ΑΝΑΦΟΡΆ

της εξαμηνιαίας εργασίας με τίτλο:

‹‹ΔΙΕΡΕΥΝΗΣΗ ΔΙΑΦΟΡΩΝ ΑΛΓΟΡΙΘΜΩΝ ΓΙΑ ΤΗ ΔΥΑΔΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΠΕΡΙΟΡΙΣΜΕΝΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ››

για το μάθημα

Προσομοίωση Φυσιολογικών Συστημάτων

Μέλη της ομάδας:

1)Αργυρώ Τσίπη 03119950

2)Κωνσταντίνα Πάνου 03120144

3)Μιχαήλ Σβεντζούρης 03118917

4)Χρήστος Καρβελάς 03120153

Περίληψη

Η συγκεκριμένη εργασία έχει ως στόχο την αξιολόγηση διαφορετικών μεθόδων δυαδικής ταξινόμησης όσον αφορά την αποτελεσματικότητά τους σε ένα μικρό σύνολο δεδομένων. Όλες οι μέθοδοι χρησιμοποιήθηκαν σε συνδυασμό με την προσέγγιση Leave One Out, εκπαίδευση δηλαδή του μοντέλου πάνω στην πλειοψηφία του συνόλου δεδομένων και δοκιμή του πάνω σε ένα δείγμα που αφαιρέθηκε από την εκπαίδευση. Τα αποτελέσματα αξιολογήθηκαν βάσει τεσσάρων αξόνων: accuracy, precision, recall, f1 score και παρουσιάζονται διαγραμματικά. Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν σχετικά μικρό περιλαμβάνοντας μόνο δέκα συμμετέχοντες που παρακολούθησαν δέκα βίντεο ο καθένας, και περιέχει δεδομένα από καταγραφές EEG (ηλεκτροεγκεφαλογράφημα) και δημογραφικές πληροφορίες των συμμετεχόντων. Η δυαδική ταξινόμηση βασίζεται στη διάκριση μεταξύ δύο κατηγοριών (σύγχυση ή όχι σύγχυση) βάσει των χαρακτηριστικών των δεδομένων.

Λέξεις Κλειδιά:

LSTM

Random Forest

Gradient Boosting

eXtreme Gradient Boosting

Leave One Out Cross Validation

Περιγραφη συνολου δεδομενων

Το σύνολο δεδομένων που αξιοποιήθηκε έχει τίτλο: "Confused student EEG brainwave data". Πρόκειται για δεδομένα ηλεκτροεγκεφαλογραφήματος που συλλέχθηκαν από 10 φοιτητές ενώ έβλεπαν βίντεο εικονικών διαλέξεων. Συνολικά υπάρχουν 20 βίντεο, 10 πάνω σε βασική άλγεβρα και γεωμετρία και άλλα 10 πάνω σε σύνθετα θέματα κβαντικής μηχανικής και έρευνας βλαστοκυττάρων. Η μετρήσεις EEG έγιναν με μονοκάναλο ασύρματο ηλεκτροεγκεφαλογράφο MindSet, ο οποίος μετρούσε δραστηριότητα στον μετωπιαίο λοβό. Κάθε φοιτητής παρακολούθησε 10 βίντεο και αξιολόγησε τη σύγχυσή του σε κλίμακα 1-7 μετά το τέλος του καθενός, ενώ στη συνέχεια η ετικέτα απλοποιήθηκε σε δυαδική (0=μη μπερδεμένος, 1=μπερδεμένος).

Οι συγγραφείς του dataset το χαρακτηρίζουν εξαιρετικά δυσπρόσιτο στην εφαρμογή δυαδικής ταξινόμησης και, ειδικά λόγω του περιορισμένου όγκου δεδομένων, αξιοποιείται κυρίως για την εξάσκηση και την αντιμετώπιση των προκλήσεων που επιφέρει το πρόβλημα αυτό. Διάφοροι φοιτητές και ερευνητές πριν από εμάς προσπάθησαν να τελειοποιήσουν τα μοντέλα τους, οπότε υπάρχουν διάφορες απόψεις επί του θέματος.

Δεν μπόρεσε να γίνει ουσιαστική στατιστική ανάλυση του δείγματος, καθώς συμμετείχαν μόλις 10 φοιτητές με παρόμοιων ηλικιών και εθνικοτήτων.

Διατυπωση ερευνητικου ερωτηματος

Στο πλαίσιο της συγκεκριμένης εργασίας αναρωτηθήκαμε σχετικά με την αποτελεσματικότητα διαφορετικών μεθόδων εκπαίδευσης πάνω στο μικρό αυτό σύνολο δεδομένων. Εξετάστηκαν διάφορες τεχνικές προεπεξεργασίας των δεδομένων ώστε να προκύψουν περισσότερα χαρακτηριστικά. Άλλο ένα ερώτημα που εξετάστηκε ήταν κατά πόσο μπορεί η διασταυρούμενη επικύρωση(cross validation) να βελτιώσει την ακρίβεια των προβλέψεων των διαφορετικών μοντέλων που εκπαιδεύτηκαν.

Μέθοδος Ανάλυσης

- Η μέθοδος ανάλυσης που χρησιμοποιήθηκε είναι η **Βαθιά Μάθηση**, με χρήση και σύγκριση των μοντέλων **GB**, **XGB**, **Random Forest** και **BiLSTM**.
- Χρησιμοποιήθηκαν **τεχνικές προεπεξεργασίας δεδομένων** όπως η **κανονικοποίηση (Normalization)** και η **μοντελοποίηση κατηγοριών (One-Hot Encoding)** για την προετοιμασία των δεδομένων.
- **Leave one out cross validation** για να εκμεταλλευτούμε στο μέγιστο το περιορισμένο σύνολο δεδομένων

Τεχνολογίες και Εργαλεία

- **Python**: Χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python για την ανάπτυξη των μοντέλων.
- **Pandas and Numpy**: Χρησιμοποιήθηκαν για την επεξεργασία και τον χειρισμό των δεδομένων.
- **Matplotlib**: Χρησιμοποιήθηκε για τη δημιουργία γραφημάτων και visualisation των αποτελεσμάτων, όπως γραμμικά διαγράμματα για την απόδοση των μοντέλων σε κάθε επανάληψη και στατιστικά διαγράμματα (ιστογράμματα και box plots).
- **Scikit-learn:**
    - Χρησιμοποιήθηκε ο **GradientBoostingClassifier** της βιβλιοθήκης με σκοπό τη δημιουργία ενός αντίστοιχου μοντέλου για classification. Το συγκεκριμένο μοντέλο κατασκευάζει μια σειρά από regression trees, όπου κάθε ένα εκπαιδεύεται με σκοπό να διορθώσει τα λάθη των προηγούμενων iterations, δίνοντας βάση στα misclassifications. Αυτό γίνεται με το optimization μιας loss function με gradient descent, βοηθώντας το μοντέλο να ελαχιστοποιήσει τα λάθη κατά τη διάρκεια του training.

      Αξίζει να αναφέρουμε πως το συγκεκριμένο μοντέλο, δεν έχει την ικανότητα να διαχειριστεί "sequential" δεδομένα, και έτσι δεν μπορούμε να αξιοποιήσουμε τις χρονικές εξαρτήσεις των δεδομένων μας. Για αυτό το λόγο, θεωρούμε πως κάθε χρονική στιγμή είναι ανεξάρτητη, και έτσι κάνουμε "flatten" τον πίνακα με τα δεδομένα μας, από 3D σε 2D. Έτσι μπορούμε να αξιοποιήσουμε την ικανότητα του μοντέλου να αποτυπώνει σύνθετες σχέσεις με το feature set.

    - Αξιοποιήθηκε για την υλοποίηση και εκπαίδευση του μοντέλου **Random Forest**. Η μέθοδος αυτή βασίζεται στη δημιουργία ενός συνόλου δέντρων αποφάσεων

(decision trees), όπου κάθε δέντρο εκπαιδεύεται σε διαφορετικό τυχαίο υποσύνολο δεδομένων και χαρακτηριστικών. Το τελικό αποτέλεσμα προκύπτει μέσω πλειοψηφίας ψήφων από τα δέντρα, εξασφαλίζοντας μεγαλύτερη ακρίβεια και ανθεκτικότητα στον υπερπροσαρμογή (overfitting).

- Στην περίπτωση των μοντέλων **XGB και BiLSTM** αξιοποιήθηκαν ο StandardScaler και ο OneHotEncoder για τη προεπεξεργασία των δεδομένων, η μέθοδος LeaveOneOut για cross validation καθώς και διάφορες συναρτήσεις για την εκτίμηση ακρίβειας.

- **XGBoost:** Χρησιμοποιήθηκε για πιο γρήγορη και απλοποιημένη εκπαίδευση του μοντέλου **XGB** αφού είναι μία βιβλιοθήκη απόλυτα συμβατή με τη scikit-learn, ενώ προσφέρει δυνατότητες όπως επιτάχυνση με GPU, περισσότερες μεταβλητές παραμέτρους για τη τελειοποίηση του μοντέλου και διαγνωστικά εργαλεία για την καλύτερη αντιμετώπιση προβλημάτων κατά την εκπαίδευση.
- **TensorFlow/Keras**: Χρησιμοποιήθηκε για την ανάπτυξη και εκπαίδευση του **BiLSTM** μοντέλου. Περιλαμβάνει τον Keras API για την εύκολη δημιουργία του νευρωνικού δικτύου.

Μεθοδολογία Εκπαίδευσης

Ανάλυση Δεδομένων

- Τα δεδομένα EEG και δημογραφικά χαρακτηριστικά των συμμετεχόντων συνδυάστηκαν για να δημιουργηθεί ένα σύνολο χαρακτηριστικών και ενιαίο dataset που περιλαμβάνει όλες τις πληροφορίες.
- Εφαρμόστηκε **κανονικοποίηση** των χαρακτηριστικών του EEG για να εξασφαλιστεί η ομοιογένεια της κλίμακας των δεδομένων, αποφεύγοντας την επιρροή που θα είχε η κλίμακα διαφορετικών χαρακτηριστικών.
- Εφαρμόστηκε **One-Hot Encoding** στις κατηγορίες φύλου και εθνικότητας για να μετατραπούν σε μορφή που μπορεί να κατανοηθεί από το μοντέλο.
- Τα δεδομένα **ομαδοποιήθηκαν** κατά SubjectID και VideoID, ενώ οι ετικέτες που χρησιμοποιήθηκαν ήταν οι ετικέτες σύγχυσης των συμμετεχόντων, οι οποίες είναι δυαδικές. Με αυτή την ομαδοποίηση δημιουργούνται 100 μοναδικά data points (10 video για καθένα από τα 10 subjects), ο μέγιστος αριθμός ο οποίος μπορούσε να προκύψει για αυτό το σύνολο δεδομένων.

Εκπαίδευση του Μοντέλου

- Χρησιμοποιήθηκε η τεχνική **Leave-One-Out Cross-Validation (LOOCV)** για να εκτιμηθεί η επίδοση του μοντέλου. Σε κάθε βήμα της LOOCV, το μοντέλο εκπαιδεύτηκε με όλα τα δεδομένα εκτός από το τρέχον δείγμα και το αποτέλεσμα του μοντέλου αξιολογήθηκε με βάση την πρόβλεψη αυτού του δείγματος.
- Το **BiLSTM μοντέλο** εκπαιδεύτηκε για 10 epochs, με την **ενημέρωση των βαρών** τους μέσω του βελτιστοποιητή **Adam** και του σφάλματος **binary crossentropy** για τη δυαδική ταξινόμηση.
- Τα **GB, XGB και RF μοντέλα** εκπαιδεύτηκαν με 100 n_estimators, δημιουργήθηκαν και εκπαιδεύτηκαν δηλαδή 100 δέντρα αποφάσεων διαδοχικά, το καθένα βελτιωμένη έκδοση του προηγούμενου.

Κριτήρια Αξιολόγησης της Επίδοσης

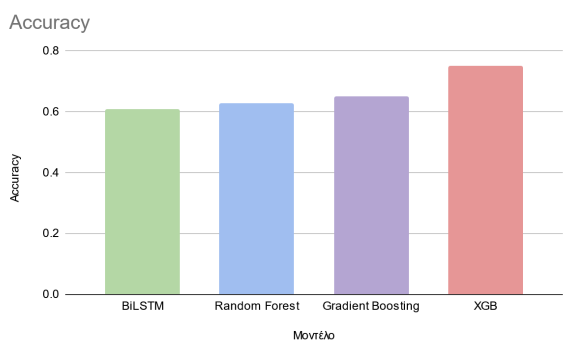
Η αξιολόγηση των μοντέλων έγινε με τη χρήση των εξής μετρικών απόδοσης που είναι κατάλληλες για προβλήματα δυαδικής ταξινόμησης:

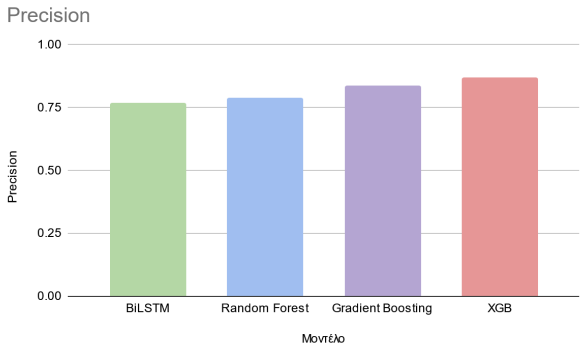
1. **Accuracy**: Η ακριβής αναλογία σωστών προβλέψεων προς το σύνολο των προβλέψεων. Αν και σε μικρά σύνολα δεδομένων η ακρίβεια μπορεί να παραπλανεί, θεωρείται βασικό μέτρο για την αξιολόγηση.
2. **Precision**: Η αναλογία των σωστών θετικών προβλέψεων προς το σύνολο των θετικών προβλέψεων. Χρησιμοποιείται για να μετρηθεί πόσο ακριβές είναι το μοντέλο όταν προβλέπει θετικά.
3. **Recall**: Η αναλογία των σωστών θετικών προβλέψεων προς το σύνολο των πραγματικών θετικών δειγμάτων. Μας λέει πόσα από τα θετικά δείγματα εντοπίζει το μοντέλο.
4. **F1-Score**: Ο αρμονικός μέσος όρος του Precision και του Recall. Χρησιμοποιείται για να δώσει μια πιο ισχυρή εικόνα της απόδοσης του μοντέλου, ειδικά σε περιπτώσεις ανισορροπίας στις κλάσεις.

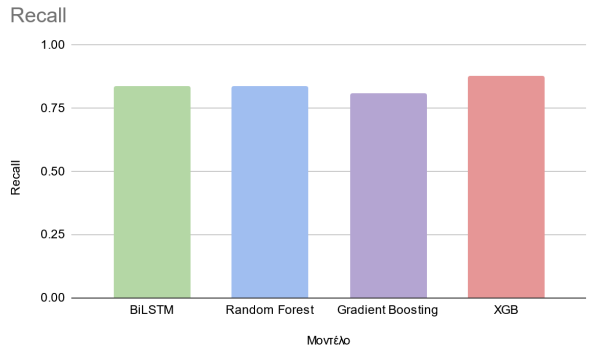
Το συγκεκριμένο σύνολο δεδομένων, λόγω της μικρής του κλίμακας, ήταν εξαιρετικά δύσκολο για binary classification. Οι συγγραφείς ορίζουν ακρίβεια άνω του ~65% ικανοποιητική για μοντέλα δυαδικής ταξινόμησης πάνω στο συγκεκριμένο σύνολο δεδομένων.

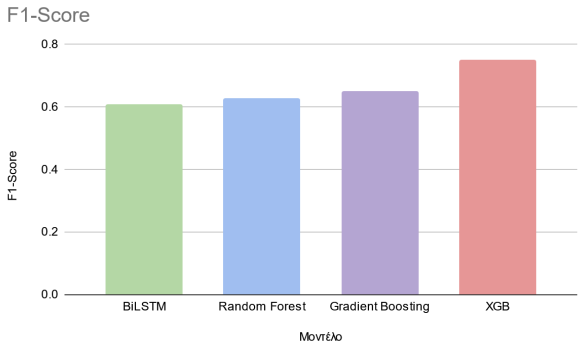
Τα αποτελέσματα της αξιολόγησης των μοντέλων παρουσιάζονται στον παρακάτω πίνακα και γραφήματα:

Μοντέλο	Accuracy	Precision	Recall	F1-Score
BiLSTM	0.61	0.77	0.84	0.61
Random Forest	0.63	0.79	0.84	0.63
Gradient Boosting	0.65	0.84	0.81	0.65
XGB	0.75	0.87	0.88	0.75



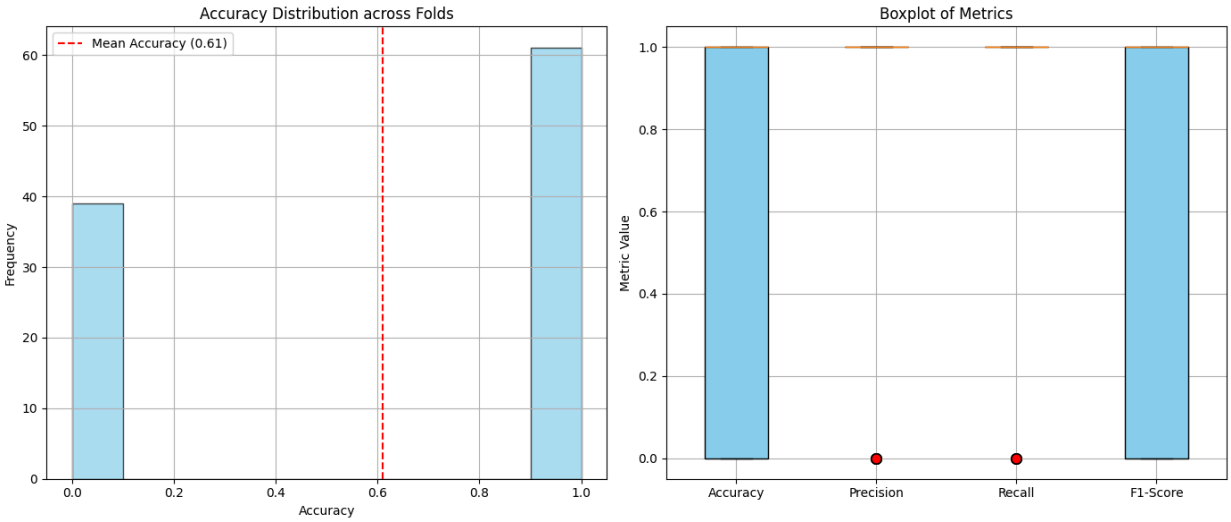




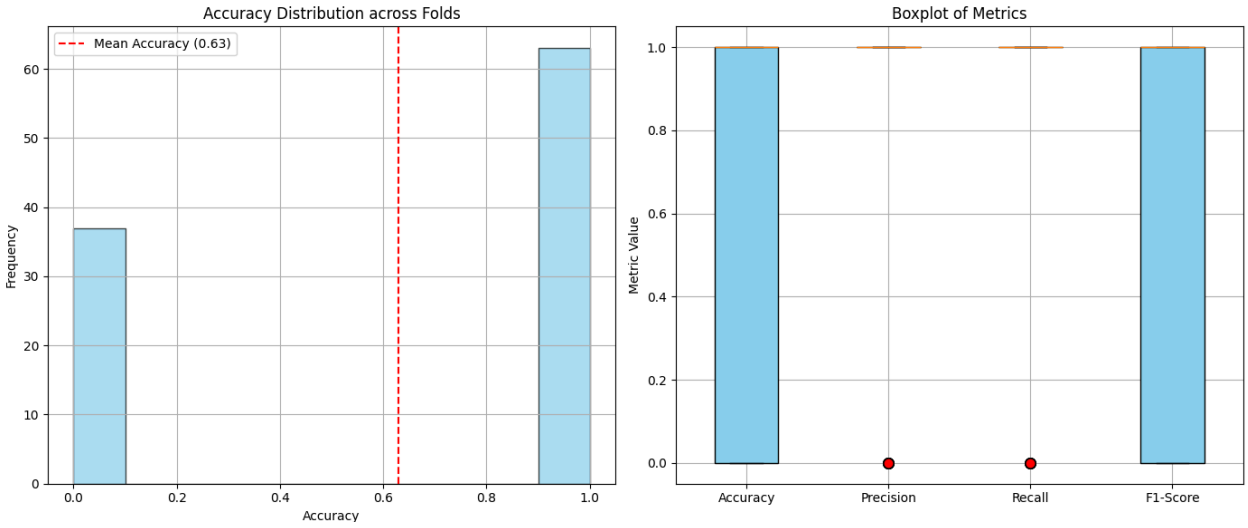


Histogram and BoxPlot για το κάθε μοντέλο:

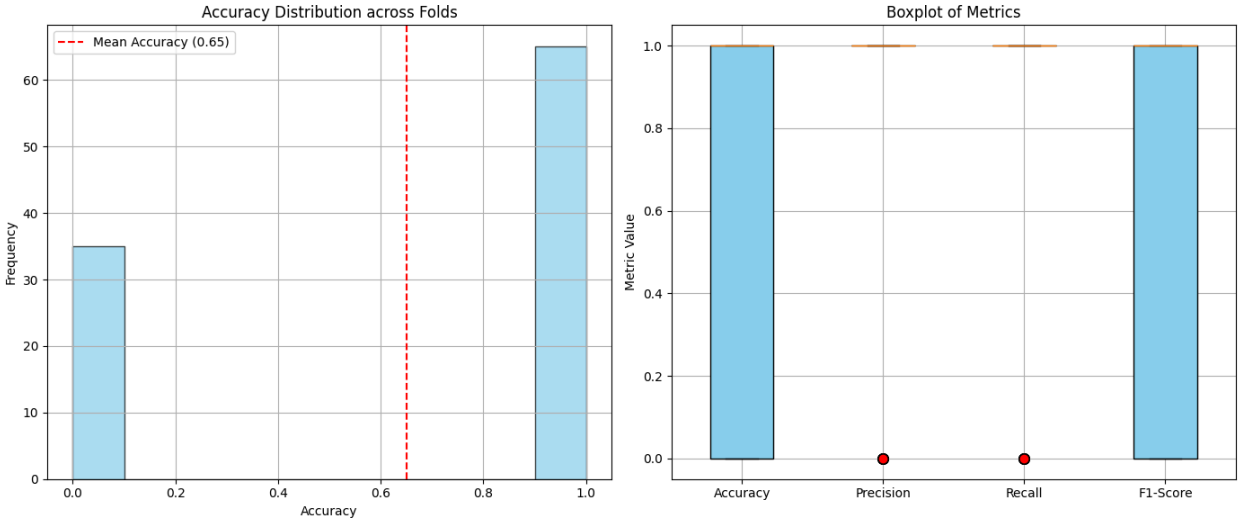
BiLSTM



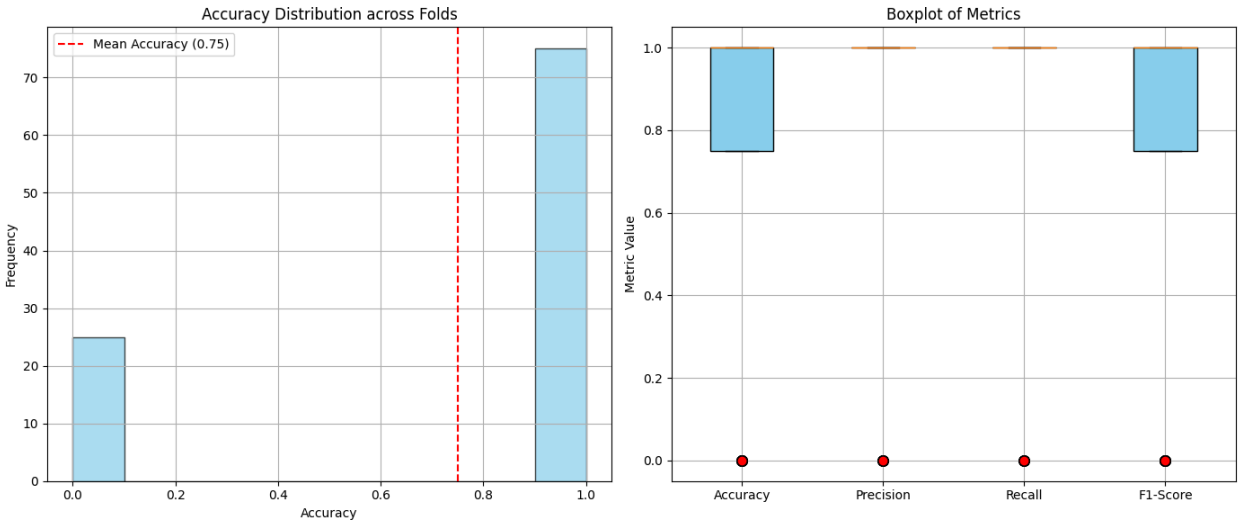
Random Forest



Gradient Boosting

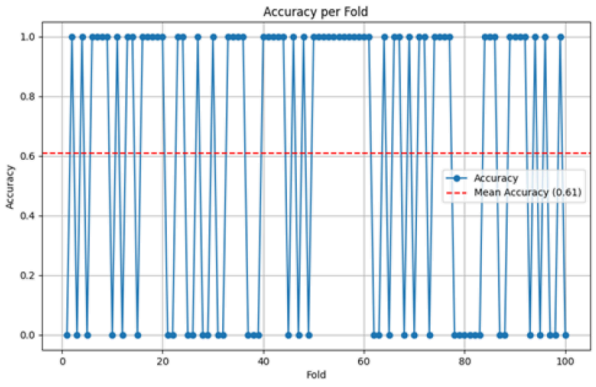


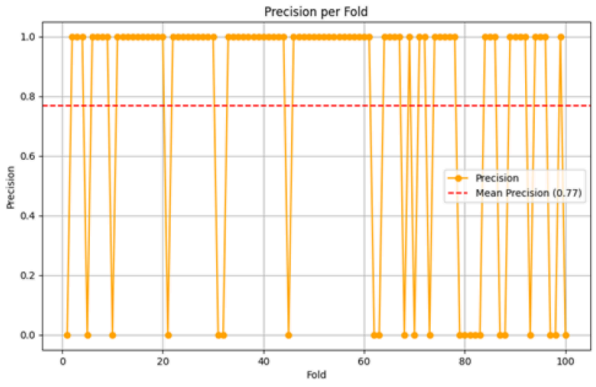
XGB

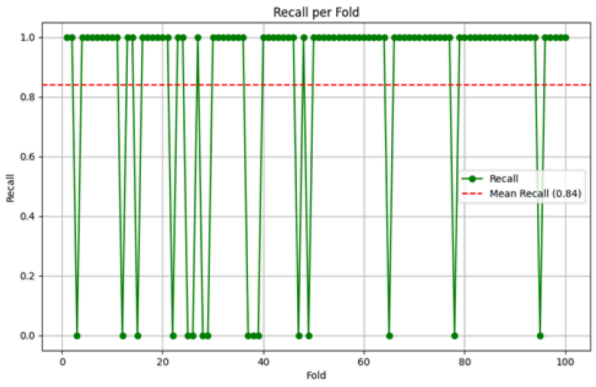


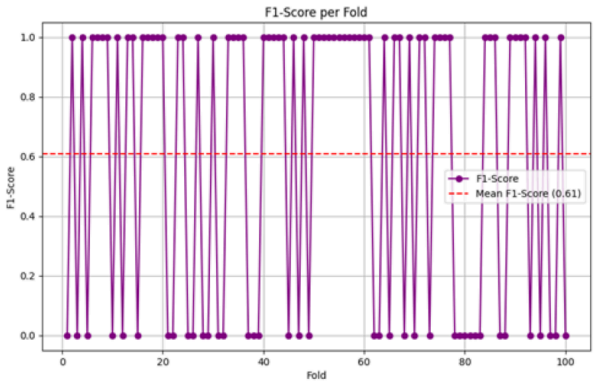
Στη συνέχεια θα παρουσιάσουμε τα Metrics per Fold για κάθε μοντέλο:

BiLSTM

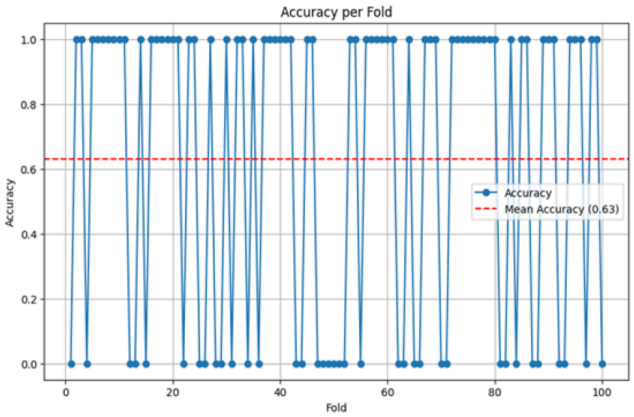


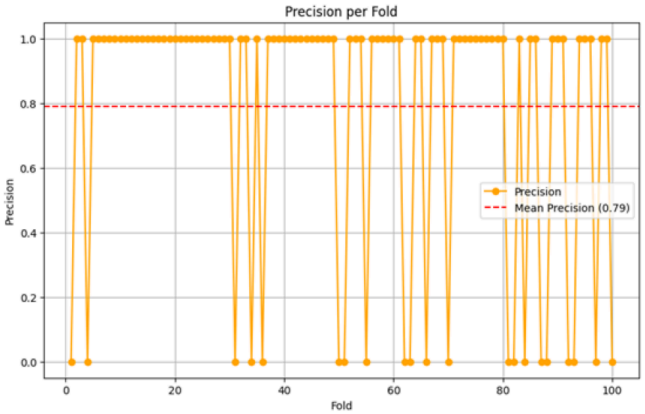


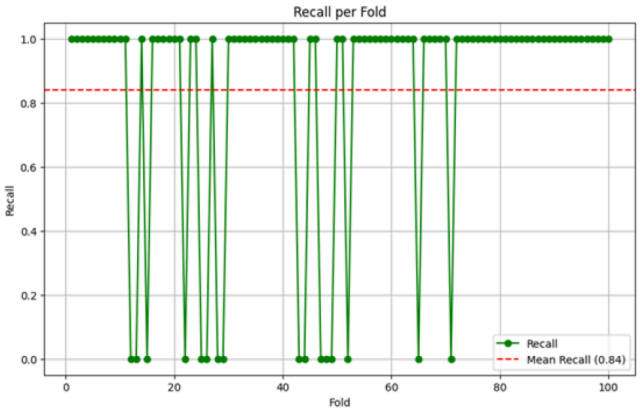


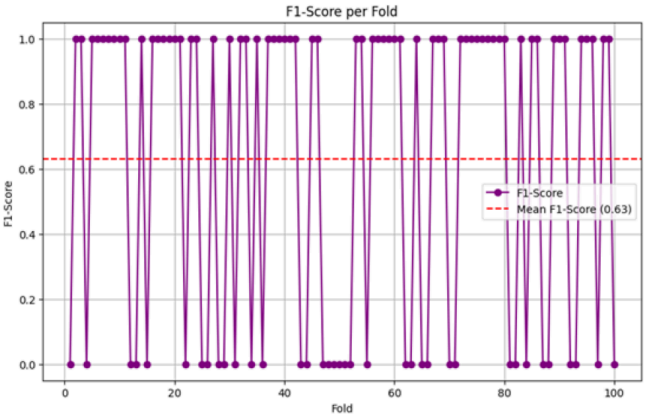


Random Forest

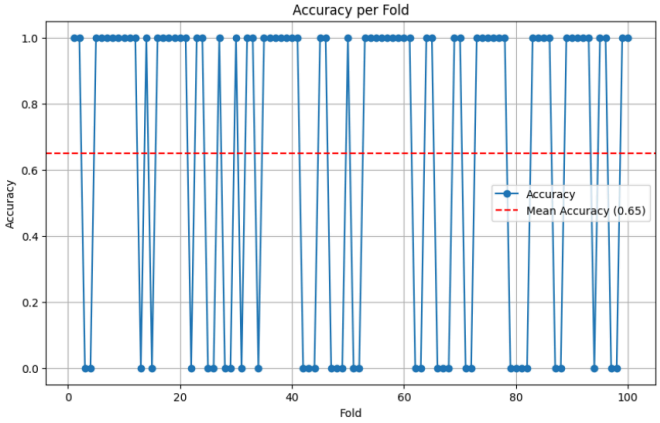


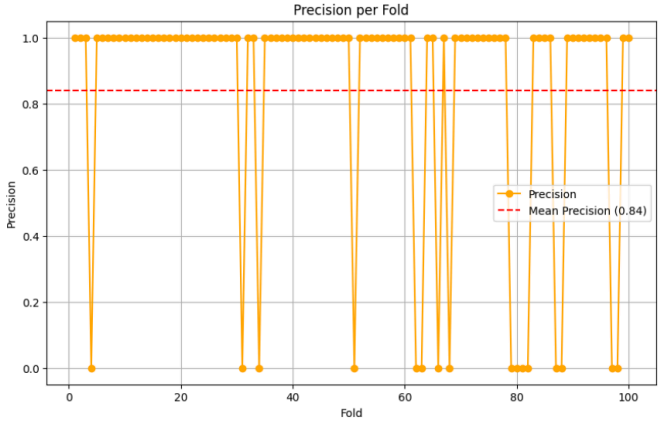


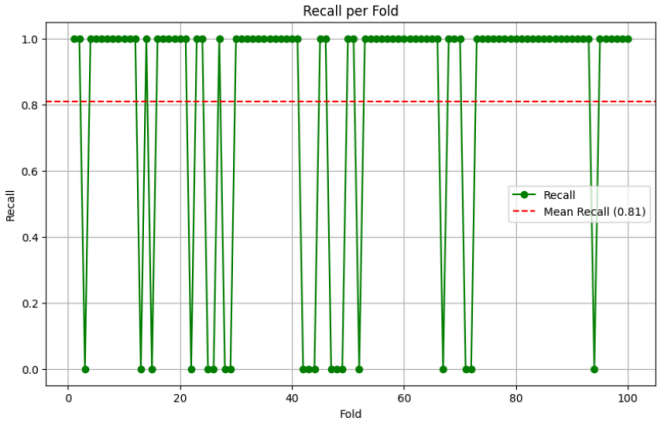


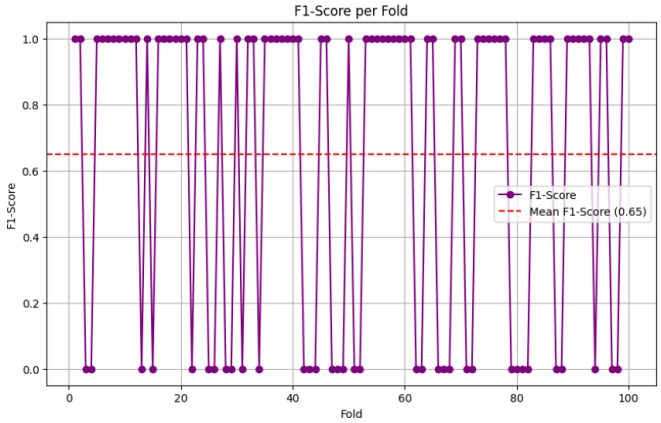


Gradient Boosting

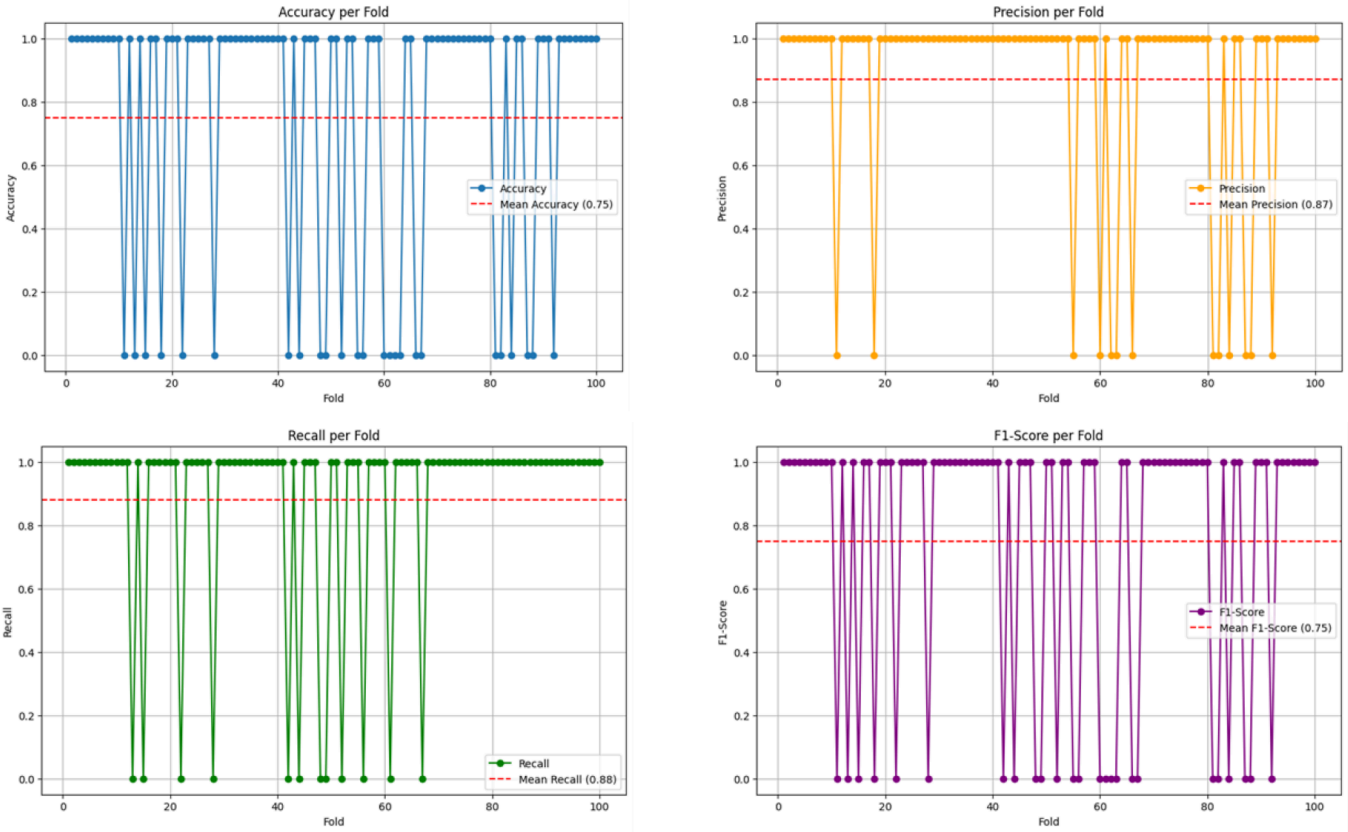








XGB



ΣΥΖΗΤΗΣΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

Κατόπιν σύγκρισης των αναλυτικών αποτελεσμάτων κάθε μεθόδου εκπαίδευσης προκύπτει ότι η πιο αποτελεσματική είναι το XGB, με μέση ακρίβεια 75%. Αν και η συγκεκριμένη ακρίβεια δεν θα ήταν ικανοποιητική σε κάποιο μεγάλο σύνολο δεδομένων, είναι πολύ μεγαλύτερη της προτεινόμενης ~65% από τους δημιουργούς του συνόλου δεδομένων. Στο kaggle οι επικρατέστερες προσεγγίσεις χρησιμοποιούσαν μεθόδους όπως k-fold για cross validation και επιτεύχθηκαν ακρίβειες άνω του 99%. Τελικά όμως πρόκειται για overfitting, καθώς τέτοιες μέθοδοι δεν συνίστανται για μικρά σύνολα δεδομένων. Η μικρότερη ακρίβεια της τάξης του 75% παρέχει υψηλές προσδοκίες στην περίπτωση εφαρμογής του μοντέλου σε νέα δεδομένα καθώς η συγκεκριμένη προσέγγιση αποφεύγει το πρόβλημα του overfitting.