**CT Medical Images**

Πραγματοποιήθηκε εξερευνητική ανάλυση δεδομένων (EDA) για το dataset CT Medical Images, εστιάζοντας σε 100 εικόνες που περιλάμβαναν πλήρη μεταδεδομένα και labels, τα οποία προέρχονταν από τα αρχεία dicom\_dir, tiff\_images και overview, το οποίο, ωστόσο, δεν αξιοποιήθηκε συνολικά, καθώς το αρχείο *full\_archive.npz* (μορφή NumPy arrays) δεν περιλάμβανε labels (π.χ 0 και 1), τα οποία είναι απαραίτητα για την εκπαίδευση ή την αξιολόγηση μοντέλων μηχανικής μάθησης. Η μέση ηλικία των ασθενών ήταν 67,48 έτη, με εύρος από 39 έως 83. Οι τιμές των pixels παρουσίαζαν σημαντική διακύμανση (από -1024 έως 3071), υποδεικνύοντας την ανάγκη για κανονικοποίηση πριν την εφαρμογή σε μοντέλα μηχανικής μάθησης. Εντοπίστηκαν επίσης ακραίες τιμές (outliers), οι οποίες υποδεικνύουν ετερογένεια στο σύνολο των δεδομένων. Καμία από τις εικόνες δεν εμφάνισε ελλείπουσες τιμές στα pixels. Στην προεπεξεργασία, εντοπίστηκαν δύο ελλείπουσες τιμές στη στήλη ContrastTag, οι οποίες αντικαταστάθηκαν με την επικρατούσα τιμή "NONE", χωρίς να επηρεάσουν την κατανομή των δεδομένων. Η τελική, καθαρή εκδοχή του συνόλου αποθηκεύτηκε στο αρχείο eda\_cleaned\_updated.csv.

**Medical MNIST**

Στο Medical MNIST, παρατηρήθηκαν βασικά στατιστικά των pixel, με μέση τιμή 94,18 και διάμεσο 123. Η τυπική απόκλιση ήταν 78,6, που δείχνει σημαντική διακύμανση στις τιμές. Οι τιμές των pixel κυμαίνονται από 0 (σκοτάδι) έως 255 (πλήρης φωτεινότητα).

**Κανονικοποίηση/Τυποποίηση χαρακτηριστικών:**

Εφαρμόσαμε τη μέθοδο StandardScaler για τυποποίηση των χαρακτηριστικών, με αποτέλεσμα κάθε pixel να έχει μέση τιμή 0 και τυπική απόκλιση 1. Εντοπίστηκαν 7.750 ακραιες τιμές (outliers), οι οποίες θεωρήθηκαν αναμενόμενες λόγω της φύσης των εικόνων και διατηρήθηκαν. Η τυποποίηση έδειξε ικανοποιητικά αποτελέσματα, χωρίς ελλείπουσες τιμές, διευκολύνοντας την περαιτέρω ανάλυση.

**Επιλογή και μετασχηματισμός χαρακτηριστικών:**

Εφαρμόσαμε τον μετασχηματισμό διακύμανσης (VarianceThreshold) για να αφαιρέσουμε χαρακτηριστικά με μικρή διακύμανση (κάτω από το 0.01), κρατώντας μόνο τα σημαντικά χαρακτηριστικά που επηρεάζουν την ανάλυση των εικόνων. Μετά την εφαρμογή του, παρατηρήσαμε ότι παρέμειναν 4096 χαρακτηριστικά, υποδεικνύοντας ότι δεν υπήρχαν χαρακτηριστικά με χαμηλή διακύμανση που να μπορούν να αφαιρεθούν. Επομένως, μπορούμε να παραλείψουμε αυτό το βήμα του feature engineering, καθώς τα δεδομένα μας έχουν ήδη αρκετή ποικιλία, και να προχωρήσουμε σε μείωση διαστάσεων για βελτίωση της απόδοσης του μοντέλου.

**Διαχωρισμός σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου**

Αρχικά, περιορίσαμε τον αριθμό των εικόνων σε κάθε κατηγορία (BreastMRI, ChestCT, CXR, Hand) σε 1000, δημιουργώντας ένα σύνολο 4000 εικόνων για ταχύτερη επεξεργασία και εκπαίδευση χωρίς να χάνεται η ποικιλία των κατηγοριών. Στη συνέχεια, διαιρέσαμε το σύνολο δεδομένων σε τρία υποσύνολα. Συγκεκριμένα, το training set (70%): 2800 εικόνες, το validation set (20%): 804 εικόνες και το test set (10%): 396 εικόνες. Η κατανομή των εικόνων στις κατηγορίες ήταν ισοκατανομημένη, με 700 εικόνες για κάθε κατηγορία στο training set, 201 στο validation set και 99 στο test set. Χρησιμοποιήσαμε τη μέθοδο stratify για να διασφαλίσουμε ότι η κατανομή των κατηγοριών ήταν η ίδια σε όλα τα σύνολα, εξασφαλίζοντας αντιπροσωπευτικότητα των δεδομένων. Το μοντέλο Λογιστικής Παλινδρόμησης εκπαιδεύτηκε και παρουσίασε εξαιρετικά αποτελέσματα, με 100% accuracy στο training set, 99,50% accuracy στο validation set και 99,75% accuracy στο test set.Το μοντέλο παρουσίασε τέλεια ακρίβεια και recall για όλες τις κατηγορίες και στα τρία σετ. Μόνο στις κατηγορίες CXR και Hand παρατηρήθηκαν ελαφρώς χαμηλότερες μετρικές, αλλά η ακρίβεια παρέμεινε υψηλή.

**Μείωση Διαστάσεων:**

**PCA**

Εφαρμόσαμε την τεχνική Principal Component Analysis (PCA) για να μειώσουμε τη διάσταση του dataset, διατηρώντας τη μεγαλύτερη δυνατή διακύμανση των δεδομένων χωρίς να χάσουμε σημαντική πληροφορία. Αφού τυποποιήσαμε τα δεδομένα του training set, χρησιμοποιήσαμε την PCA και κρατήσαμε τις πρώτες 50 κύριες συνιστώσες. Τα σχήματα των δεδομένων μετά την εφαρμογή της PCA είναι: Training set: (2800, 50), Validation set: (804, 50), Test set: (396, 50). Οι πρώτες 50 κύριες συνιστώσες εξηγούν το 90% της συνολικής διακύμανσης των δεδομένων. Ανακατασκευάσαμε εικόνες από τα μειωμένα δεδομένα και παρατηρήσαμε ότι οι ανακατασκευασμένες εικόνες είναι αρκετά παρόμοιες με τις πρωτότυπες. Οπτικοποιώντας τα δεδομένα, διαπιστώσαμε ότι η PCA βοήθησε στη σαφή διάκριση των κατηγοριών. Όταν εκπαιδεύσαμε το μοντέλο Λογιστικής Παλινδρόμησης με τα δεδομένα PCA, παρατηρήσαμε μια μικρή μείωση στην ακρίβεια (στο validation set από 99.50% σε 99.25% και από 99.75% σε 99.49% στο test set), ενώ η ακρίβεια στο training set παρέμεινε 100%. Παρά τη μικρή αυτή μείωση, τα αποτελέσματα παραμένουν πολύ καλά και το μοντέλο ήταν αποδοτικό.

Το PCA συνέβαλε στη μείωση της υπολογιστικής επιβάρυνσης, καθιστώντας την επεξεργασία των δεδομένων πιο αποδοτική. Τα confusion matrices έδειξαν κάποια μικρά σφάλματα ταξινόμησης στις κατηγορίες CXR και Hand, αλλά οι επιδόσεις παραμένουν πολύ καλές για τις περισσότερες κατηγορίες.

**LDA**

Η LDA μείωσε τις διαστάσεις των δεδομένων σε μόνο 2 διαστάσεις και το μοντέλο Λογιστικής Παλινδρόμησης (Logistic Regression) συνέχισε να αποδίδει καλά. Έδωσε στο training set: Ακρίβεια 100% (τέλεια προσαρμογή στα δεδομένα εκπαίδευσης). Σε σύγκριση όμως με την PCA, η LDA οδήγησε σε μικρότερη απόδοση στο validation set (93.66% έναντι 99.25%) και στο test set (92.42% έναντι 99.49%). Στο validation set, οι κατηγορίες CXR, ChestCT, και Hand παρουσίασαν μικρές μειώσεις στην ακρίβεια, αλλά η precision και recall παρέμειναν υψηλές. Επίσης, στο test set υπήρξε μεγαλύτερη πτώση στις μετρικές, ιδίως για τις κατηγορίες ChestCT και Hand.

Συγκριτικά, όμως με την PCA, η απόδοση του LDA είναι ελαφρώς χαμηλότερη. Αυτό μπορεί να οφείλεται στο ότι η PCA είναι πιο ευέλικτη σε πιο σύνθετα δεδομένα ή σε κατηγορίες με μεγαλύτερη αβεβαιότητα.

**t-SNE**

Εφαρμόζοντας την τεχνική t-SNE για μείωση διαστάσεων, παρατηρήσαμε ότι οι κατηγορίες διαχωρίζονται σαφώς σε δύο διαστάσεις για τα σύνολα εκπαίδευσης, επικύρωσης και δοκιμής. Αν και το t-SNE αποδεικνύεται χρήσιμο για την οπτικοποίηση και αναδεικνύει τις ιδιαίτερες σχέσεις μεταξύ των κατηγοριών, δεν συμβάλλει άμεσα στη βελτίωση των μετρικών του μοντέλου, καθώς δεν είναι κατάλληλο για την εκπαίδευση μοντέλων.

**LDA πάνω σε δεδομένα που έχουν υποστεί PCA**

Εφαρμόσαμε, επίσης, την τεχνική Linear Discriminant Analysis (LDA) πάνω σε δεδομένα που είχαν πρώτα υποστεί Principal Component Analysis (PCA), για να εξετάσουμε αν η συνδυασμένη χρήση των δύο μεθόδων θα μπορούσε να βελτιώσει την απόδοση του μοντέλου. Η εφαρμογή LDA μετά την PCA οδήγησε σε απόδοση στο train set 99.25% (χαμηλότερη από εκείνη που δίνει η εφαρμογή μόνο PCA) , στο validation set 99.00% (επίσης χαμηλότερη) και στο test set 98.99% (επίσης χαμηλότερη) .

Συμπερασματικά, από τις δοκιμές που πραγματοποιήσαμε καταλαβαίνουμε πως, η χρήση PCA μόνο φαίνεται να είναι η πιο αποδοτική μέθοδος για τη μείωση των διαστάσεων και την βελτιστοποίηση της απόδοσης σε νέα δεδομένα, καθώς έχει την υψηλότερη ακρίβεια στο test set χωρίς να προκαλεί σημαντική μείωση στην απόδοση. Για την βελτιστοποίηση του μοντέλου με PCA, εφαρμόζουμε την τεχνική Grid Search με Cross Validation για την επιλογή των καλύτερων υπερπαραμέτρων, χρησιμοποιώντας το Logistic Regression στα δεδομένα που έχουν υποστεί PCA.

**Κλασικοί Αλγόριθμοι Ταξινόμησης**

**Logistic Regression**

Χρησιμοποιήσαμε 5-fold cross-validation για να εξασφαλίσουμε τη γενίκευση του μοντέλου, και οι καλύτεροι υπερπαράμετροι που προέκυψαν ήταν: C = 0.1, penalty = 'l1', solver = 'liblinear'

Με τις συγκεκριμένες υπερπαράμετρους, το μοντέλο Logistic Regression πέτυχε accuracy 99.80% στο test set και ROC-AUC score 99.81% , υποδεικνύοντας εξαιρετική απόδοση στην διάκριση των κλάσεων. Οι τιμές precision και recall πλησίασαν το 1.00 για όλες τις κατηγορίες, ενώ η f1-score ήταν επίσης κοντά στο 1.00, δείχνοντας τη συνολική καλή απόδοση του μοντέλου. Συνεπώς, η εφαρμογή του Grid Search συνέβαλε στην εξαιρετική απόδοση του μοντέλου, ενώ η χρήση της τεχνικής PCA βοήθησε στη μείωση της διάστασης των δεδομένων χωρίς να επηρεάσει σημαντικά την ακρίβεια του μοντέλου.

Επιπλέον, οπτικοποιήσαμε τα decision boundaries στον 2D χώρο των πρώτων δύο κυριότερων συνιστωσών του PCA για να κατανοήσουμε πώς το μοντέλο ταξινομεί τα δεδομένα. Η οπτικοποίηση αυτή επιτρέπει τη σαφή διάκριση των κατηγοριών και βοηθά να κατανοήσουμε καλύτερα πώς το μοντέλο θα συμπεριφέρεται σε άγνωστα δεδομένα.

**Support Vector Machines**

Για την αξιολόγηση της απόδοσης του μοντέλου Support Vector Machine (SVM), εφαρμόσαμε τη μέθοδο SVM με γραμμικό πυρήνα (kernel='linear') πάνω σε δεδομένα που είχαν υποστεί Principal Component Analysis (PCA). Τα αποτελέσματα έδειξαν ότι το μοντέλο πετυχαίνει 100% ακρίβεια στο training set, 99.50% στο validation set και 99.49% στο test set, υποδεικνύοντας ότι το SVM γενικεύει καλά σε άγνωστα δεδομένα.

Οι επιδόσεις του μοντέλου ήταν πολύ καλές, με τα classification reports να δείχνουν υψηλές τιμές precision, recall και f1-score για όλες τις κατηγορίες. Οι confusion matrices επιβεβαίωσαν ότι τα λάθη ταξινόμησης ήταν ελάχιστα, δείχνοντας ότι το μοντέλο είχε εξαιρετική ακρίβεια στις προβλέψεις του.

Στη συνέχεια, με την τεχνική Grid Search βρήκαμε υπερπαραμέτρους για το μοντέλο SVM με πυρήνα RBF (Radial Basis Function). Εξετάσαμε διάφορους συνδυασμούς παραμέτρων, όπως C, τύπους πυρήνα (linear, rbf, poly) και παράμετρο gamma για τον RBF πυρήνα. Με την καλύτερη συνδυασμένη ρύθμιση C=1, gamma='scale' και kernel='rbf', το μοντέλο παρουσίασε εξαιρετικά αποτελέσματα με 100% ακρίβεια στο training set, 99.88% στο validation set και 99.75% στο test set.

Η AUC (Area Under the Curve) για όλες τις κλάσεις είναι στο 1, επιβεβαιώνοντας την εξαιρετική ικανότητα διάκρισης του μοντέλου. Επιπλέον, οπτικοποιήσαμε τα όρια απόφασης (decision boundaries) σε 2 διαστάσεις, για να κατανοήσουμε πώς το μοντέλο ταξινομεί τα δεδομένα στον εκπαιδευτικό χώρο και πώς διαχωρίζει τις κατηγορίες.

**k-Κοντινότεροι Γείτονες (k-NN)**

Επιλέξαμε τον αριθμό γειτόνων k = 5, καθώς αυτός ο αριθμός συνήθως εξασφαλίζει ισχυρή απόδοση χωρίς να εξαρτάται υπερβολικά από πολύ μικρό ή πολύ μεγάλο σύνολο γειτόνων. Το μοντέλο εκπαιδεύτηκε και αξιολογήθηκε στα σύνολα εκπαίδευσης, επικύρωσης και δοκιμής, με τα εξής αποτελέσματα: train accuracy: 99.64%, validation accuracy: 99.50%, test accuracy: 99.24%, δηλαδή απέδωσε εξαιρετικά τόσο στο σύνολο εκπαίδευσης όσο και στα άγνωστα δεδομένα (validation και test sets), με υψηλές τιμές για τους δείκτες precision, recall και f1-score σε όλες τις κατηγορίες. Οι confusion matrices επιβεβαίωσαν την υψηλή απόδοση, με ελάχιστα λάθη, κυρίως ανάμεσα σε κατηγορίες που είναι πιο δύσκολο να διακριθούν (π.χ. CXR και Hand).

Με βελτιστοποίηση υπερπαραμέτρων την τεχνική GridSearchCV με 5-fold cross-validation βρήκαμε τις βέλτιστες παραμέτρους: n\_neighbors = 3, algorithm = auto, metric = euclidean, weights = uniform. Με τις βελτιστοποιημένες παραμέτρους, το μοντέλο k-NN έδωσε train accuracy: 99.75%, validation accuracy: 99.75%, test accuracy: 99.75%. Τα classification reports επιβεβαίωσαν εξαιρετική απόδοση με πολύ υψηλές τιμές για precision, recall και f1-score σε όλες τις κατηγορίες. Επίσης, τα confusion matrices υποδεικνύουν ότι το μοντέλο έκανε πολύ λίγα λάθη. Η οπτικοποίηση ROC curves για κάθε κατηγορία έδειξε ότι το μοντέλο έχει εξαιρετική ικανότητα διάκρισης, με τις τιμές AUC (Area Under the Curve) να είναι πολύ κοντά στο 1, επιβεβαιώνοντας την εξαιρετική διάκριση των κατηγοριών από το μοντέλο. Επιπλέον, η οπτικοποίηση των ορίων απόφασης (decision boundaries) παρείχε μια σαφή εικόνα για το πώς το μοντέλο διαχωρίζει τις κατηγορίες στον εκπαιδευτικό χώρο.

**Naive Bayes**

Η ακρίβεια στο training set ήταν 98.21% , στο validation set 98.51% και στο test set 98.23%. Το classification report για όλα τα σύνολα δεδομένων έδειξε υψηλές τιμές για precision, recall και f1-score, με τις περισσότερες κατηγορίες να επιτυγχάνουν πολύ καλή ακρίβεια. Τα confusion matrices επιβεβαίωσαν ότι τα λάθη ήταν ελάχιστα, κυρίως μεταξύ κατηγοριών που ήταν πιο δύσκολο να διακριθούν, όπως CXR και Hand. Επιπλέον, χρησιμοποιήσαμε Grid Search για να βρούμε τις βέλτιστες παραμέτρους του μοντέλου, με αποτέλεσμα την βελτιστοποίηση της παραμέτρου var\_smoothing. Η εφαρμογή της συγκεκριμένης βελτίωσης ανέβασε την ακρίβεια του μοντέλου σχεδόν στο training set (98.36%) , αλλά όχι στο validation set (98.51%) και στο test set (98.23%). Τέλος, τα classification reports, οι confusion matrices, η ROC curve και το AUC , καθώς και η οπτικοποίηση των ορίων απόφασης, επιβεβαίωσαν την εξαιρετική ικανότητα του μοντέλου να διακρίνει σωστά τις κατηγορίες.

**Βαθιά Νευρωνικά Δίκτυα**

**Απλό πολυεπίπεδο νευρωνικό δίκτυο (MLP)**

Επιλέξαμε δύο κρυφά επίπεδα, καθένα από τα οποία έχει 128 νευρώνες. Η συγκεκριμένη επιλογή βασίζεται στην ιδέα ότι για πιο περίπλοκες κατανομές δεδομένων, τα πιο βαθιά δίκτυα (με περισσότερα επίπεδα και νευρώνες) μπορούν να μάθουν πιο σύνθετες σχέσεις. Το dropout χρησιμοποιείται για την αποφυγή της υπερπροσαρμογής (overfitting), επιτρέποντας στο μοντέλο να γενικεύει καλύτερα σε άγνωστα δεδομένα. Επιλέγουμε τη συνάρτηση ενεργοποίησης ReLU λόγω των καλών επιδόσεων σε πολλά προβλήματα και της απλότητάς της. Το L2 regularization (μέσω του weight\_decay στον Adam optimizer) αποτρέπει την υπερβολική προσαρμογή στις ειδικές περιπτώσεις του training set.

Αρχικά, οι ετικέτες του dataset κωδικοποιήθηκαν σε αριθμητική μορφή μέσω του LabelEncoder, καθώς είναι απαραίτητο για τη λειτουργία του μοντέλου ταξινόμησης, με PCA στις εικόνες και το τελευταίο επίπεδο να έχει τον αριθμό εξόδων ίσο με τον αριθμό των κατηγοριών. Στο πλαίσιο της κανονικοποίησης, χρησιμοποιήθηκε Dropout με ποσοστό 30% και L2 κανονικοποίηση μέσω του weight\_decay, ώστε να μειωθεί ο κίνδυνος υπερπροσαρμογής (overfitting).

Η εκπαίδευση του μοντέλου πραγματοποιήθηκε για 20 εποχές, χρησιμοποιώντας τον αλγόριθμο Adam και την συνάρτηση κόστους CrossEntropyLoss, κατάλληλη για πολυκατηγορηματική ταξινόμηση. Τα αποτελέσματα της εκπαίδευσης έδειξαν πως στην 20ή εποχή, η ακρίβεια για το training set ήταν 97.89%, για το validation set 99.13%, και για το test set 99.49%. Η απώλεια μειώθηκε σημαντικά κατά τη διάρκεια της εκπαίδευσης, υποδεικνύοντας συνεχιζόμενη βελτίωση του μοντέλου. Τα γραφήματα που δημιουργήθηκαν για τις ακρίβειες και τις απώλειες επιβεβαιώνουν ότι το μοντέλο εκπαιδεύεται σωστά, χωρίς υπερπροσαρμογή, και καταφέρνει να γενικεύσει πολύ καλά στα δεδομένα τεστ.

Για τη βελτιστοποίηση του μοντέλου, πραγματοποιήθηκαν πειράματα με διαφορετικές υπερπαραμέτρους, όπως συναρτήσεις ενεργοποίησης (ReLU, LeakyReLU, Tanh), αριθμό κρυφών επιπέδων και νευρώνων, ρυθμούς μάθησης και ποσοστά dropout. Η εκπαίδευση έγινε για 20 εποχές, με χρήση του Adam optimizer και L2 κανονικοποίησης. Το καλύτερο μοντέλο προέκυψε από τη ρύθμιση με δύο κρυφά επίπεδα (256, 128), τη συνάρτηση ενεργοποίησης ReLU, ρυθμό μάθησης 0.01 και ποσοστό dropout 0.5, επιτυγχάνοντας τελικά ακρίβεια 99.68% στην εκπαίδευση, 99.50% στο validation και 98.99% στο test set. Οι τεχνικές κανονικοποίησης και τα διαφορετικά επίπεδα dropout συνέβαλαν στην αποφυγή υπερπροσαρμογής και στην ενίσχυση της γενίκευσης του μοντέλου, επιτυγχάνοντας εξαιρετικά αποτελέσματα. Τα ROC Curves και οι υπολογισμοί AUC επιβεβαίωσαν τη μεγάλη ικανότητα διάκρισης του μοντέλου για κάθε κατηγορία. Τα classification reports και οι confusion matrices για τα test, train, και validation σύνολα δεδομένων δείχνουν ότι το μοντέλο έχει άριστη απόδοση, χωρίς υπερπροσαρμογή, και κατατάσσει σωστά τις κατηγορίες, με precision και recall στο 1. Αυτά τα αποτελέσματα δείχνουν ότι το μοντέλο είναι αποτελεσματικό και γενικεύεται καλά σε νέα δεδομένα.

**Συνελικτικό Νευρωνικό Δίκτυο (CNN) για δεδομένα εικόνας**

Δημιουργήθηκε ένα CNN με 3 επίπεδα συνελίξεων και δύο πλήρως συνδεδεμένα επίπεδα, με σκοπό την αναγνώριση των χαρακτηριστικών των εικόνων και την κατηγοριοποίησή τους. Η επιλογή των τριών επιπέδων συνελίξεων βασίστηκε στην ανάγκη να ανιχνεύσουμε τόσο τα τοπικά όσο και τα σύνθετα χαρακτηριστικά των εικόνων, με το πρώτο επίπεδο να εξάγει βασικά χαρακτηριστικά όπως άκρα και υφές, το δεύτερο να ανιχνεύει σχήματα και πιο σύνθετα μοτίβα, και το τρίτο να συνδυάζει όλα αυτά τα χαρακτηριστικά για να κατανοήσει πιο περίπλοκες σχέσεις στην εικόνα. Τα δύο πλήρως συνδεδεμένα επίπεδα (fc1, fc2) επιλέχθηκαν για την κατηγοριοποίηση των εξαγόμενων χαρακτηριστικών και την αποδοτική ταξινόμηση των εικόνων στις αντίστοιχες κατηγορίες τους.

Κατά τη διάρκεια της εκπαίδευσης, το CNN παρουσίασε εξαιρετικά αποτελέσματα. Η ακρίβεια του μοντέλου κατά την εκπαίδευση έφτασε το 100% κατά τις τελευταίες εποχές, ενώ η απώλεια (loss) μειώθηκε σημαντικά, φτάνοντας σχεδόν στο μηδέν. Στην επικύρωση και τη δοκιμή, η ακρίβεια του μοντέλου ήταν επίσης εξαιρετική, φτάνοντας το 99.75%, κάτι που δείχνει ότι το μοντέλο κατέχει την ικανότητα να γενικεύει σε νέα δεδομένα και να αποφεύγει την υπερπροσαρμογή.

Στη συνέχεια, προχωρήσαμε σε διάφορους πειραματισμούς για να βελτιώσουμε την απόδοσή του. Συγκεκριμένα, εφαρμόσαμε Dropout με ποσοστό 50% πριν από το πλήρως συνδεδεμένο επίπεδο του δικτύου, προκειμένου να μειώσουμε την εξάρτηση από οποιοδήποτε συγκεκριμένο νευρώνα και να ενισχύσουμε τη συνολική ικανότητα του δικτύου να γενικεύει σε νέα δεδομένα. Επιπλέον, πειραματιστήκαμε με διαφορετικές συναρτήσεις ενεργοποίησης, όπως οι LeakyReLU και ELU, πέρα από τη συνήθη ReLU, για να εξετάσουμε την επίδρασή τους στην απόδοση του μοντέλου. Οι συγκεκριμένες συναρτήσεις ενεργοποίησης επιλέχθηκαν για να αντιμετωπίσουν πιθανά προβλήματα όπως το "vanishing gradient" που μπορεί να προκύψει από τη χρήση της ReLU σε βαθιά δίκτυα. Επίσης, εξετάσαμε την επιλογή του ρυθμού μάθησης (learning rate), καθώς έχει σημαντική επίδραση στη σύγκλιση του μοντέλου. Δοκιμάσαμε δύο τιμές για τον ρυθμό μάθησης (0.001 και 0.0001), προκειμένου να βρούμε την καλύτερη ρύθμιση για την εκπαίδευση του δικτύου. Επιπλέον, προσθέσαμε L2 Regularization (Weight Decay) στον Adam optimizer για να περιορίσουμε την υπερβολική αύξηση των βαρών, ενισχύοντας τη γενίκευση του μοντέλου.

Στην εκπαίδευση, καταγράψαμε τις καμπύλες απώλειας και ακρίβειας για κάθε πείραμα, τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα επικύρωσης, με στόχο να αναλύσουμε την επίδραση των παραμέτρων στην απόδοση του μοντέλου. Προέκυψε πως οι καλύτερες υπερπαράμετροι είναι: αριθμός νευρώνων = [256, 128], activation function = 'ReLU', συντελεστής κανονικοποίησης L2 = 0.01, dropout = 0.5). Για την αξιολόγηση του μοντέλου, υπολογίσαμε το test loss και την test accuracy στο σύνολο δοκιμής (test set), τα οποία έδειξαν εξαιρετική απόδοση με test loss = 0.0232 και test accuracy = 99.75%.

Χρησιμοποιώντας classification report, confusion matrix και ROC curve, παρατηρούμε πως το μοντέλο έχει άριστη απόδοση με ακρίβεια 100% σε όλα τα σύνολα και AUC=1.00 για κάθε κατηγορία. Συνεπώς το μοντέλο κατατάσσει σωστά όλες τις κατηγορίες με τέλεια ακρίβεια και εξαιρετική ικανότητα διάκρισης μεταξύ των κλάσεων.

ΠΑΡΑΡΤΗΜΑ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Accuracy | Without dimensional reduction | PCA | LDA | t-SNE | LDA after PCA |
| Training | 1.0000 | 1.0000 | 1.0000 | 0.9907 | 0.9925 |
| Validation | 0.9950 | 0.9925 | 0.9366 | 0.9751 | 0.9900 |
| Test | 0.9975 | 0.9949 | 0.9242 | 0.0126 | 0.9899 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Accuracy  Test set | Logistic Regression | Support Vector Machines | k-Κοντινότεροι Γείτονες (k-NN) | Naive Bayes |
| Πριν το Grid Search | 0.9949 | 0.9949 | 0.9924 | 0.9823 |
| Μετά το Grid Search | 0.9980 | 0.9975 | 0.9975 | 0.9823 |

Βιβλιογραφία

<https://scikit-learn.org/stable/modules/lda_qda.html>

<https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

<https://www.w3schools.com/python/python_ml_knn.asp>

<https://scikit-learn.org/stable/modules/naive_bayes.html>