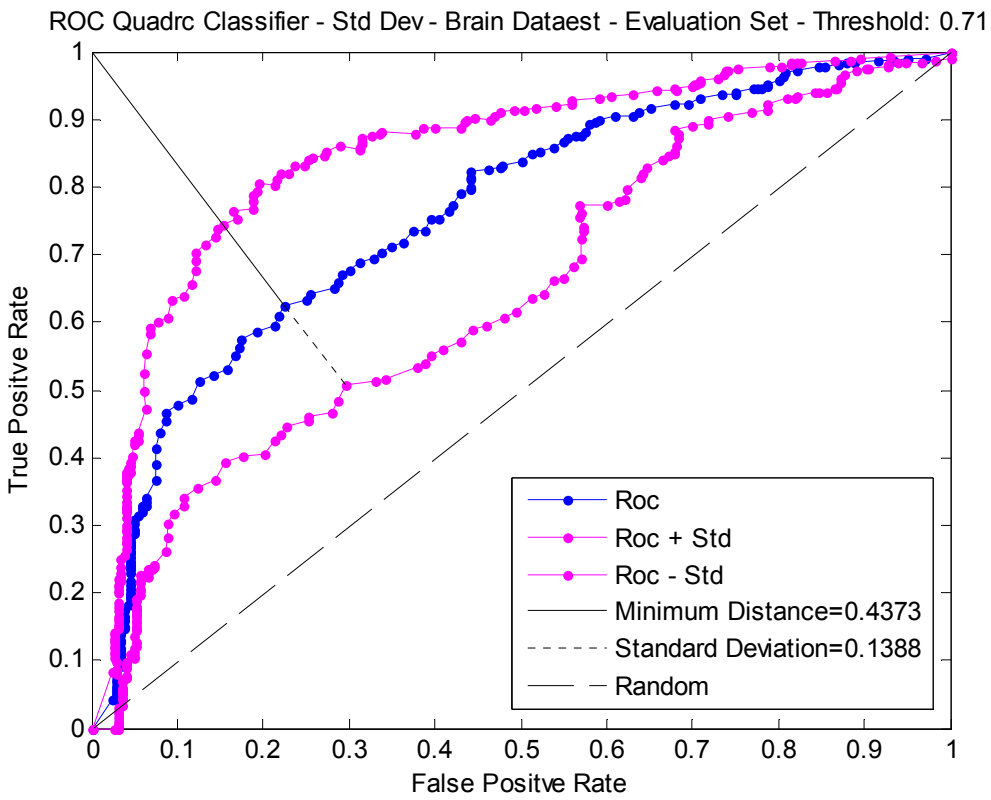

ROC Polyc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.75
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4009
Standard Deviation=0.0183
Random

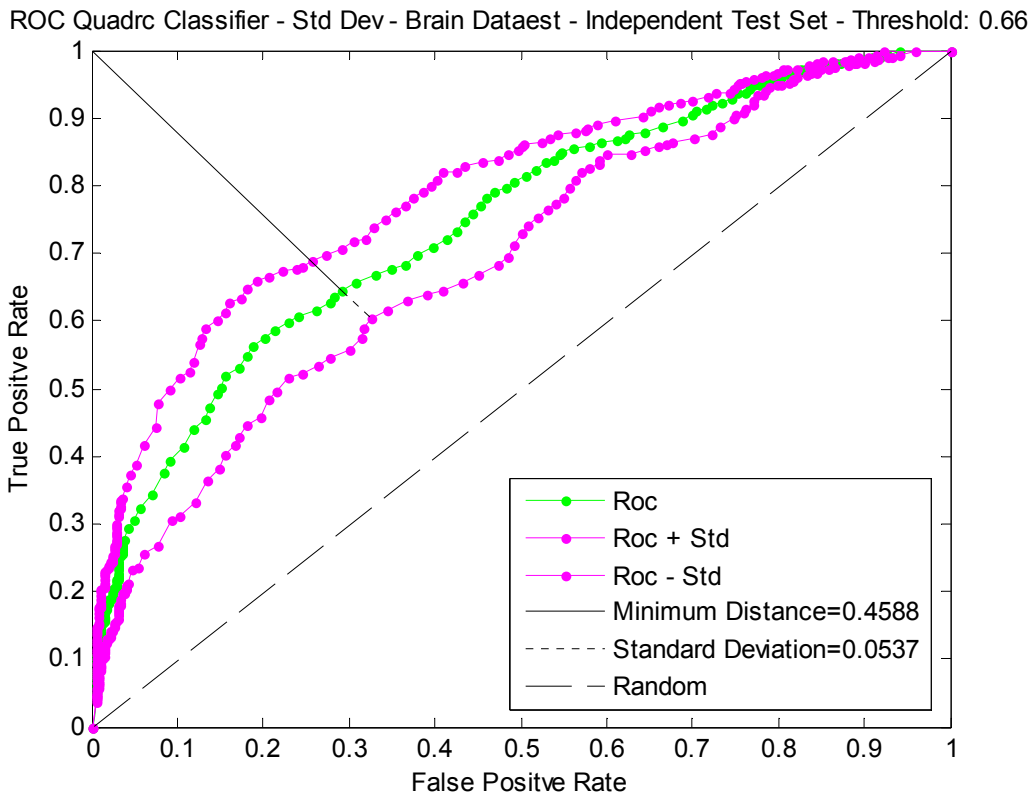
Qdc


ROC Qdc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.82
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4341
Standard Deviation=0.1318
Random

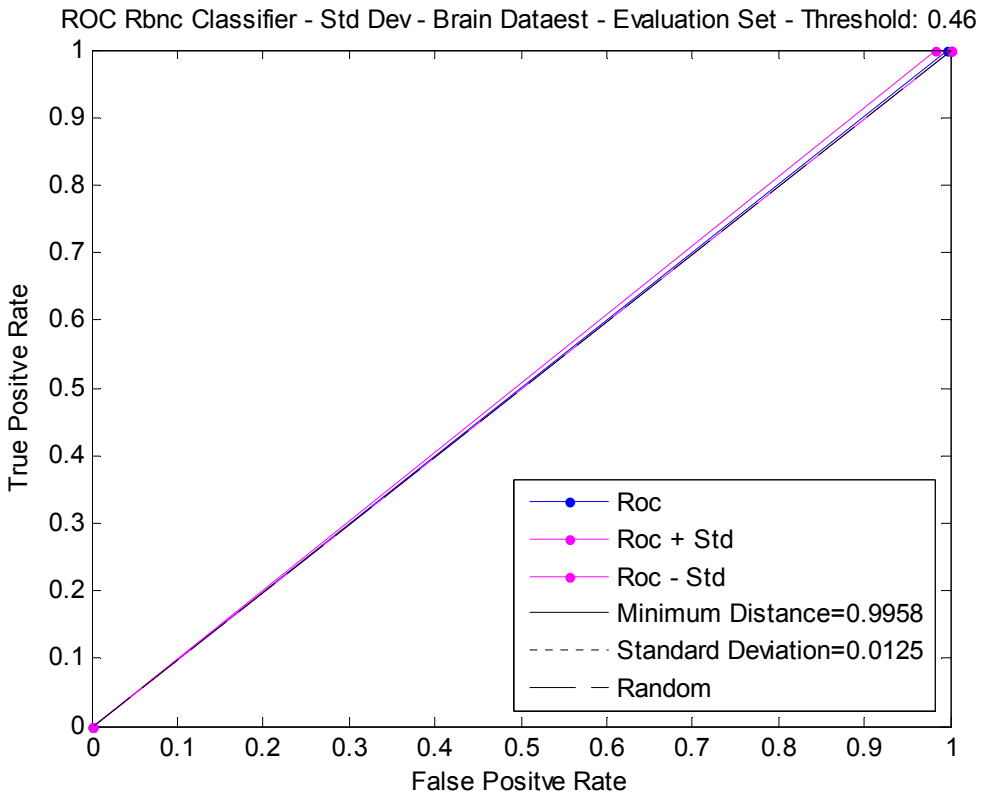

ROC Qdc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.82
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4568
Standard Deviation=0.0507
Random

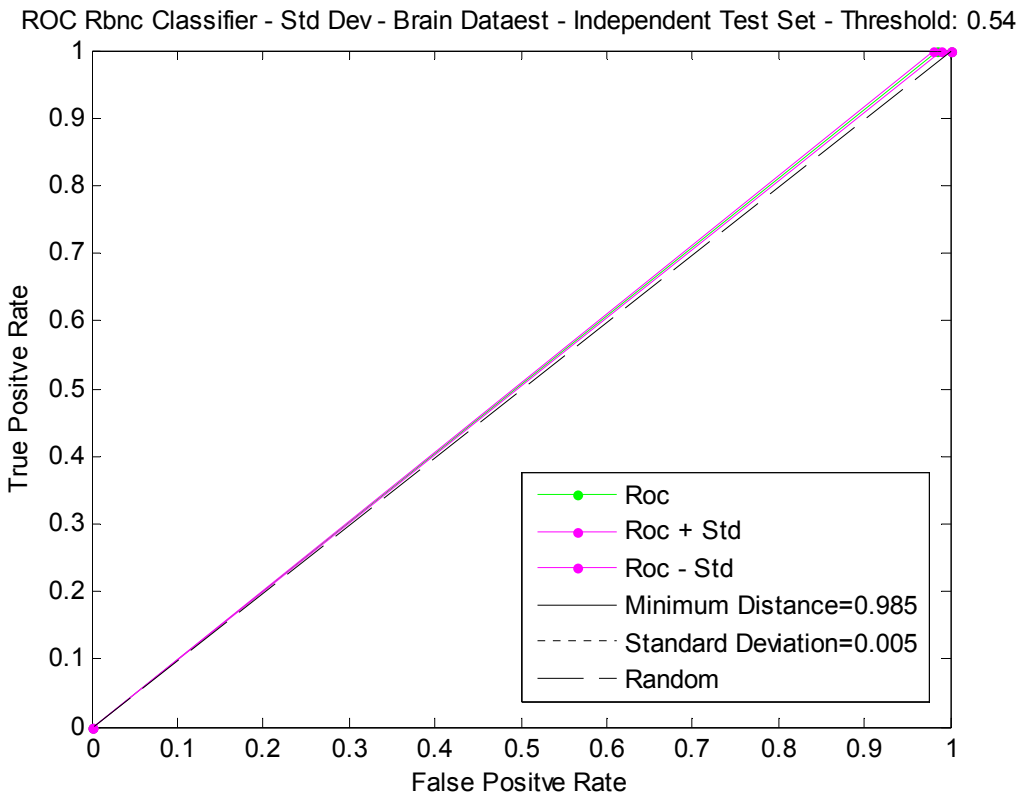
Quadrc


ROC Quadrc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.71
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4373
Standard Deviation=0.1388
Random

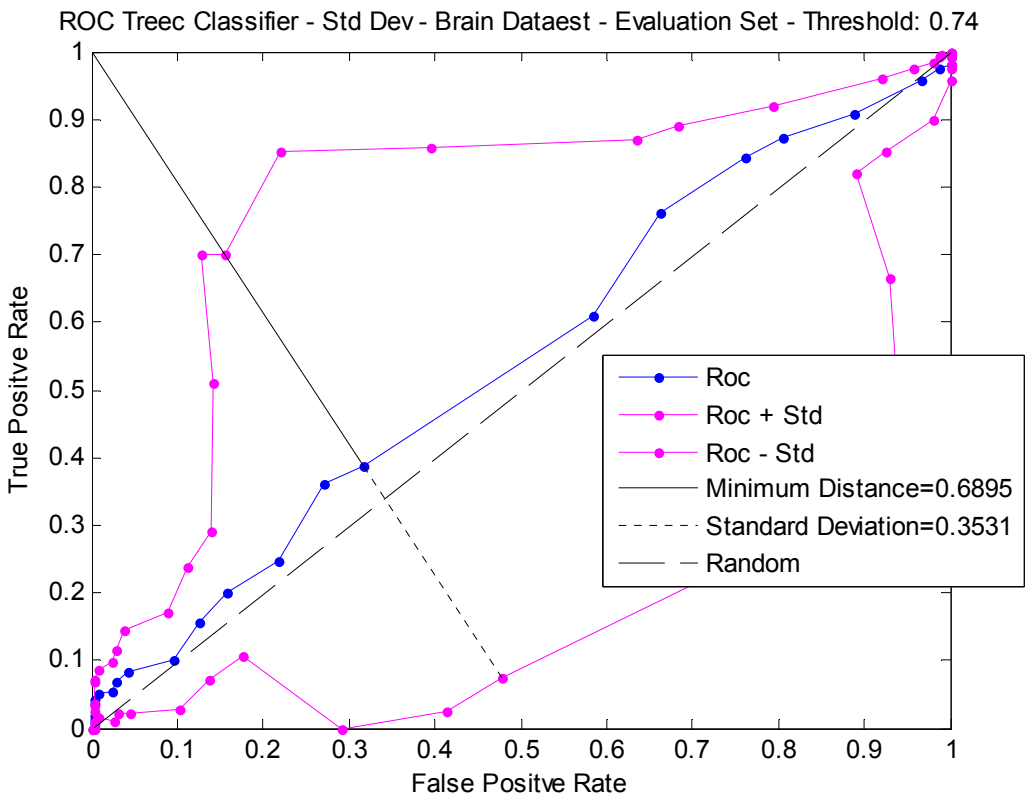

ROC Quadrc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.66
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4588
Standard Deviation=0.0537
Random

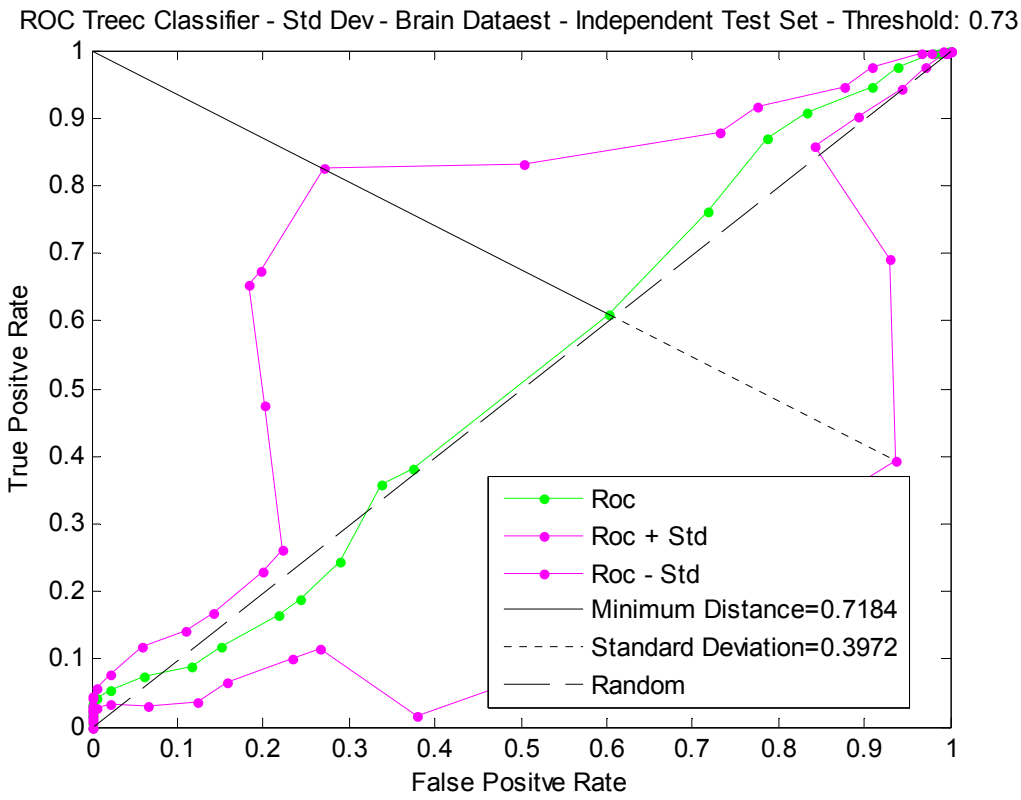
Rbnc


ROC Rbnc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.46
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.9958
Standard Deviation=0.0125
Random

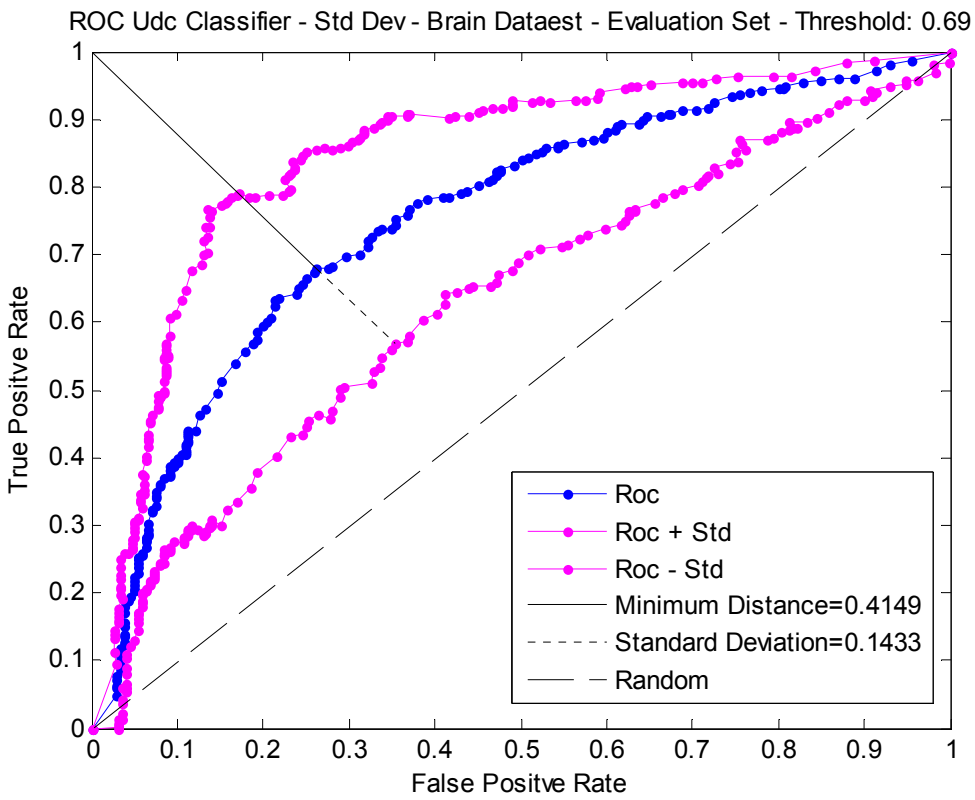

ROC Rbnc Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.54
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.985
Standard Deviation=0.005
Random

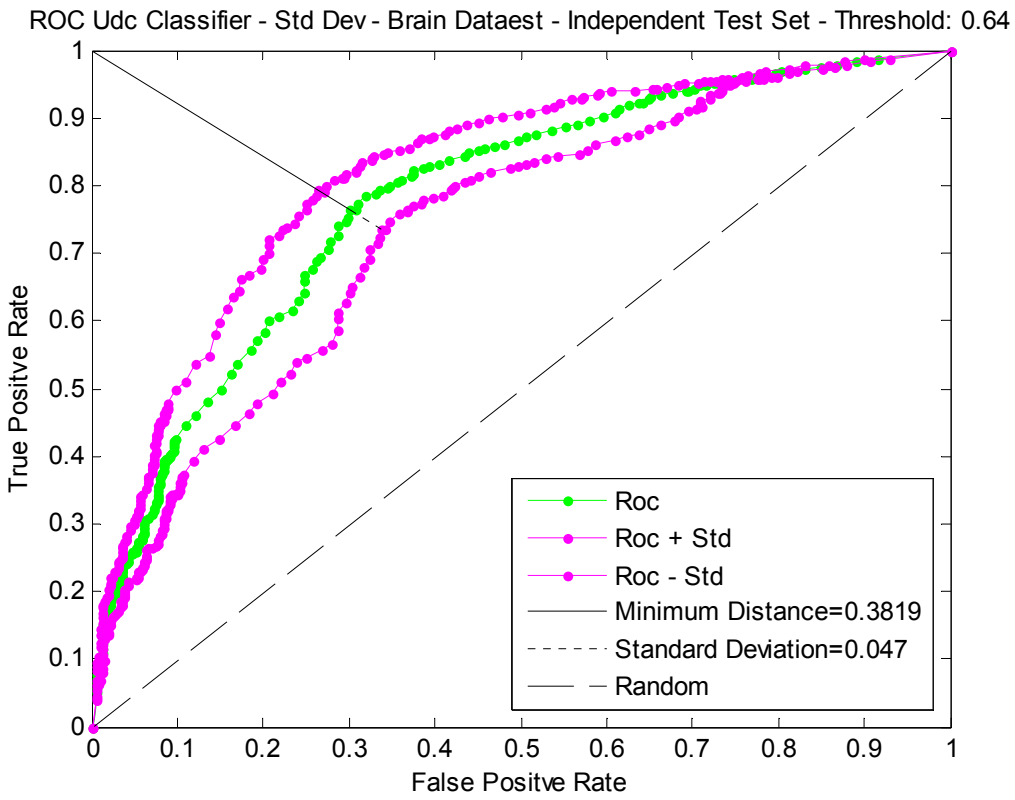
Treec


ROC Treec Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.74
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.6895
Standard Deviation=0.3531
Random


ROC Treec Classifier - Std Dev - Brain Dataest - Independent Test Set - Threshold: 0.73
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.7184
Standard Deviation=0.3972
Random

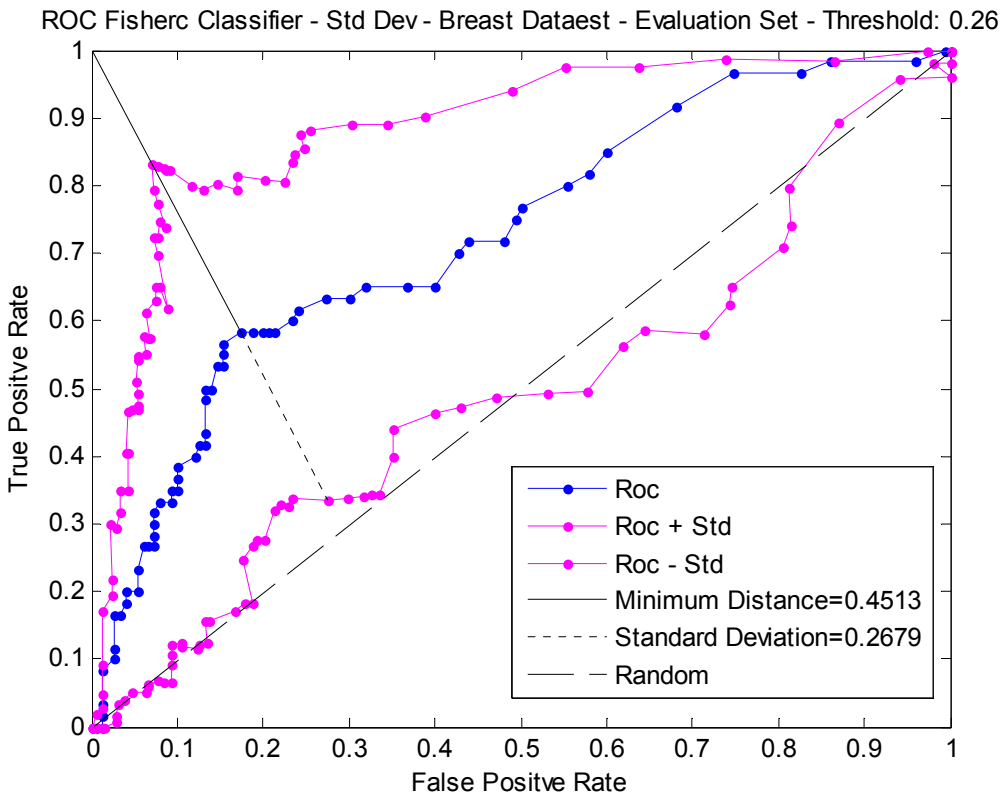
Udc

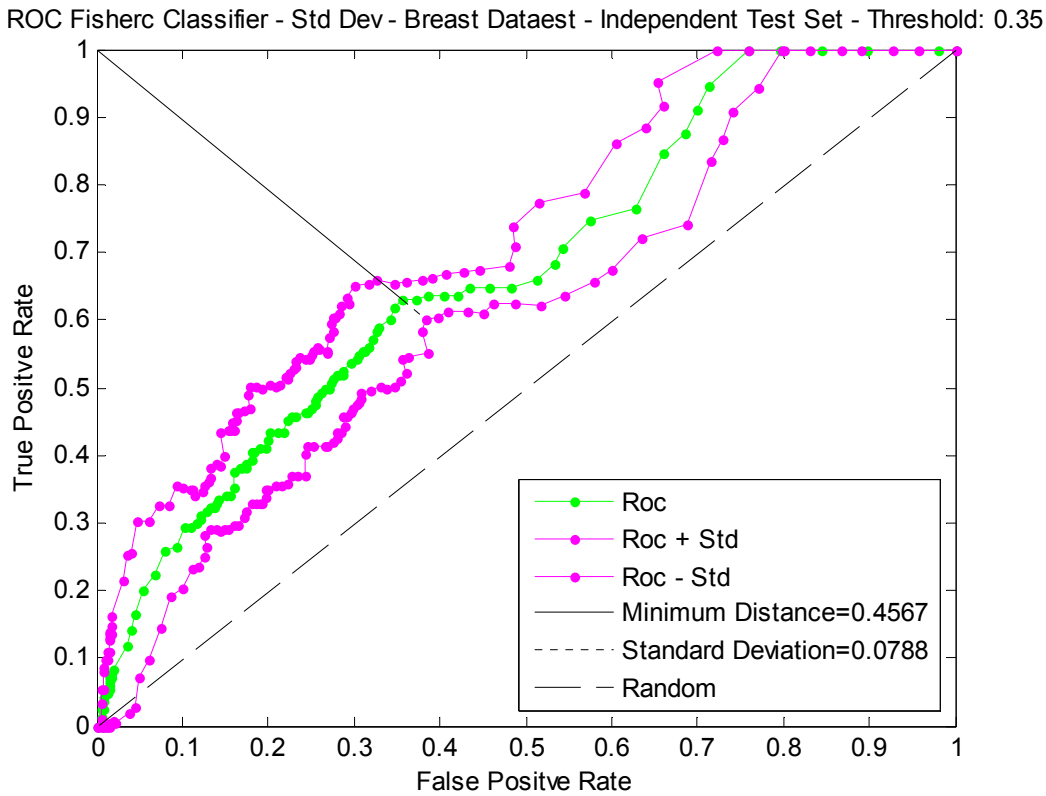

ROC Udc Classifier - Std Dev - Brain Dataest - Evaluation Set - Threshold: 0.69
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4149
Standard Deviation=0.1433
Random



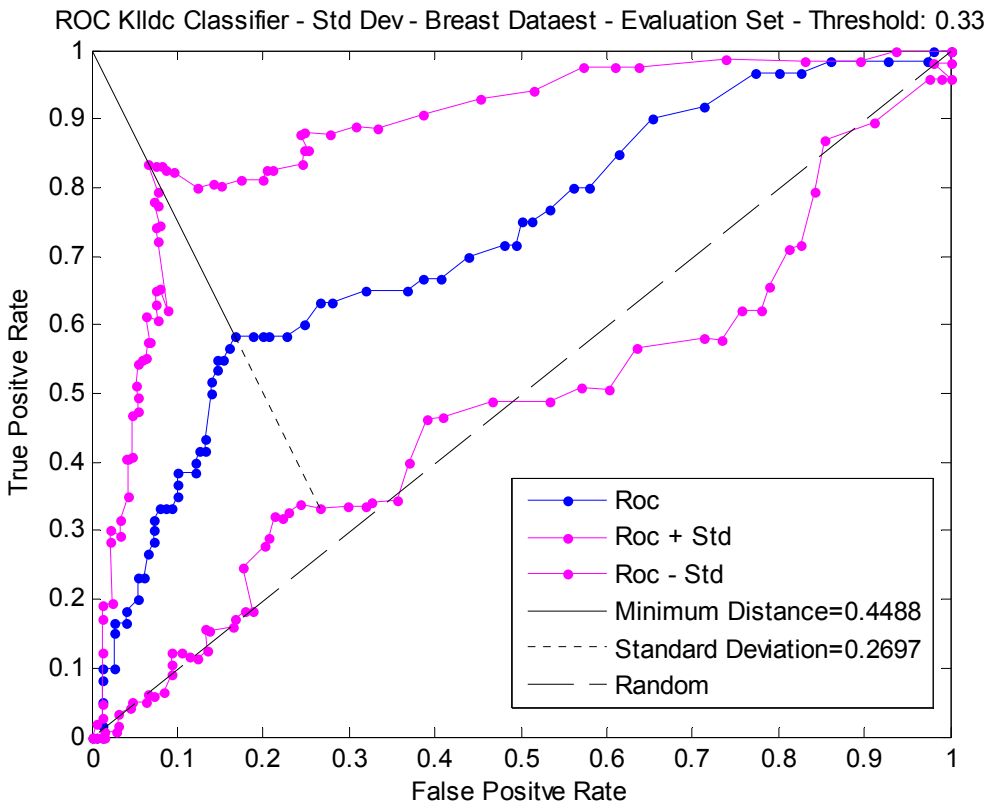

### 8.1.3. Breast Cancer Dataset

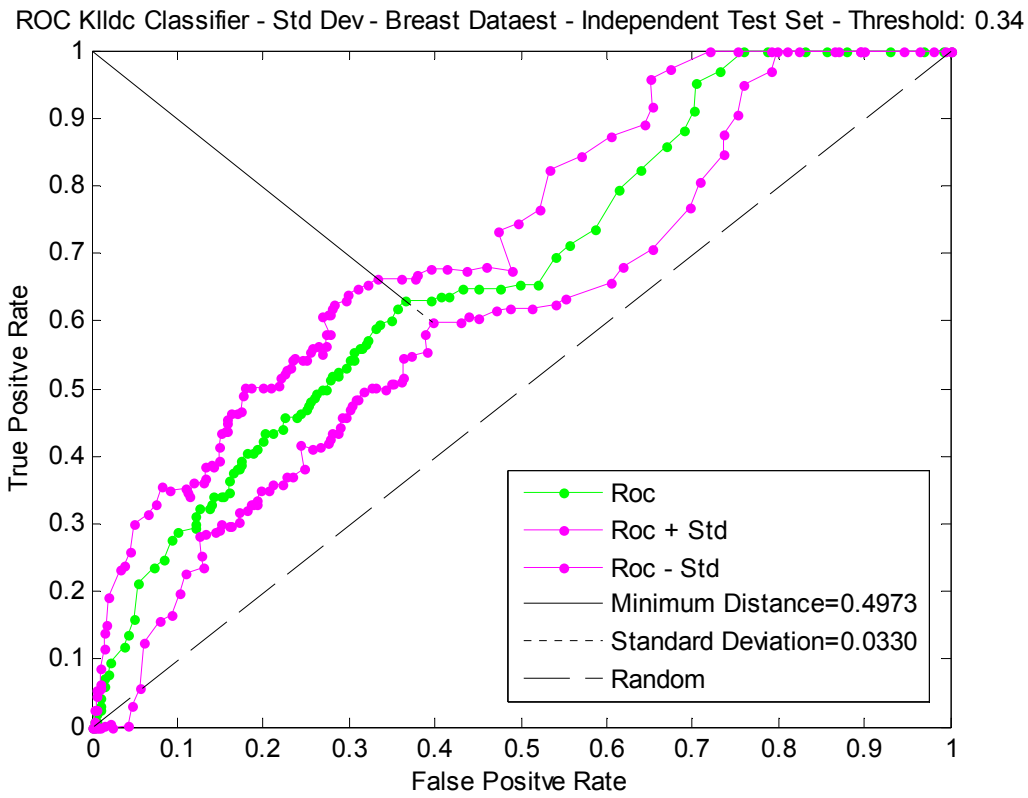
Fisherc


ROC Fisherc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.26
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4513
Standard Deviation=0.2679
Random

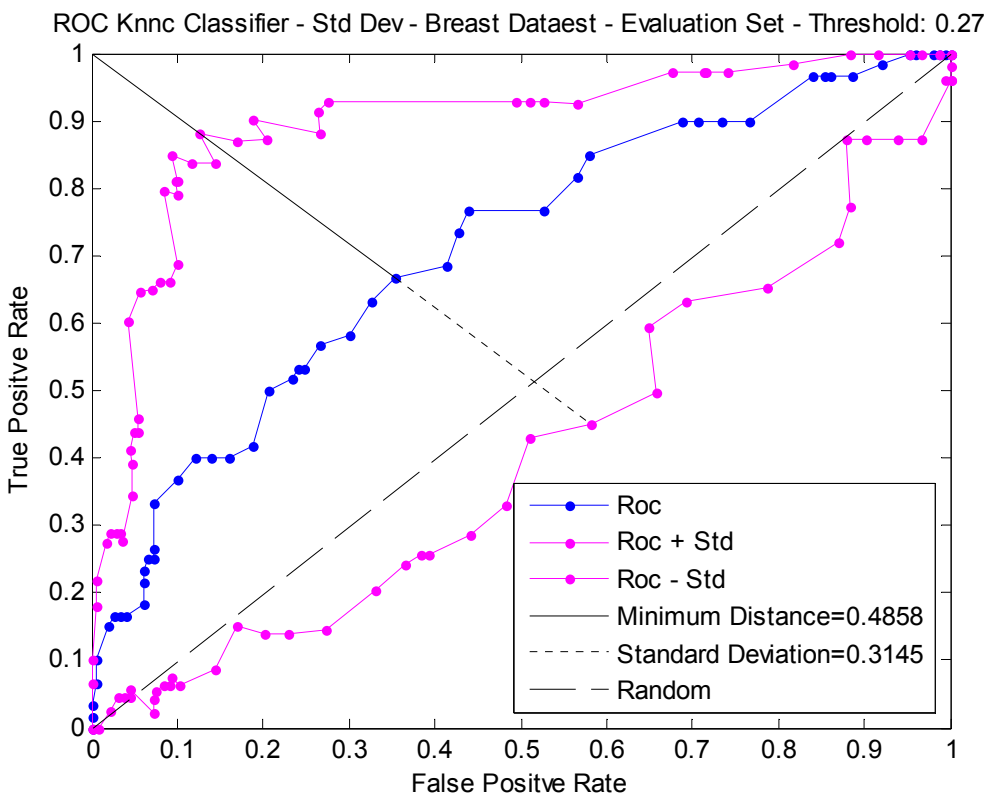

ROC Fisherc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.35
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4567
Standard Deviation=0.0788
Random

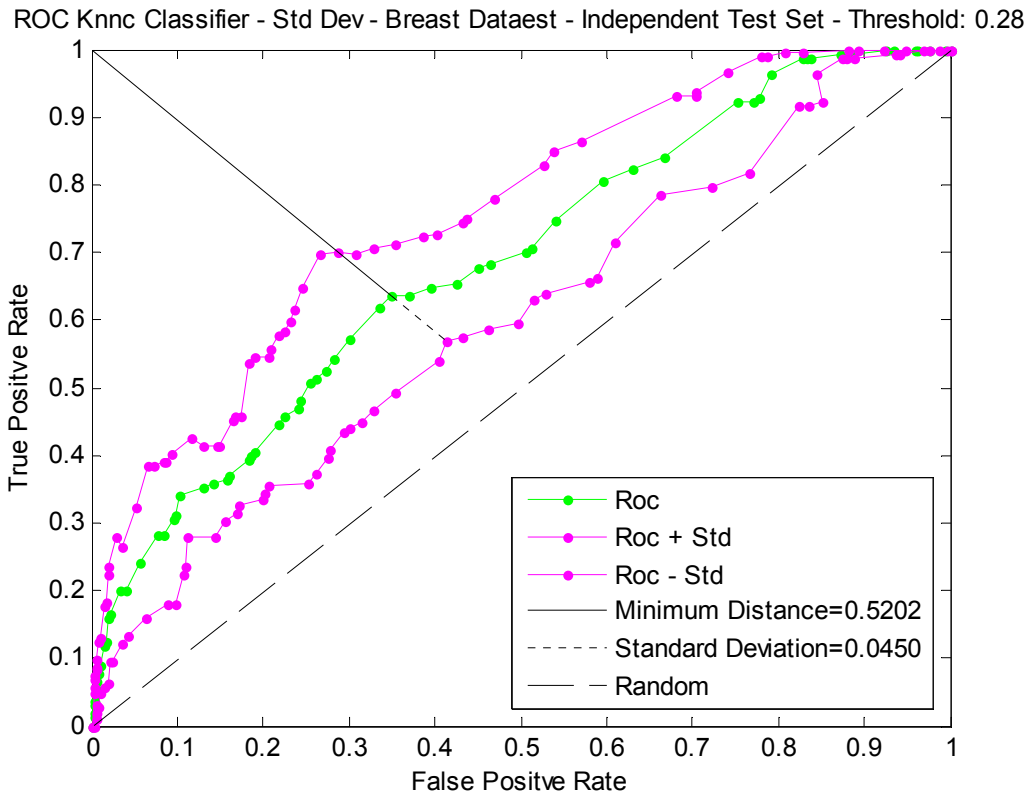
Klldc


ROC Klldc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.33
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4488
Standard Deviation=0.2697
Random

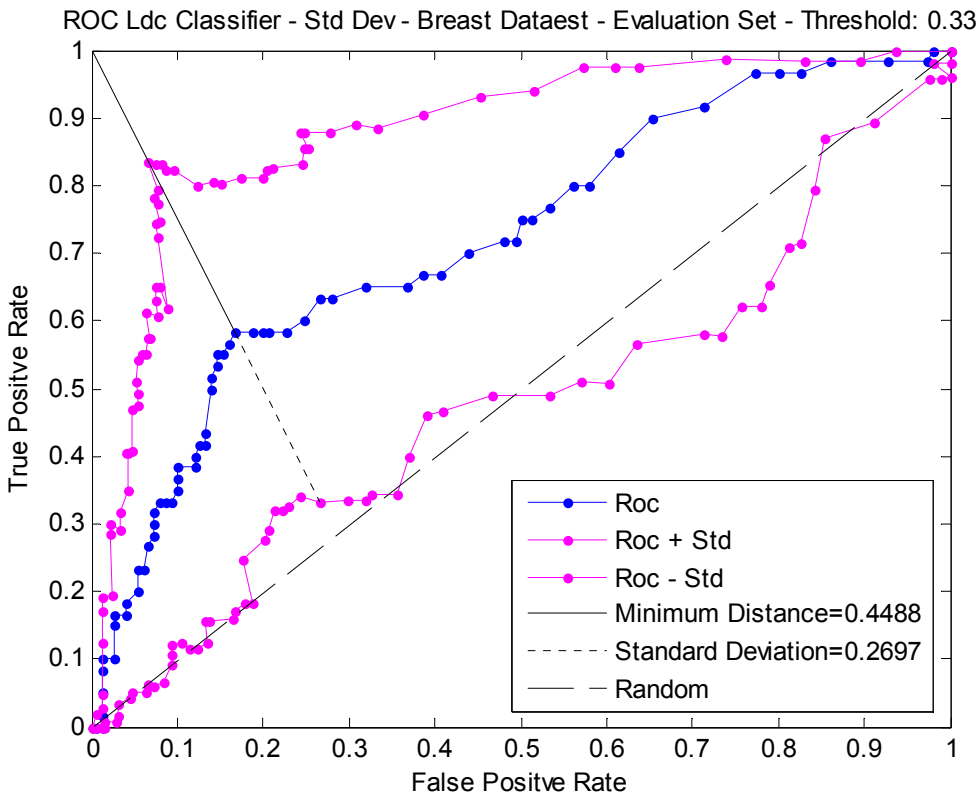

ROC Klldc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.34
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4973
Standard Deviation=0.0330
Random

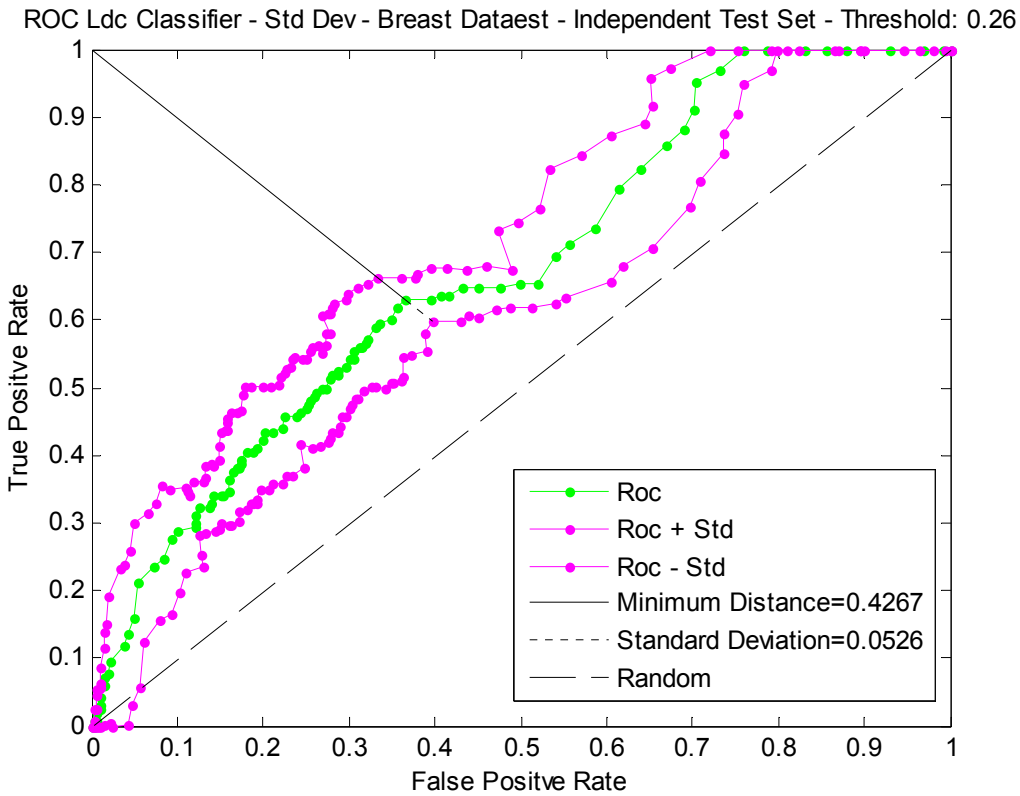
Knnc


ROC Knnc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.27
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4858
Standard Deviation=0.3145
Random

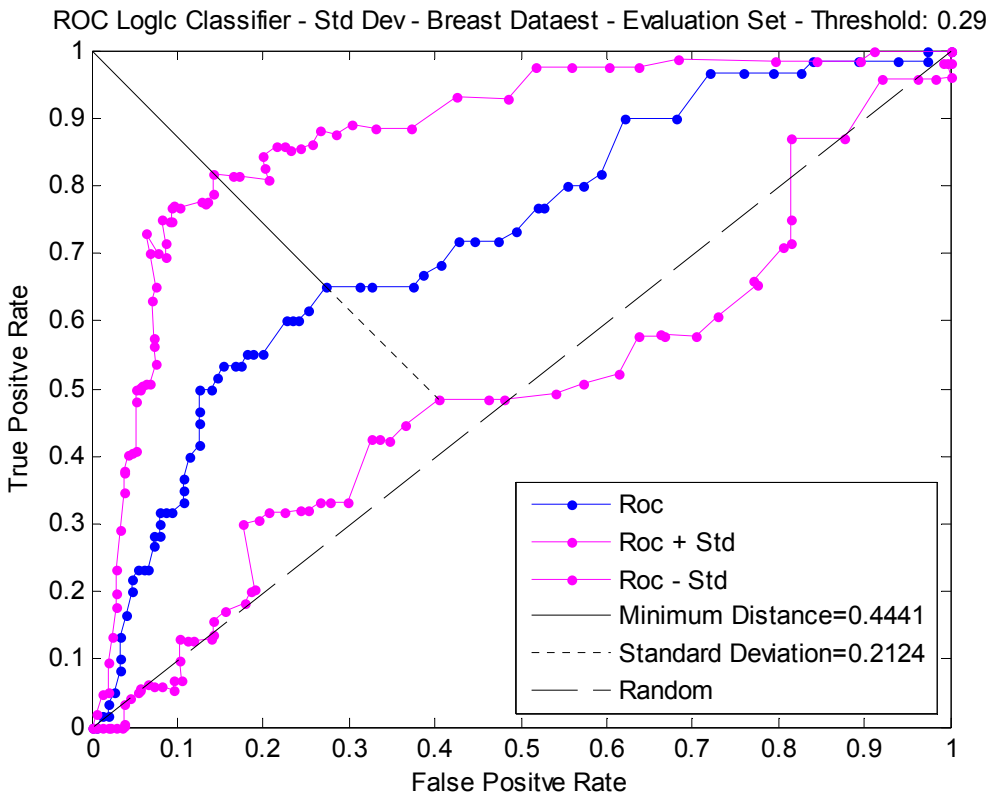

ROC Knnc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.28
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5202
Standard Deviation=0.0450
Random

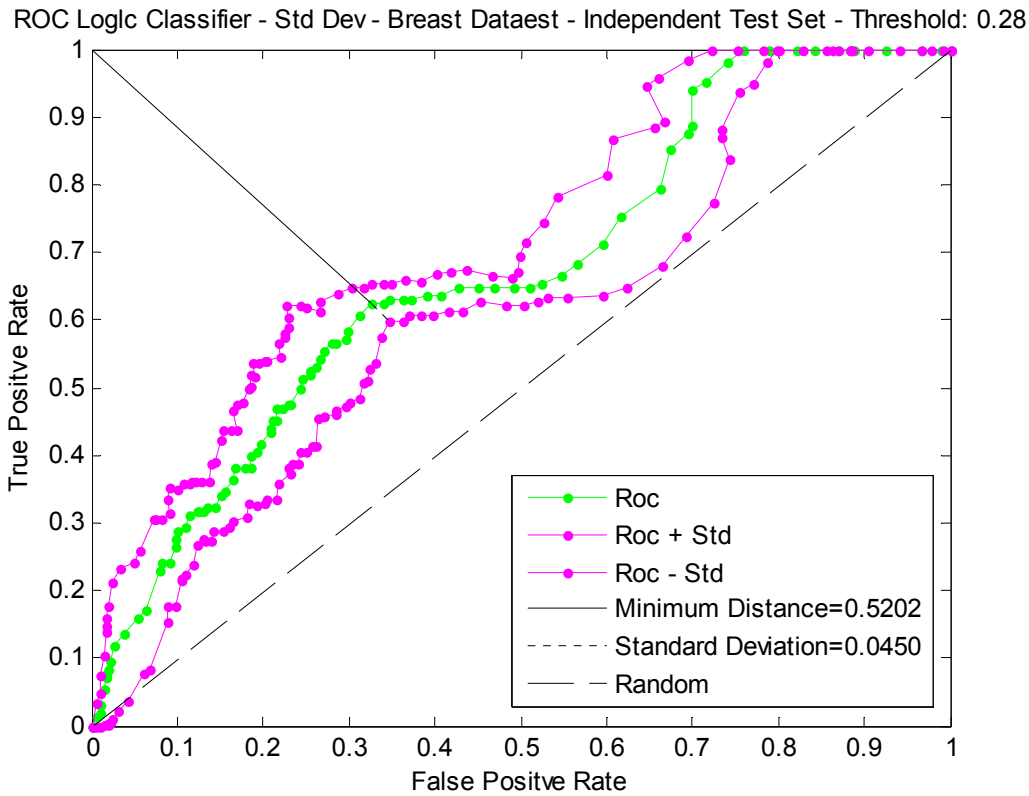
Ldc


ROC Ldc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.33
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4488
Standard Deviation=0.2697
Random

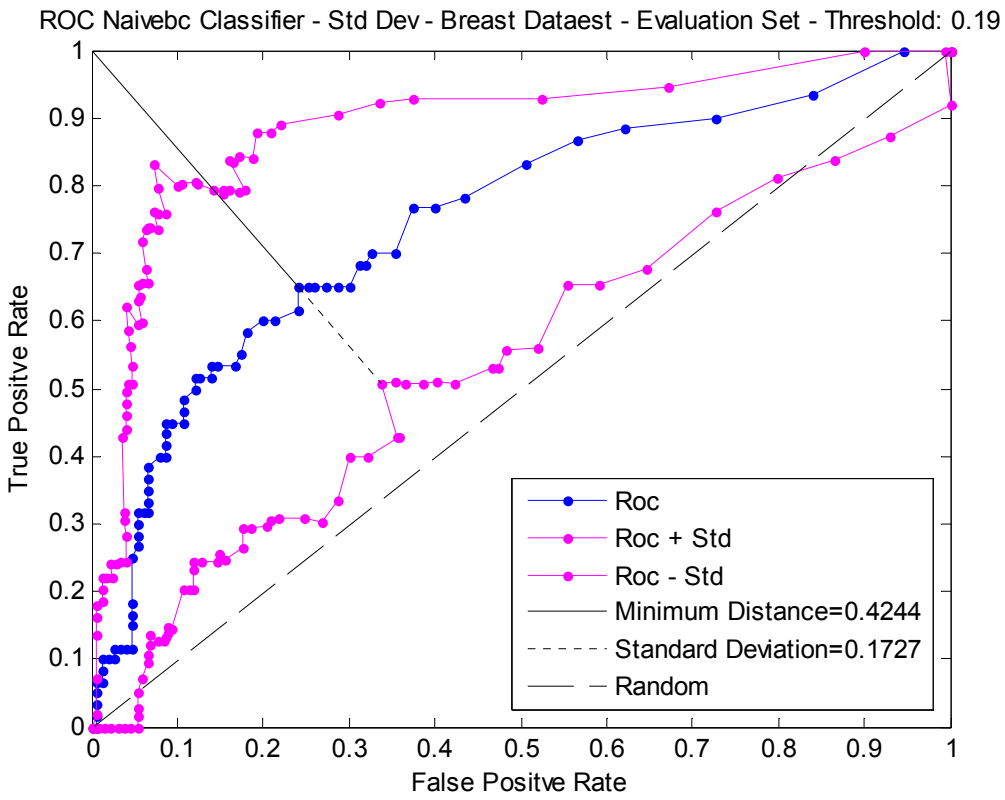

ROC Ldc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.26
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4267
Standard Deviation=0.0526
Random

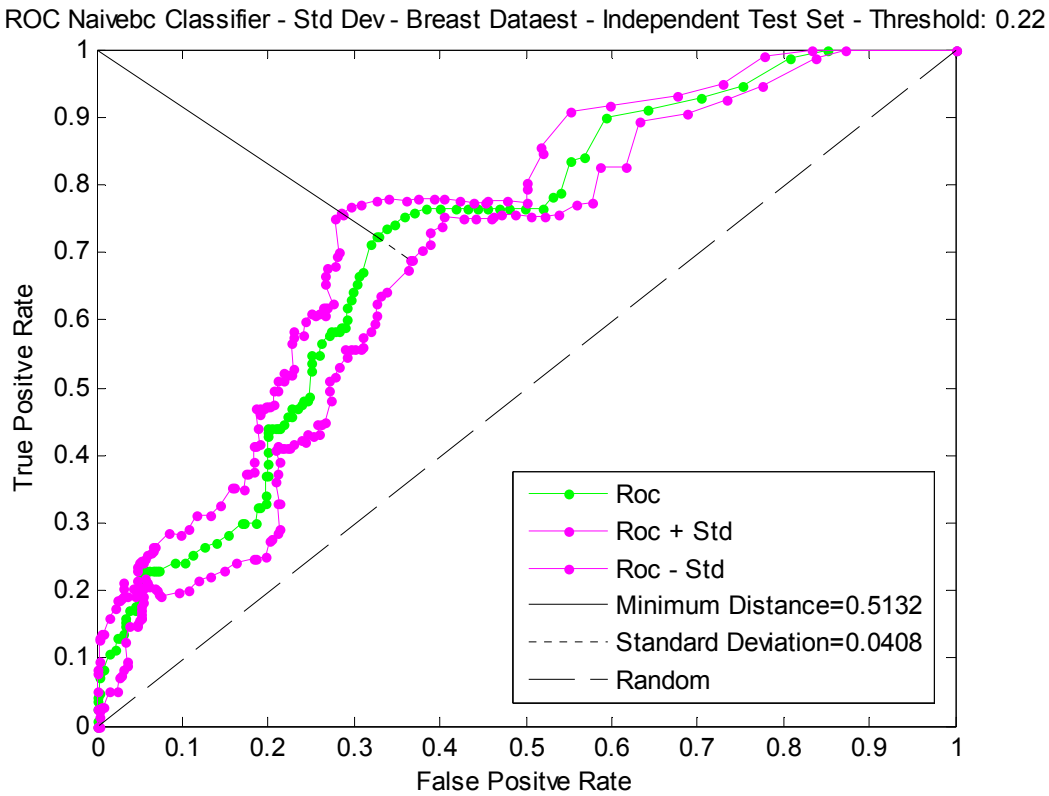
Loglc


ROC Loglc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.29
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4441
Standard Deviation=0.2124
Random

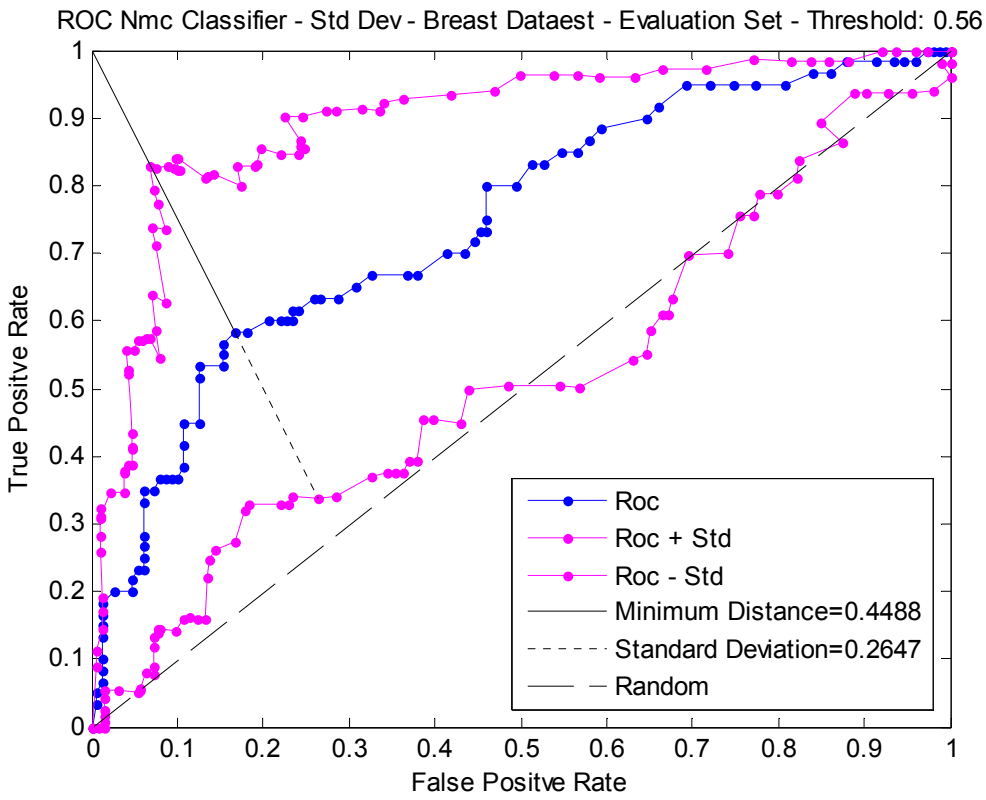

ROC Loglc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.28
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5202
Standard Deviation=0.0450
Random

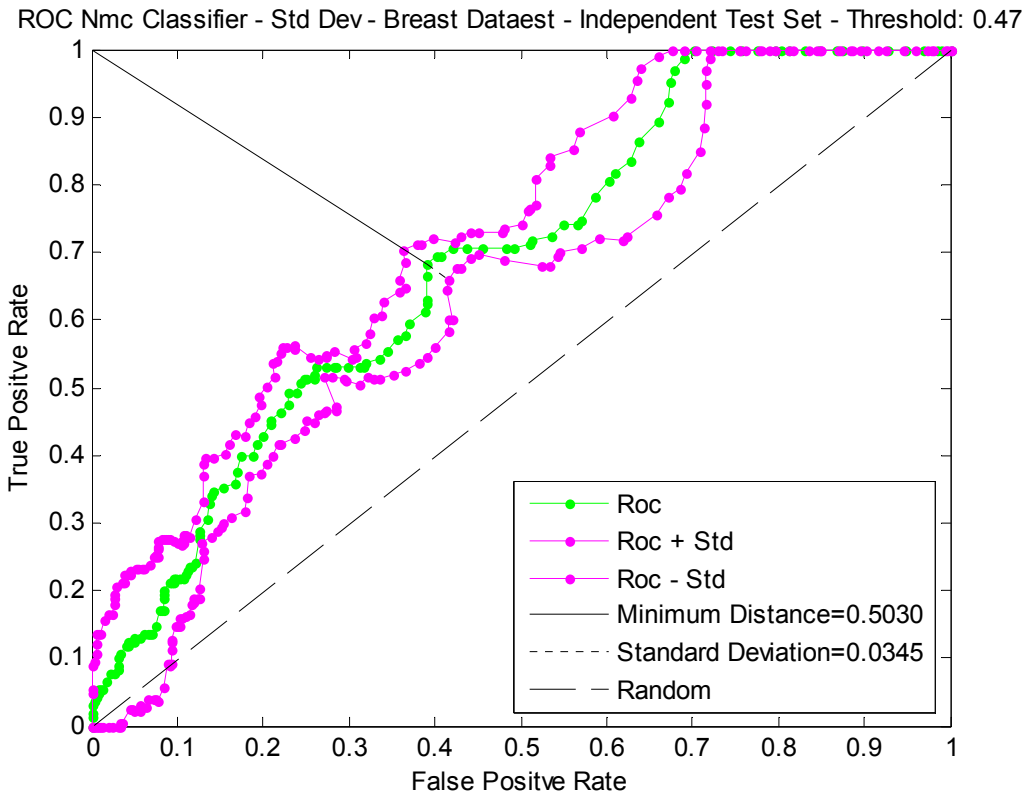
Naivebc


ROC Naivebc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.19
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4244
Standard Deviation=0.1727
Random

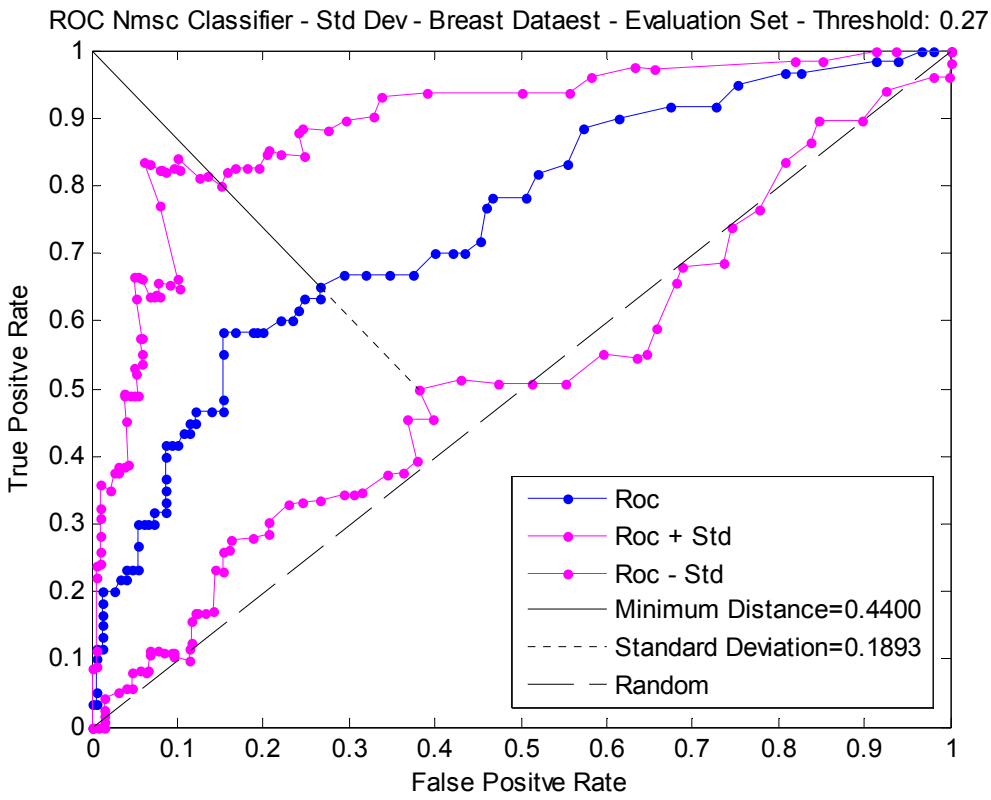

ROC Naivebc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.22
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5132
Standard Deviation=0.0408
Random

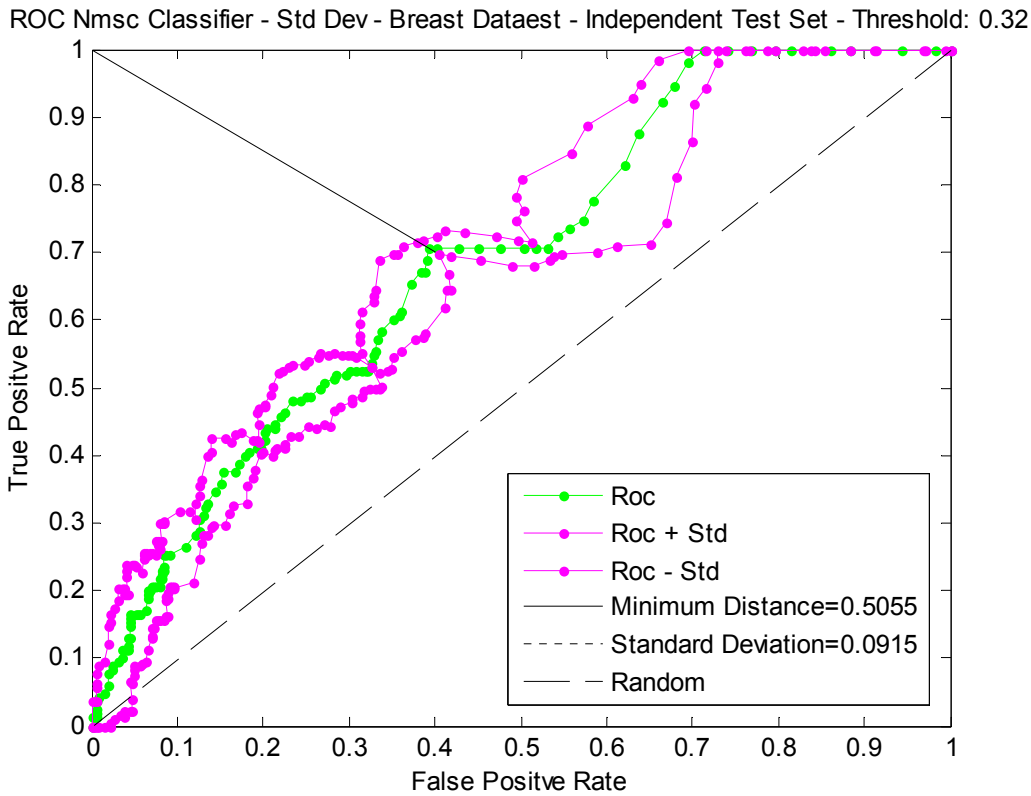
Nmc


ROC Nmc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.56
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4488
Standard Deviation=0.2647
Random

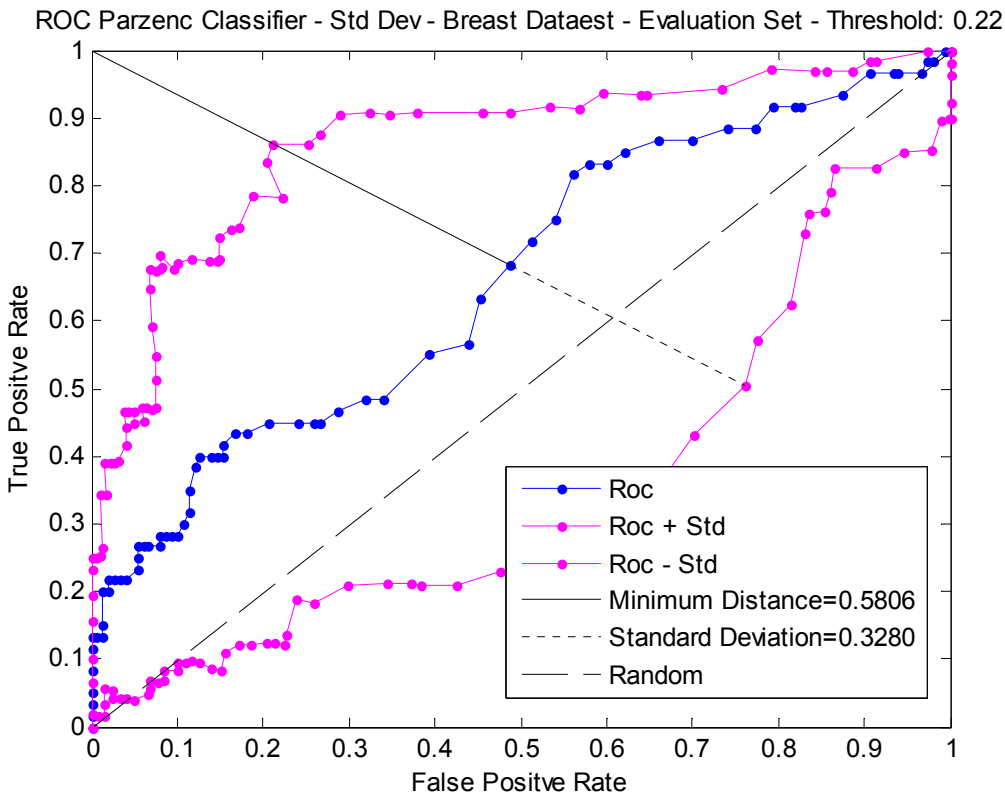

ROC Nmc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.47
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5030
Standard Deviation=0.0345
Random

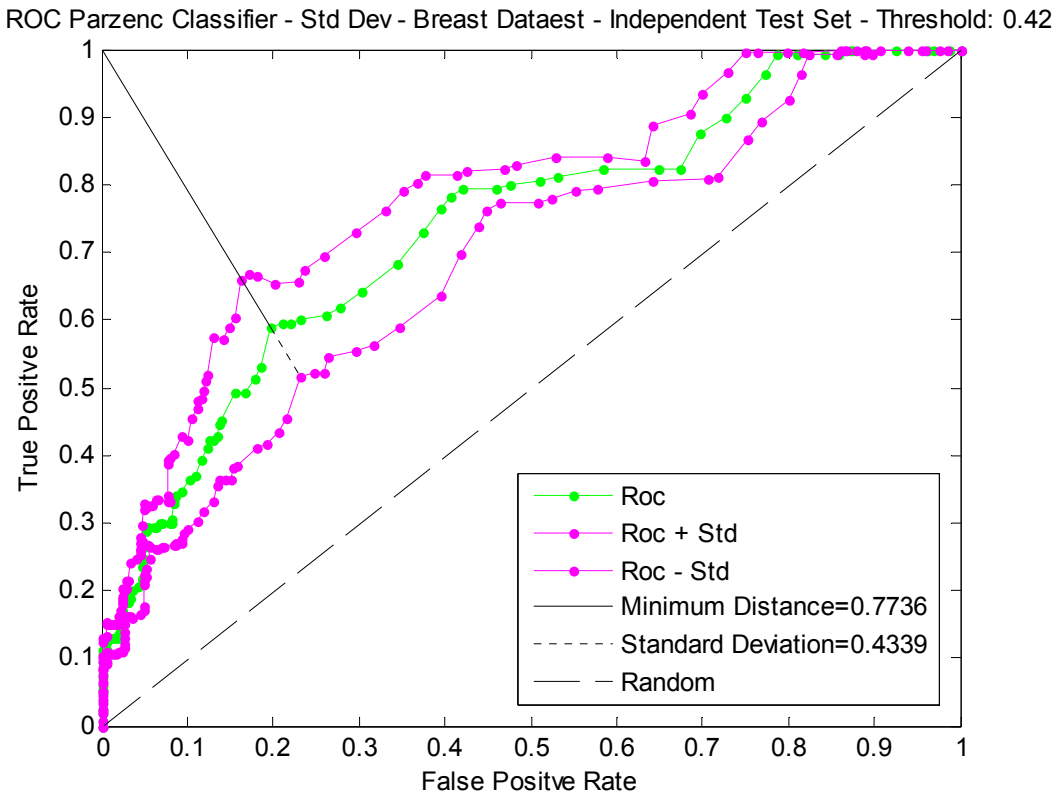
Nmsc


ROC Nmsc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.27
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4400
Standard Deviation=0.1893
Random

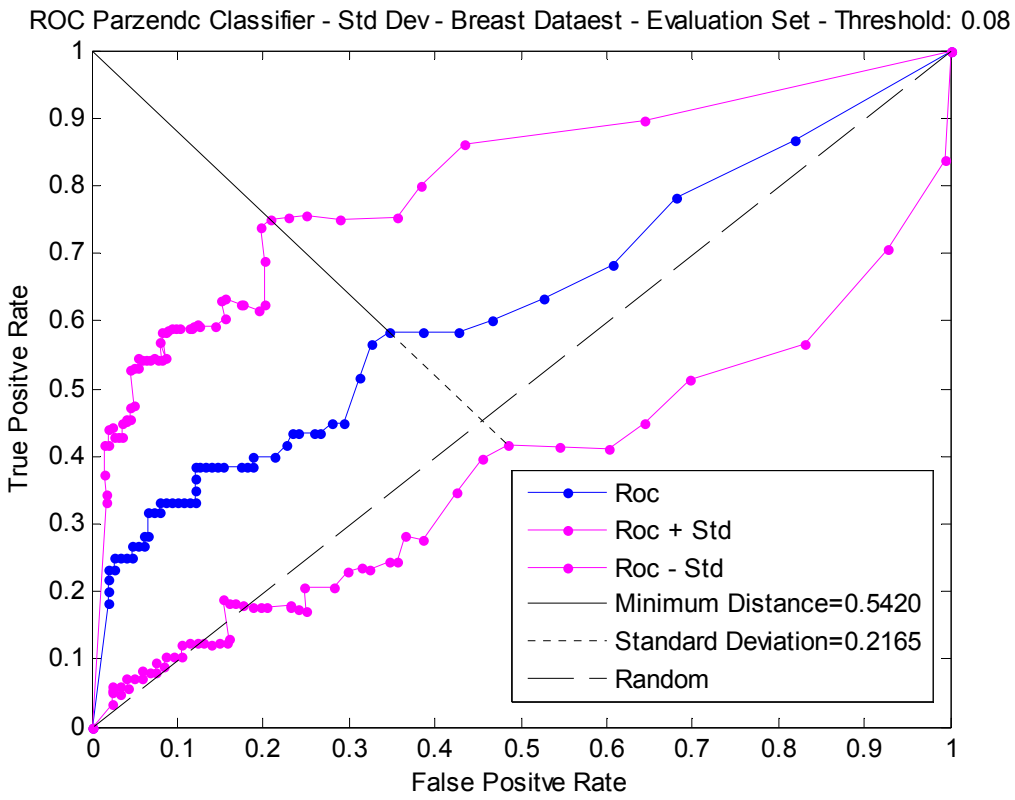

ROC Nmsc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.32
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5055
Standard Deviation=0.0915
Random

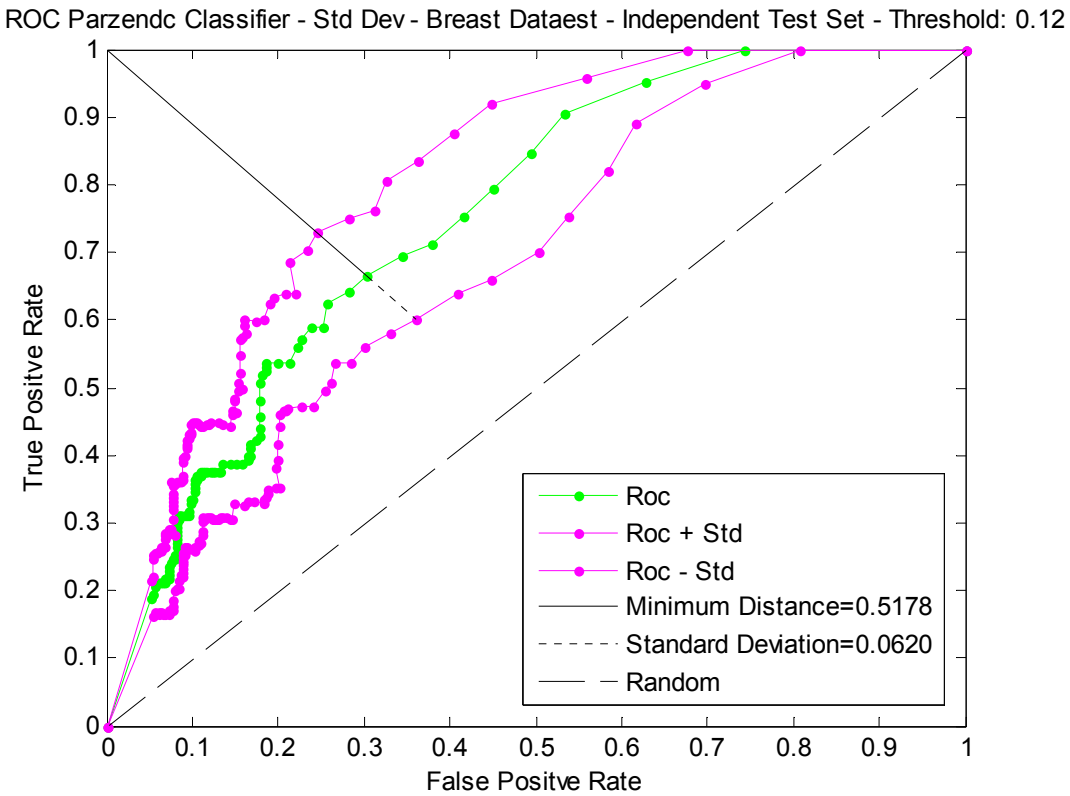
Parzenc


ROC Parzenc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.22
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5806
Standard Deviation=0.3280
Random

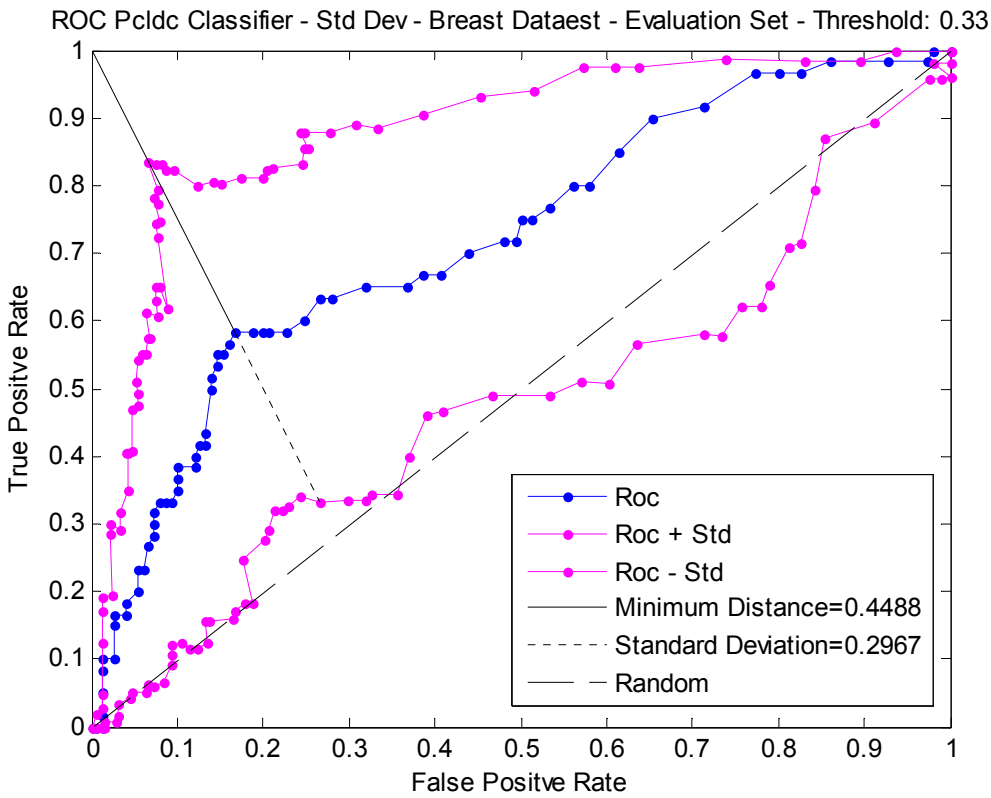

ROC Parzenc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.42
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.7736
Standard Deviation=0.4339
Random

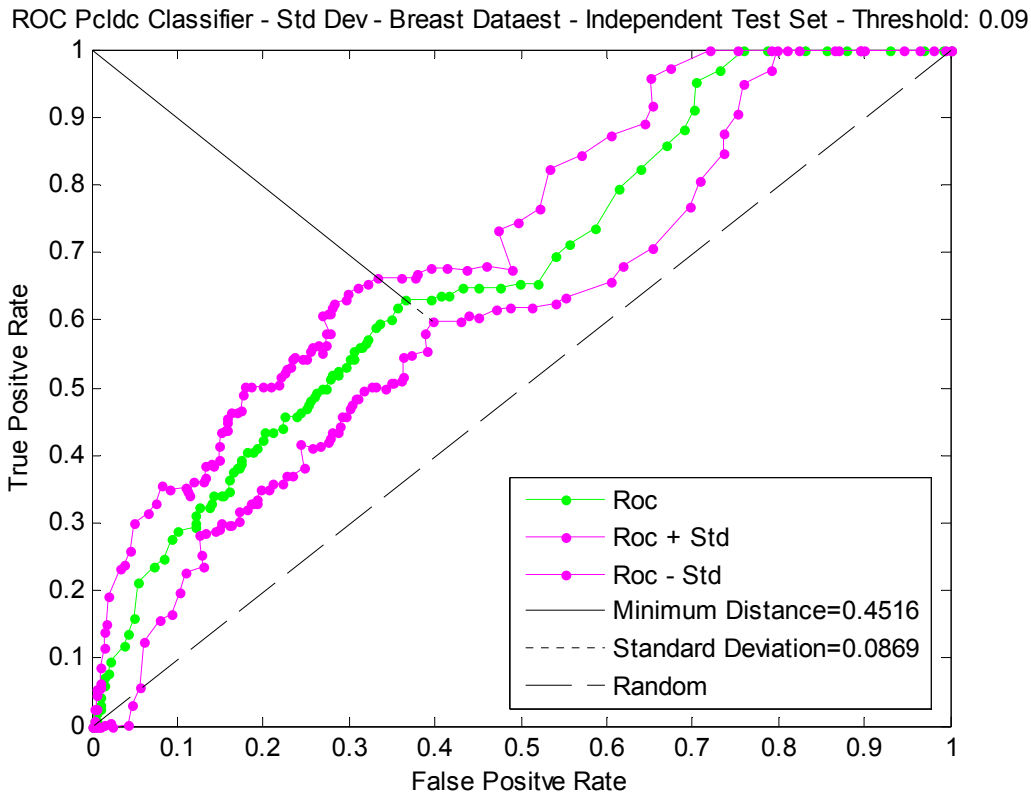
Parzendc


ROC Parzendc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.08
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5420
Standard Deviation=0.2165
Random

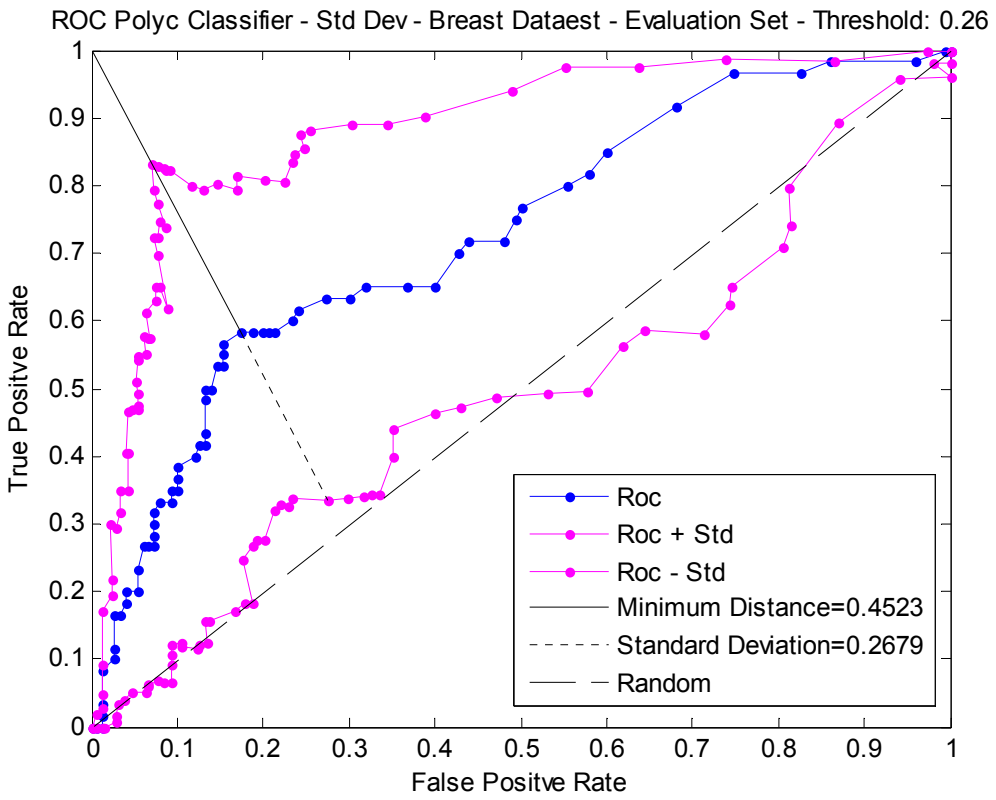

ROC Parzendc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.12
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5178
Standard Deviation=0.0620
Random

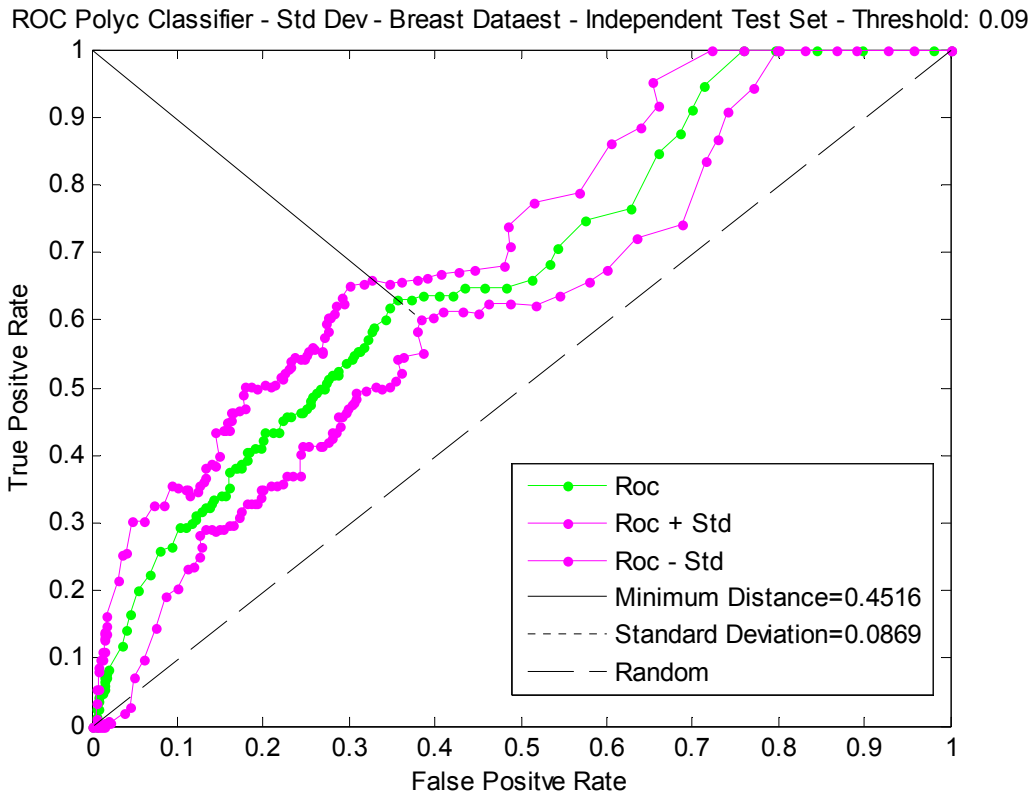
Pcldc


ROC Pcldc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.33
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4488
Standard Deviation=0.2967
Random

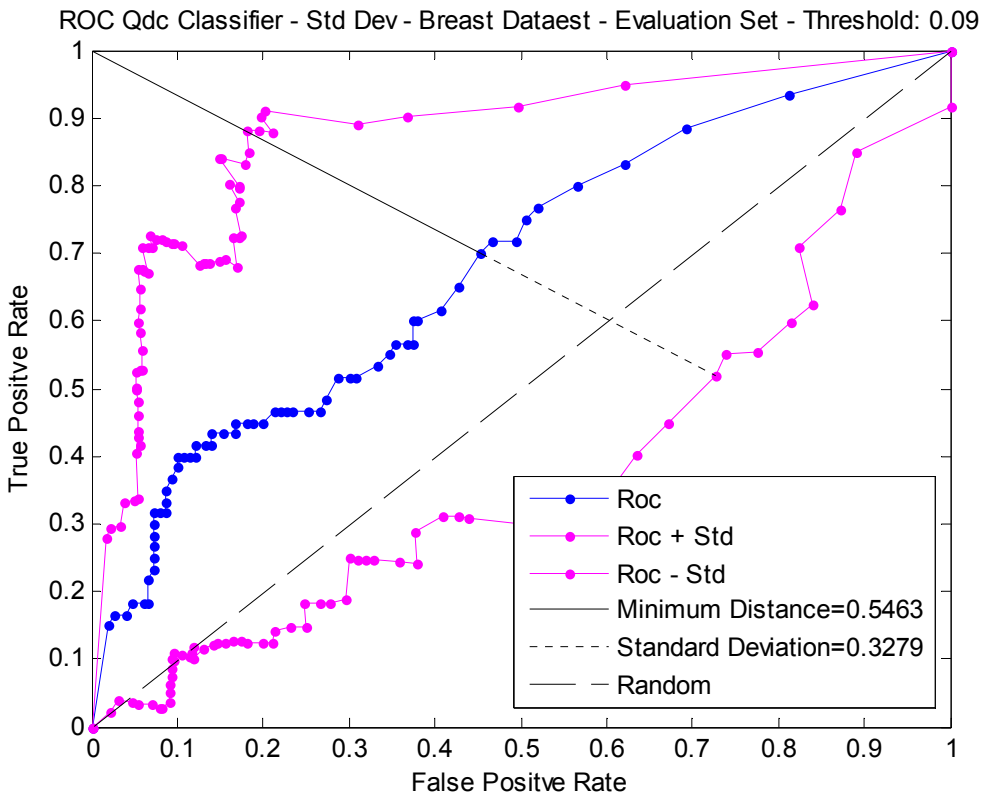

ROC Pcldc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.09
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4516
Standard Deviation=0.0869
Random

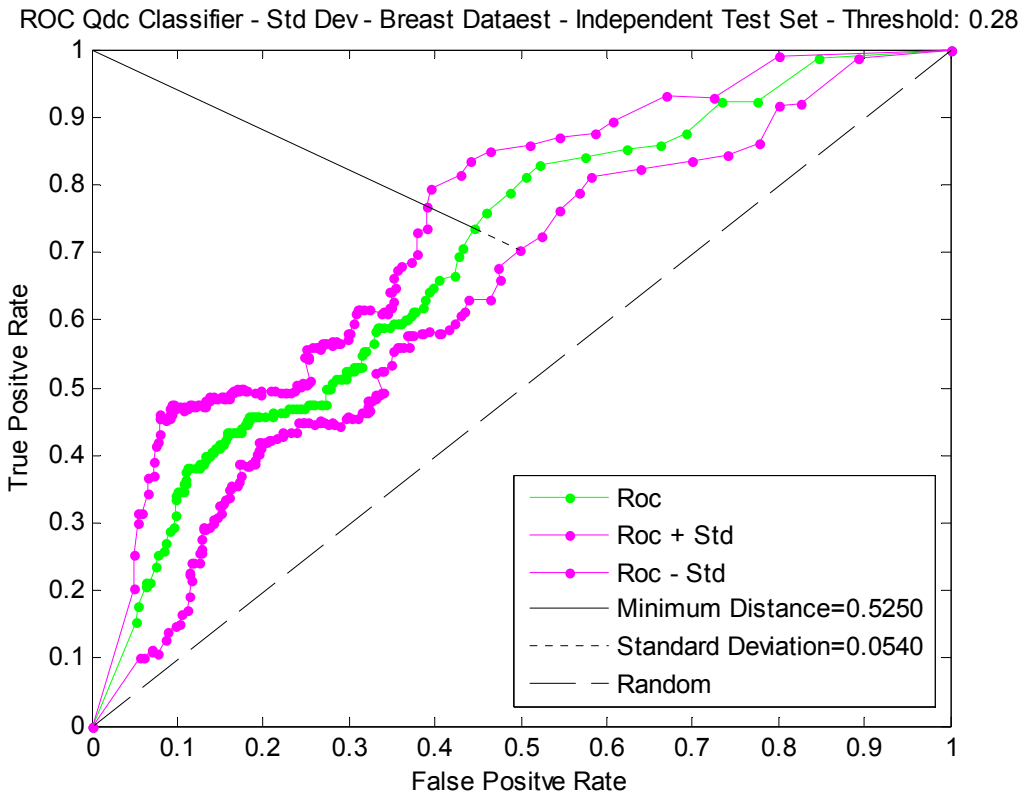
Polyc


ROC Polyc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.26
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4523
Standard Deviation=0.2679
Random

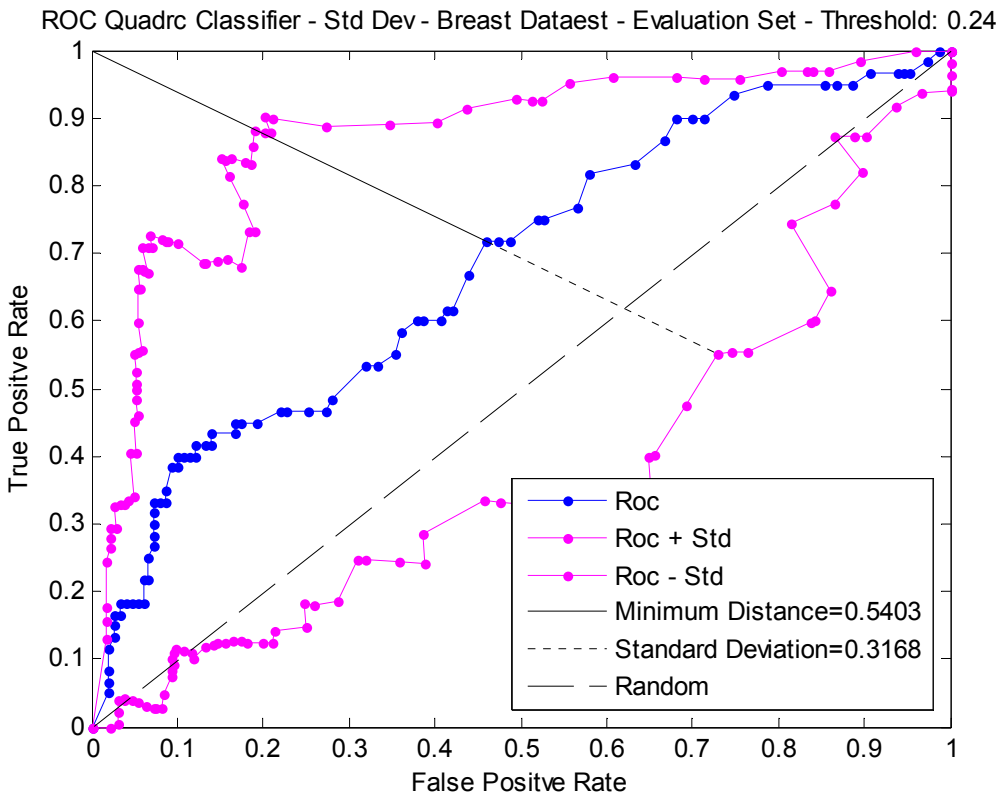

ROC Polyc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.09
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4516
Standard Deviation=0.0869
Random

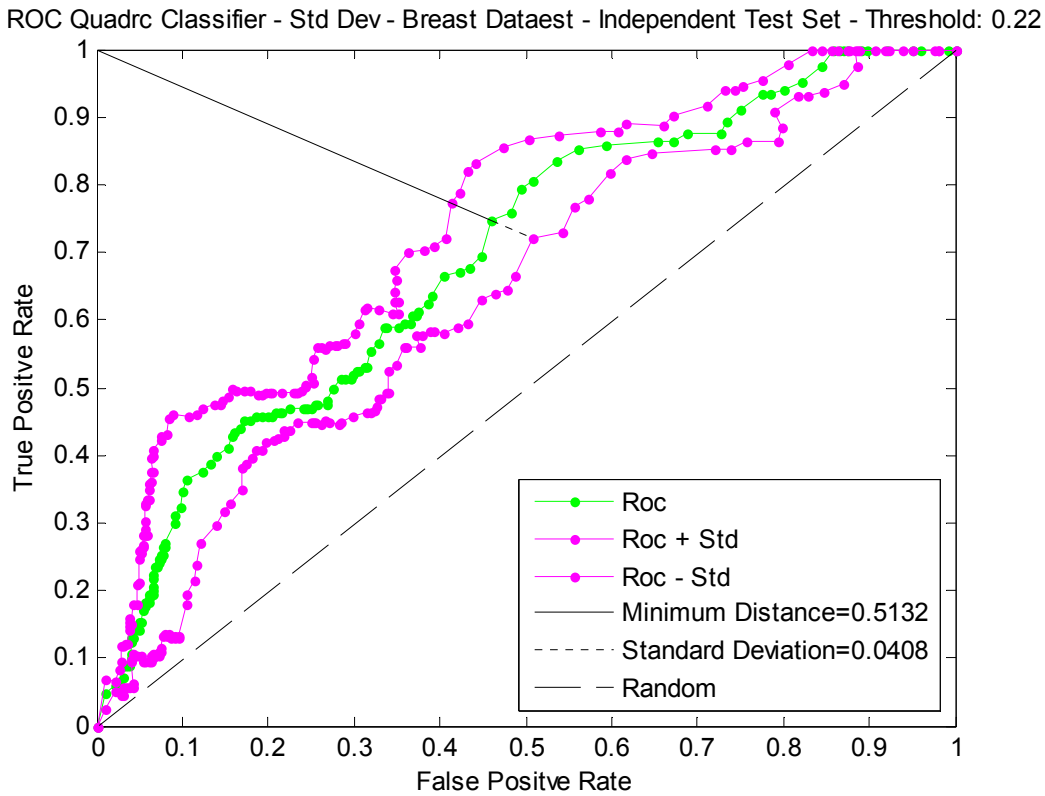
Qdc


ROC Qdc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.09
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5463
Standard Deviation=0.3279
Random

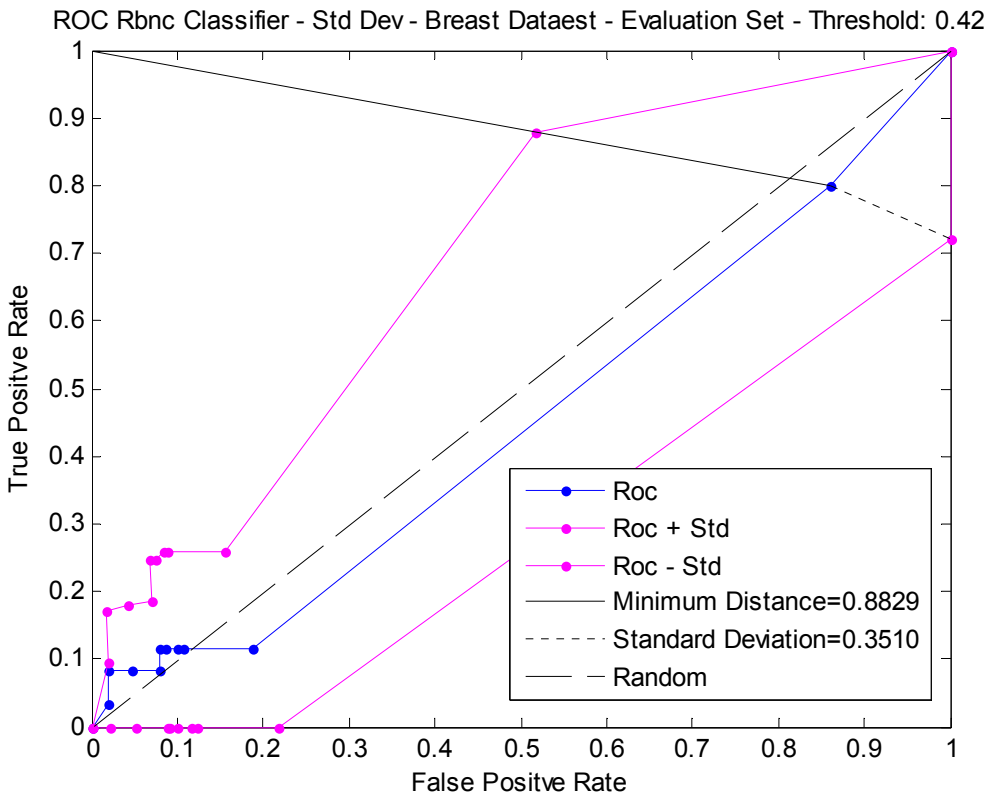

ROC Qdc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.28
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5250
Standard Deviation=0.0540
Random

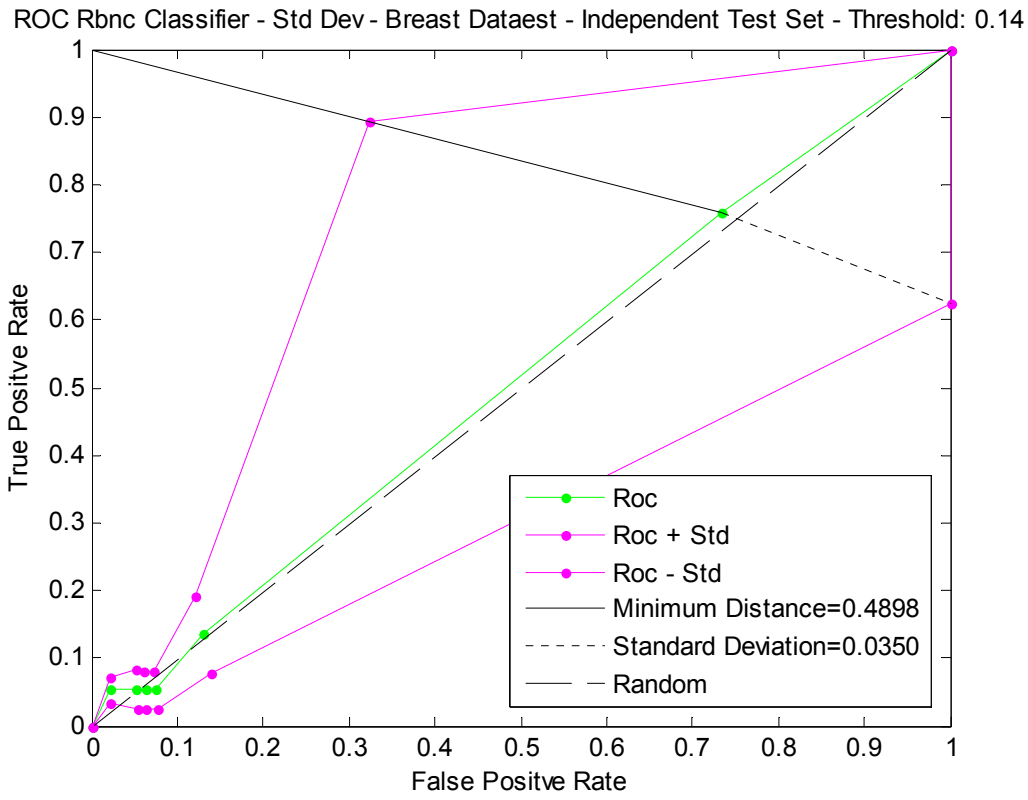
Quadrc


ROC Quadrc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.24
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5403
Standard Deviation=0.3168
Random

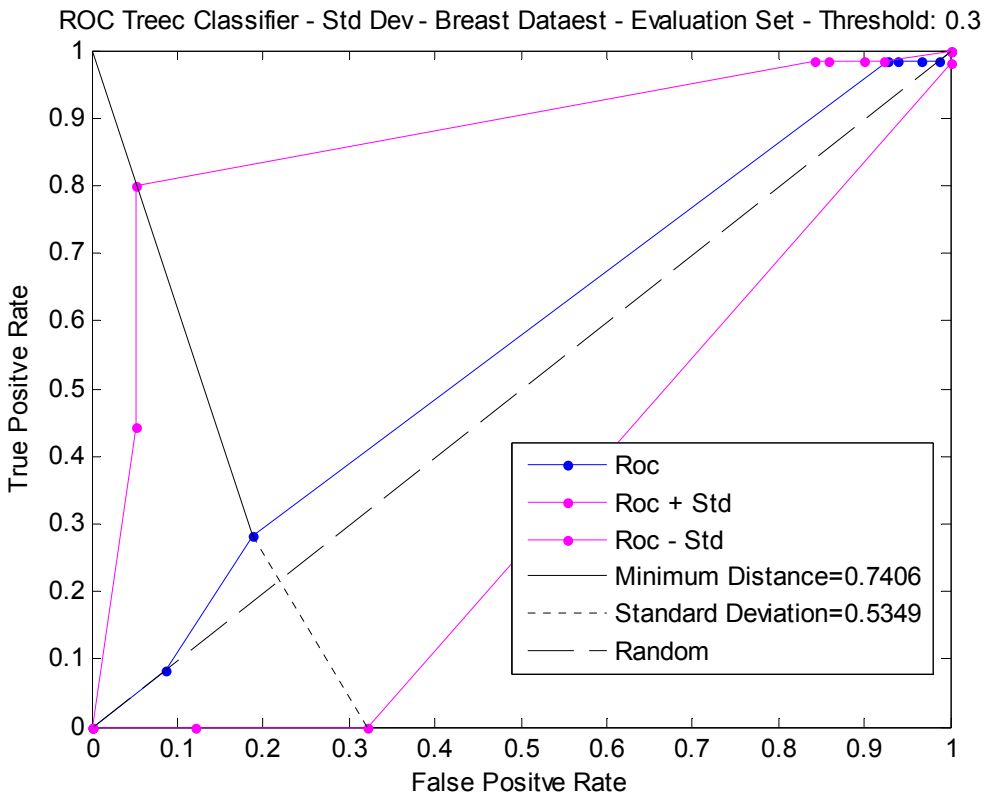

ROC Quadrc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.22
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5132
Standard Deviation=0.0408
Random

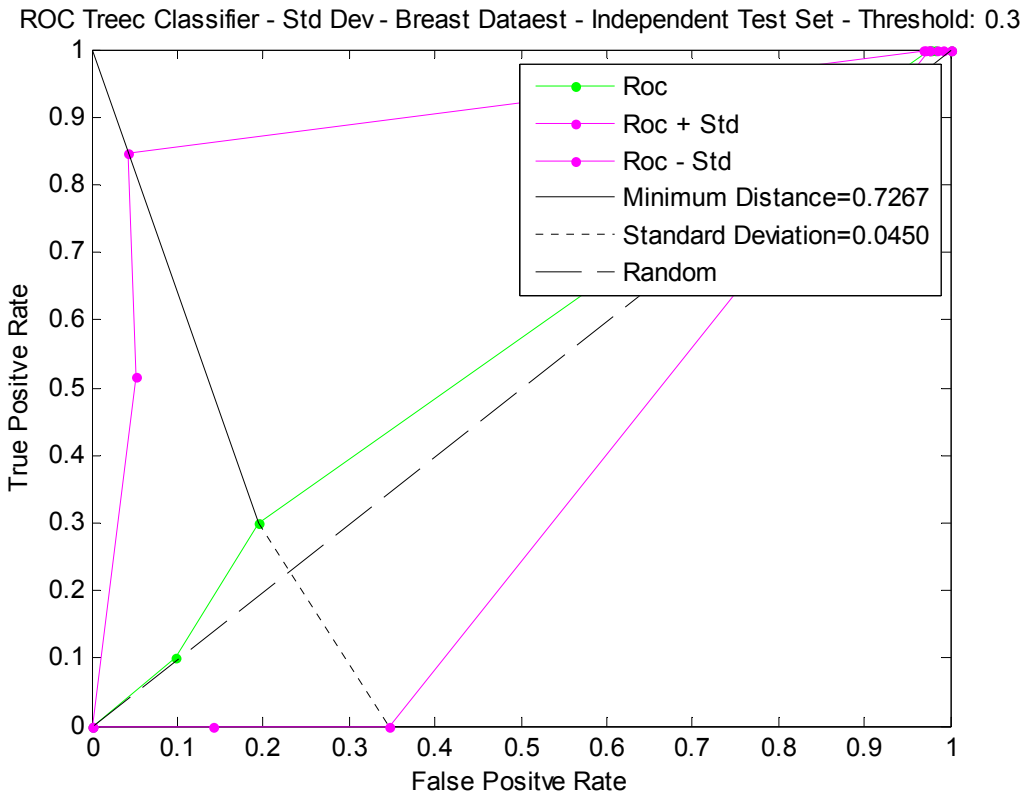
Rbnc


ROC Rbnc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.42
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.8829
Standard Deviation=0.3510
Random

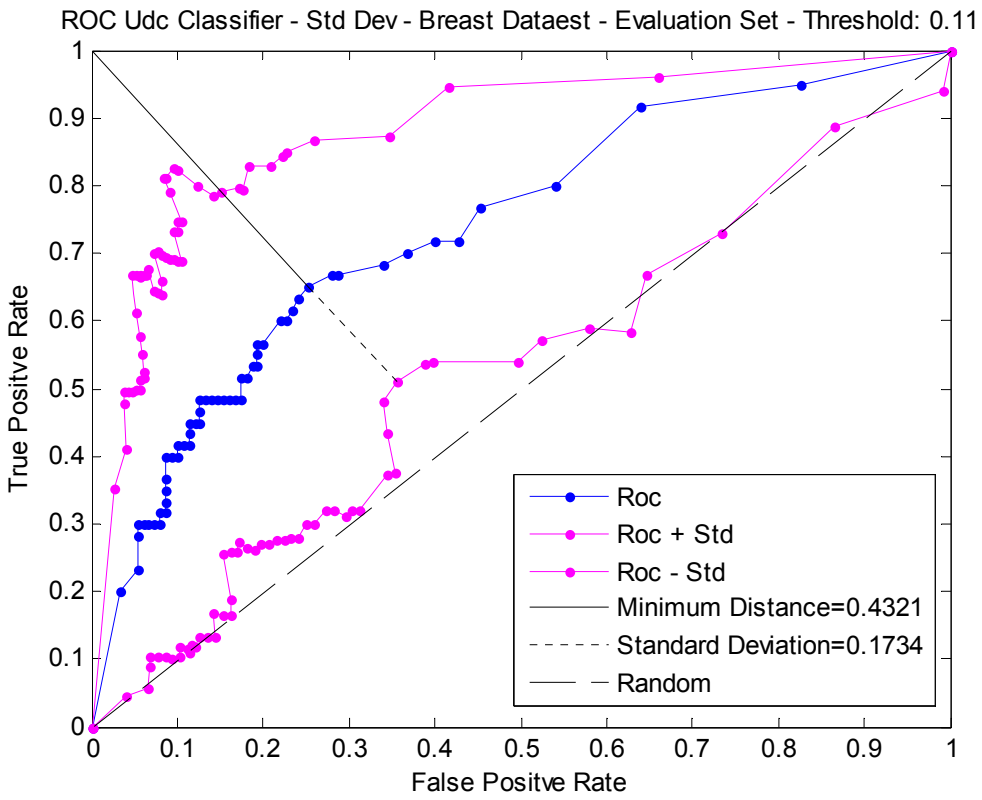

ROC Rbnc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.14
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4898
Standard Deviation=0.0350
Random

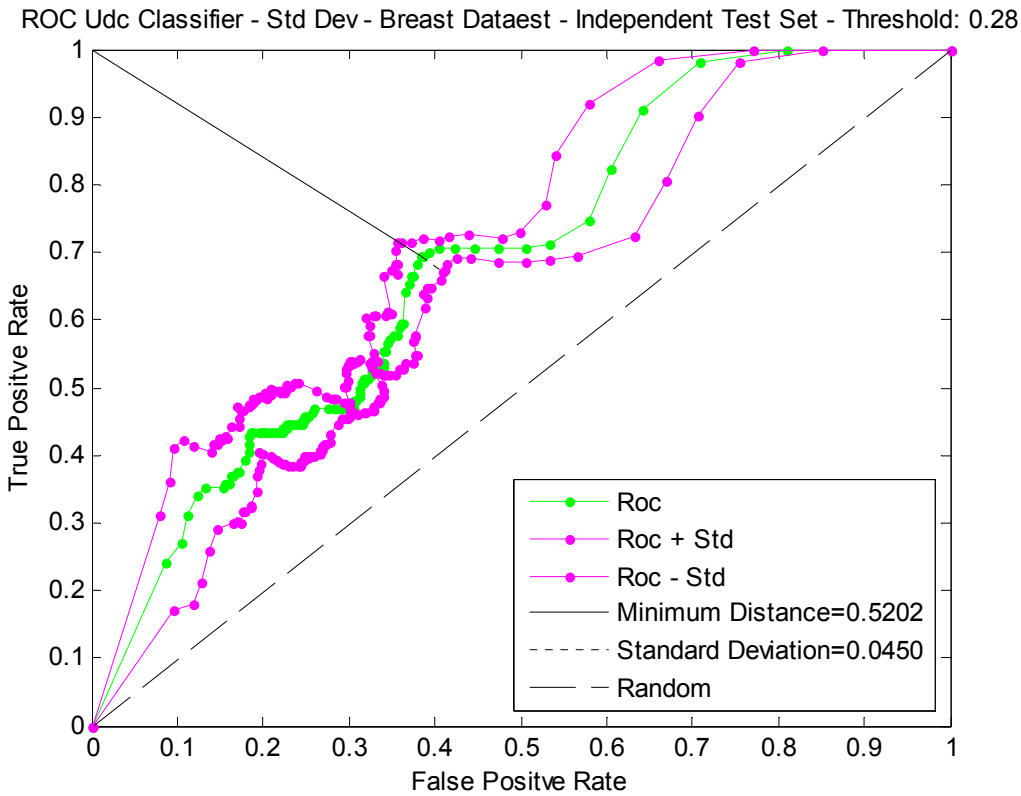
Treec


ROC Treec Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.3
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.7406
Standard Deviation=0.5349
Random


ROC Treec Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.3
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.7267
Standard Deviation=0.0450
Random

Udc


ROC Udc Classifier - Std Dev - Breast Dataest - Evaluation Set - Threshold: 0.11
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.4321
Standard Deviation=0.1734
Random


ROC Udc Classifier - Std Dev - Breast Dataest - Independent Test Set - Threshold: 0.28
True Positve Rate
False Positve Rate
Roc
Roc + Std
Roc - Std
Minimum Distance=0.5202
Standard Deviation=0.0450
Random