VisDiceX^2: Σύγχρονες Τεχνολογίες Μηχανικής Όρασης. Μια Αναλυτική Επισκόπηση και Εφαρμοσμένη Έρευνα με Ζάρια

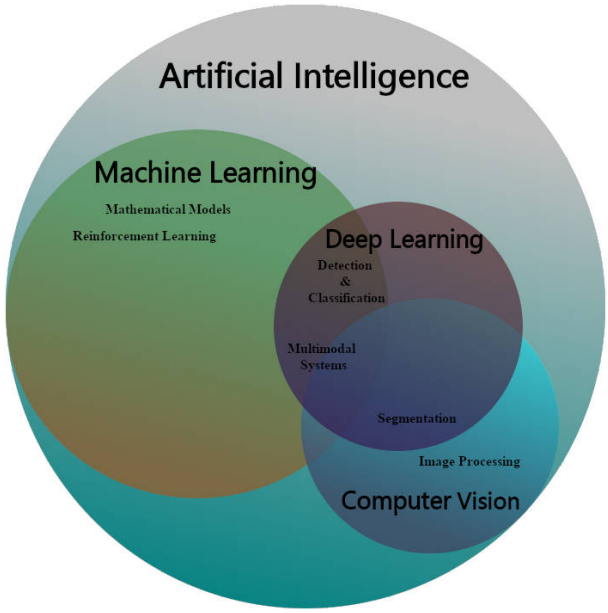
1st Dimitrios Mpouziotas
Department of Computer Science & Networking
University of Ioannina
Arta, Greece
mpouziotasd@gmail.com

***Περίληψη*—Η Τεχνητή Νοημοσύνη, αποτελεί έναν ευρύ τομέα εξερεύνησης και έρευνας με ένα μεγάλο υποτομέα της να είναι η Μηχανική όραση. Επικεντρώνεται στην πρόβλεψη και δημιουργία νέων πληροφοριών μέσω της αναγνώρισης προτύπων από διάφορους τύπους δεδομένων (modalities) με την χρήση νευρωνικών δικτύων). Η Μηχανική Όραση, αξιοποιεί τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) για την εξαγωγή πληροφορίας και την αναγνώριση προτύπων σε εικόνες και βίντεο. Συγκεκριμένα, εξειδικεύει στην ανίχνευση αντικειμένων στις εικόνες. Η παρούσα μελέτη πραγματοποιεί μια ανασκόπηση σε σύγχρονες τεχνικές μηχανικής όρασης (όπως Image Classification, Object Detection, Object Tracking, Semantic Segmentation, Visual Question Answering). Σε συνδυασμό με την ανασκόπηση, εκτελείται ένα πείραμα το οποίο έχει ως στόχο την εκπαίδευση ενός μοντέλο με τεχνικές ανίχνευσης αντικειμένων (Object Detection) με σκοπό την ανίχνευση πλευρών ζαριών, όπου η κάθε πλευρά, απεικονίζει και ένα νούμερο. Στο πείραμα αυτό αξιολογούνται τέσσερα διαφορετικά μοντέλα ανίχνευσης αντικειμένων της οικογένειας YOLO [1] (You Only Look Once): yolov8-M, yolov8-X, yolov10-M, yolov10-X. Τέλος, αναπτύχθηκε μια εφαρμογή που επιτρέπει στους χρήστες να εκτελούν τρείς βασικές λειτουργίες: Συμπέρασμα (Inference), Εκπαίδευση (Training) και Αξιολόγηση (Evaluation).**

***Index Terms*—Artificial Intelligence, Computer Vision, Object Detection, Dice Detection, YOLO (You Only Look Once),**

I. Εισαγωγή

Η τεχνητή νοημοσύνη έχει εμφανίσει ραγδαία ανάπτυξη σε εφαρμογές που έχουν αναπτυχθεί μέχρι στιγμής στη βιβλιογραφία, όπως και τα νευρωνικά δίκτυα που έφεραν την επανάσταση σε δύσκολα επιλύσιμα προβλήματα. Οι εφαρμογές της Τεχνητής Νοημοσύνης αυτοματοποιούν πολλές διαδικασίες που παραδοσιακά απαιτούσαν ανθρώπινη παρέμβαση, ενώ παράλληλα προσφέρουν λύσεις σε προβλήματα που δεν είναι προσεγγίσιμα από τον άνθρωπο. Ένας μεγάλος υποτομέας της Τεχνητής Νοημοσύνης αποτελεί τη μηχανική όραση, που συμφέρει αρκετά στην αυτοματοποίηση αρκετών διαδικασιών μειώνοντας σημαντικά την ανθρώπινη παρέμβαση. Η εικόνα 1, εμφανίζει την συνολική δομή της τεχνητής νοημοσύνης, όπως και τους αντίστοιχους υποτομείς του.

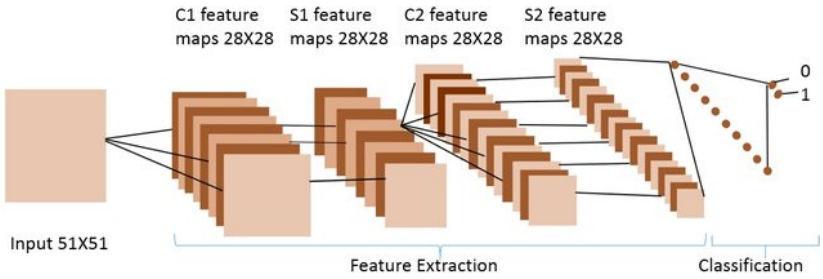


Σχήμα 1: Ένα διάγραμμα που εικονογραφεί την δομή της τεχνητής νοημοσύνης.

Στην έρευνα αυτή, θα κάνουμε μια σημαντική επισκόπηση στις τεχνολογίες της μηχανικής όρασης και παράλληλα, θα υλοποιήσουμε ένα έξυπνο σύστημα, το οποίο λειτουργεί αποτελεσματικά σε πραγματικό χρόνο.

II. Η Μηχανική Όραση

Η μηχανική όραση, όπως αναφέρθηκε προηγουμένως, αποτελεί έναν εκτενή υποτομέα της τεχνητής νοημοσύνης, ο οποίος περιλαμβάνει και τους δικούς του υποτομείς. Προτού διευκρινίσουμε τον κάθε υποτομέα της μηχανικής όρασης, πρέπει να κατανοήσουμε τον λόγο που την χρησιμοποιούμε, τι προβλήματα επιθυμούμε να λύσουμε και τι επιτυγχάνουμε με αυτές τις τεχνολογίες. Σκοπός αυτού του τομέα είναι η αναγνώριση προτύπων και χαρακτηριστικών σε εικόνες ή βίντεο. Αν και η μηχανική όραση συχνά συγκρίνεται με την επεξεργασία εικόνας, οι δύο τομείς δεν ταυτίζονται, καθώς η επεξεργασία εικόνας αποτελεί υποστηρικτικό εργαλείο για τη μηχανική όραση. Οι ποιο βασικές και γνωστές τεχνολογίες



Σχήμα 2: Βασική Δομή ενός Συνελικτικού Νευρωνικού Δικτύου. Source: [2].

που αξιοποιούνται στη μηχανική όραση, είναι τα συνελικτικά νευρωνικά δίκτυα, τα οποία εξειδικεύουν στην εξαγωγή πληροφορίας από εικόνα. Για παράδειγμα, για την ανίχνευση ενός αντικειμένου σε μια εικόνα, το συνελικτικό νευρωνικό δίκτυο, με υπόθεση ότι έχει προ-εκπαιδευτή στο να ανιχνεύει ένα είδος αντικειμένου, εξάγει σημαντικά χαρακτηριστικά από την εικόνα. Με βάση αυτά που έχει εκπαιδευτεί, υπολογίζει μια συνάφεια για μια ανίχνευση που εντόπισε, αν ταιριάζει με τα χαρακτηριστικά που έχει εκπαιδευτεί, τότε αυτό το αποκαλούμε '*Πρόβλεψη*' ή διαφορετικά '`Detection`.

Α. Συνελικτικά Νευρωνικά Δίκτυα

Η βασική δομή ενός Συνελικτικού Νευρωνικού Δικτύου αποτελείται από δύο βασικά στάδια: Το επίπεδο εξαγωγής χαρακτηριστικών (Feature Extraction) και το επίπεδο ταξινόμησης [2].

Στην απλούστερη περίπτωση, το Συνελικτικό Νευρωνικό Δίκτυο (Convolutional Neural Network - CNN) αποτελείται από δύο επίπεδα: Το επίπεδο εξαγωγής χαρακτηριστικών (Feature Extraction) και το επίπεδο ταξινόμησης (Classification Layer) [2].

Το σχήμα 2, εμφανίζει τα διάφορα στάδια του Συνελικτικού Νευρωνικού Δικτύου. Το πρώτο στάδιο η εισαγωγή μιας εικόνας και την εξαγωγή χαρακτηριστικών από αυτήν και με τελικό στόχο το επίπεδο ταξινόμησης, στο οποίο γίνεται και η πρόβλεψη.

Στα επόμενα σημεία της εργασίας, θα χρησιμοποιήσουμε τις έννοιες "Training" για την εκπαίδευση των νευρωνικών δικτύων και "Inference" οταν βγάζουμε συμπέρασμα από αυτά.

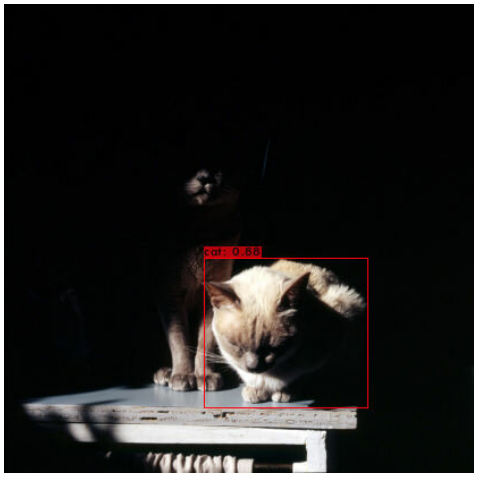
III. Σύγχρονες Τεχνικές και Μεθοδολογίες

Η μηχανική όραση αποτελείται από ένα μεγάλο εύρος μεθοδολογιών με την κάθε μεθοδολογία να ολοκληρώνει τον δικό της στόχο.

- **Image Classification (Ταξινόμηση Εικόνας)**: Η κλασσική περίπτωση της μηχανικής όρασης, στην οποία το CNN ταξινομεί την εικόνα με μια κατηγορία από ένα σύνολο κατηγοριών που έχει εκπαιδευτή. Η τεχνική αυτή υλοποιείται σε όλη την εικόνα.

  Μια αρκετά φημισμένη έρευνα το 2012, που έφερε επανάσταση στη μηχανική όραση [3] δημιούργησε ένα μεγάλο CNN μοντέλο με το όνομα ImageNet. Το μοντέλο εκπαιδεύτηκε σε ένα dataset που αποτελείται από 15 εκατομμύρια εικόνες και περίπου 22.000 κατηγορίες.

- **Object Detection (Ανίχνευση Αντικειμένων)**: Η διαδικασία η οποία ανιχνεύονται πολλαπλά αντικείμενα σε μια εικόνα. Αξιοποιείται συχνά και για τον διαχωρισμό δύο ή περισσότερον ειδών αντικειμένων (Πχ Γάτα ή Σκύλος). Κατά το inference, το μοντέλο οριοθετεί κουτιά (Bounding Boxes) στο κάθε ανιχνευμένο αντικείμενο στην εικόνα 3.
  Μια έρευνα χρησιμοποίησε τεχνικές μηχανικής όρασης για την ανίχνευση άγριας ζωής με drone στον Αμβρακικό Κόλπο.



Σχήμα 3: Παράδειγμα ανίχνευσης αντικειμένου μιας γάτας. Source: [4]

  Μια μελέτη [5] εστίασε στην ανίχνευση αντικειμένων της πανίδας του Αμβρακικού Κόλπου, με ιδιαίτερη έμφαση στην καταγραφή και παρακολούθηση της ορνιθοπανίδας.

- **Object Tracking (Παρακολούθηση Αντικειμένων)**: Η τεχνική αυτή έχει ως στόχο την συνεχής παρακολούθηση ενός ή περισσότερων αντικειμένων σε μια ροή από εικόνες ή αλλιώς βίντεο. Αυτό επιτυγχάνεται με το να τοποθετείται μια αριθμητική ετικέτα (ID) και να επικρατείτε στο ίδιο αντικείμενο στις επόμενες εικόνες. Η τεχνική αυτή συχνά χρησιμοποιείται χέρι με χέρι με την ανίχνευση αντικειμένων.

Η ίδια έρευνα που εστίασε στην παρακολούθηση της ορνιθοπανίδας συνδύασε τεχνικές Object Tracking με σκοπό την εκτίμηση του πληθυσμού της ορνιθοπανίδας [6].

- **Semantic Segmentation (Σημασιολογική Τμηματοποίηση)**: Μια τεχνική που κατηγοριοποιεί αποτελεσματικά τις περιοχές μιας εικόνας σε διακριτές κατηγορίες. Για παράδειγμα, σε μια εικόνα, τα pixel που αντικατροπτίζουν σε ένα άτομο θα χαρακτηριστιούν με έναν αριθμητικό δείκτη που αναφέρεται στην κατηγορία ”Person”, ενώ τα pixel fνός δρόμου θα χαρακτηριστιούν με έναν αριθμητικό δείκτη που αναφέρεται στην κατηγορία ”Δρόμος”. Το αποτέλεσμα που παράγεται είναι ένα πολύγωνο που επικαλύπτει την περιοχή των αντικειμένων.

  Ένα μοντέλο με το όνομα SegFormer [7] εκπαιδεύτηκε σε ένα dataset με 250 κατηγορίες (Dataset: ADE20k [8])

- **Visual Question Answering (Οπτική Απαντητική Ερώτηση)**: Μια σχετικά νέα τεχνική η οποία αξιοποιεί Multi-modality τεχνικές για την δημιουργία συσχέτισης μεταξύ κειμένου και εικόνας ταυτόχρονα. Ο χρήστης μπορεί να κάνει μια ερώτηση σε μια εικόνα και να το περάσει στο μοντέλο, το μοντέλο ύστερα θα επιστρέψει την απάντηση με βάση τα στοιχεία που υπάρχουν στην εικόνα.
  Στην συνέχεια της έρευνας της παρακολούθησης της ορνιθοπανίδας, το μοντέλο

IV. Το Πείραμα

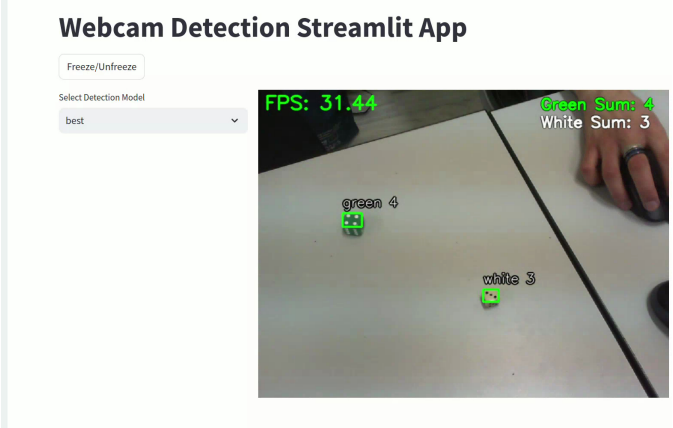
A. Στόχος Πειράματος

Στόχος του πειράματος είναι η χρήση μοντέλων ανίχνευσης και τμηματοποίησης για την ανίχνευση των ζαριών και των αριθμών τους. Η μάσκα που παράγεται από τη διαδικασία τμηματοποίησης αξιοποιείται για την εξαγωγή του χρώματος κάθε ζαριού. Με βάση το χρώμα, υπολογίζεται το άθροισμα των αριθμών για κάθε ομάδα ζαριών με ίδιο χρώμα.

Παράλληλα, η απόδοση και ακρίβεια των μοντέλων αξιολογούνται με ένα dataset ζαριών, με τη χρήση μετρικών όπως mAP (mean Average Precision) και mR (mean Recall) για την ανίχνευση αντικειμένων.

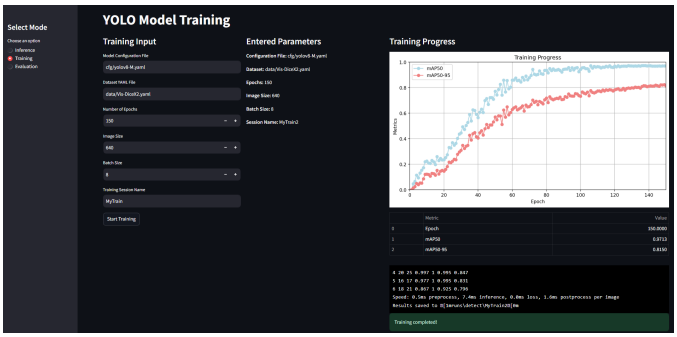
Στόχος του πειράματος είναι η ανάπτυξη μιας web-based εφαρμογής υψηλού επιπέδου με την χρήση του εργαλείου Streamlit. Η εφαρμογή επιτυγχάνει τρεις βασικές λειτουργίες: Inference, Training και Evaluation με την χρήση μοντέλων μηχανικής όρασης, συγκεκριμένα στην τεχνική ανίχνευσης αντικειμένων.

- **Inference**: Η διαδικασία τεκμηρίωσης ενός εκπαιδευμένου μοντέλου. Στην εφαρμογή υλοποιήθηκε ένα dropdown το οποίο επιτρέπει τον χρήστη να επιλέξη από ένα σύνολο προ εκπαιδευμένων μοντέλων μηχανικής όρασης.



Σχήμα 4: Inference: Αποτελέσματα ανίχνευσης ζαριών.

- **Training**: Εκπαίδευση μοντέλου μηχανικής όρασης με βάση διάφορων παραμέτρων. Ο χρήστης δαχτυλογραφεί τους παραμέτρους για την εκπαίδευση του μοντέλου και μπορεί να προβάλει σε πραγματικό χρόνο την εξέλιξη της εκπαίδευσης στην εφαρμογή στη μορφή γραφημάτων και πίνακα.



Σχήμα 5: Training: mAP (50 & 50-95) και αποτελέσματα εκπαίδευσης με γράφημα.

- **Evaluation**: Αξιολόγηση μοντέλων και προβολή αποτελεσμάτων μοντέλων ανίχνευσης αντικειμένων σε γράφημα.

Σχήμα 6: Evaluation: Αποτελέσματα σύγκρισης μοντέλων mAP50-95 ως προς latency.

Ταυτόχρονα, στο πείραμα αξιοποιήθηκαν τεχνικές ανίχνευσης χρωμάτων με την χρήση απλών τεχνικών επεξεργασίας εικόνας.

B. Dataset Ζαριών

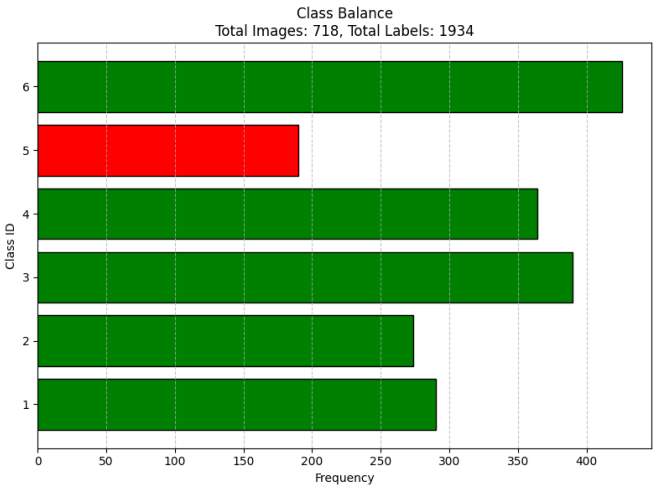
Για την εκπαίδευση και αξιολόγηση των μοντέλων χρησιμοποιήθηκε ένα dataset που περιλαμβάνει εικόνες με χρωμα-

τιστά ζάρια σε διάφορες γωνίες και συνθήκες φωτισμού από το Roboflow [9]. Αυτό μας είναι απαραίτητο με σκοπό να επιλύσουμε το πρόβλημα που τα δεδομένα κατά το τεστάρισμα διαφέρουν από τα training δεδομένα. Η δομή του dataset περιγράφεται στον Πίνακα I. Το dataset αποτελεί 359 εικόνες και 967 με 2.7 ετικέτες ανά εικόνα (μέσο). Οι ετικέτες αντιστοιχούν στις πλευρές των ζαριών, δηλαδή στους αριθμούς από το 1 έως το 6. Ωστόσο, τόσο το dataset όσο και το μοντέλο θεωρούν τις κλάσεις ως αριθμητικές τιμές που ξεκινούν από το 0, συνεπώς οι κλάσεις είναι από το 0 έως το 5.

Σύνολο Εικόνων	Σύνολο Κατηγοριών	Σύνολο Ετικετών
718	1934	6

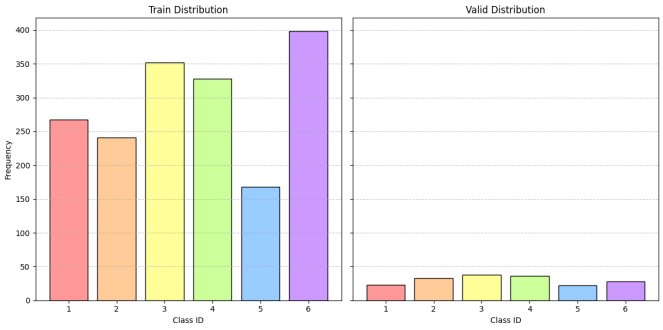
Πίνακας I: Η δομή του dataset ”6 Sided Dice Dataset” [9]

Η κατανομή της κάθε κατηγορίας για όλο το dataset περιγράφεται στο Σχήμα 7. Μπορούμε να παρατηρήσουμε πως υπάρχει ανισορροπία στις κατηγορίες ειδικά για το αριθμό ζαριού ’5’ με μόνο 190 συνολικές ετικέτες.



Σχήμα 7: Η αρχική κατανομή της κάθε κατηγορίας του dataset [9]

Το dataset κατανεμήθηκε σε δύο διαφορετικά τμήματα, Validation και Training. Η κατανομή του Validation dataset είναι 10% το κάθε ένα και το training dataset 80% όλου του dataset. Η εικόνα 8 παρουσιάζει την κατανομή της κάθε κατηγορίας των test, valid και training dataset.



Σχήμα 8: Η κατανομή της κάθε κατηγορίας των test, valid και train dataset.

Το διανεμημένο dataset που εμφανίζεται στην εικόνα αξιοποιήθηκε για την εκπαίδευση των μοντέλων του πειράματος και ύστερα για το τεστάρισμα του.

Β. Μοντέλα Πειράματος

Κατά την ανάπτυξη του πειράματος, εκπαιδεύτηκαν και αξιολογήθηκαν δύο διαφορετικά μοντέλα ανίχνευσης αντικειμένων YOLOv8 [10] & YOLOv10 [11]. Επιπλέον για το κάθε ένα χρησιμοποιήθηκαν διαφορετικά μεγέθη με το postfix ’M’ (Medium) & ’X’ (X-Large). Το dataset που αξιοποιήθηκε είναι ένα σύνολο από εικόνες με ζάρια και ετικέτες με την κάθε πλευρά του ζαριού. Η αξιολόγηση των μοντέλων πραγματοποιήθηκε με τη μέτρηση της ακρίβειας (Precision), της ανάκλησης (Recall) και του δείκτη mAP (Mean Average Precision) για κάθε μοντέλο, ώστε να επιλεχθεί το βέλτιστο για την εργασία.

Τα μοντέλα που αξιοποιήθηκαν στο πείραμα περιγράφονται στον παρακάτω πίνακα II. Ο πίνακας περιγράφει για το κάθε μοντέλο τα στοιχεία όπως το task (μέθοδος) που εκτελεί το κάθε μοντέλο, το dataset που αξιολογήθηκε όπως, την ακρίβεια του (mAP) και τέλος τους συνολικούς παραμέτρους και τα Flops.

Μοντέλο	**Task**	**mAP^{val}** 50-95	**Dataset**	**Parameters** (M)	**Flops** (G)
YOLOv8-M	Object Detection	50.2%	MSCOCO [12]	25.9	78.9
YOLOv8-X	-||-	53.9%	MSCOCO	68.2	257.8
YOLOv10-M	-||-	51.1%	MSCOCO	15.4	59.1
YOLOv10-X	-||-	54.4	MSCOCO	29.9	160.4

Πίνακας II: Μοντέλα Ανίχνευσης Αντικειμένων: Παράμετροι, ταχύτητες και GFlops. Πηγές Δεδομένων: [10] [11]

Η βασική διαφορά των δύο μοντέλων YOLOv8 και v10 είναι αρχιτεκτονική τους, όπως και ο τρόπος ο οποίος εκπαιδεύονται τα δύο μοντέλα. Αλλά αυτό που διαφέρει περισσότερο το ένα μοντέλο με το άλλο είναι πως ότι το YOLOv10 δεν χρησιμοποιεί τεχνικές NMS [13] (Non-Maximum Supression) κατά το Inference. Η NMS τεχνική αξιοποιείται στο τελευταίο βήμα κατά το Inference, με σκοπό να φιλτράρει τις ανιχνεύσεις και να χρησιμοποιήσει μόνο την πιο σχετική ανίχνευση, Σχήμα 9.

Σχήμα 9: Παράδειγμα Τεχνικής Non-Maximum Suppression σε συνδυασμό με την ανίχνευση ανθρώπων. Source: [13]

Η εξαίρεση της NMS συνάρτησης στο YOLOv10 όπως και η διαφορετική αρχιτεκτονική του, είχε ως συμπέρασμα το v10 μοντέλο να απαιτεί περισσότερο χρόνο να εκπαιδευτεί και να πλησιάσει το "Convergence" [1].

V. Αποτελέσματα Πειράματος

A. Μετρήσεις Αξιολόγησης Μοντέλων

Για την αξιολόγηση της ακρίβειας των μοντέλων η ορθότητα των ανιχνεύσεων κατηγοριοποιείται σε τέσσερις καταστάσεις και ύστερα με αυτές υπολογίζεται η συνάφεια της απόδοσης και ακρίβειας του μοντέλου. Υπάρχουν τρεις βασικές μετρήσεις αξιολόγησης [14] της απόδοσης και ακρίβειας μοντέλων, αυτές αποτελούν το Mean Average Precision (**mAP**), Mean Recall (**mR**) και το Intersection over Union (**IoU**).

- **True Positive** - TP (Αληθινό Θετικό): Η ανίχνευση η οποία η κλάση είναι σωστή σύμφωνα με το dataset και ισχύει ≥ 50 IoU.
- **False Positive** - FP (Ψευδές Θετικό): Η ανίχνευση η οποία η κλάση είναι λάθος σύμφωνα με το dataset και ισχύει ≥ 50 IoU.
- **True Negative** - TN (Αληθινό Αρνητικό): Η αρνητική ανίχνευση η οποία ανιχνεύθηκε σωστά σύμφωνα με το dataset.
- **False Negative** - FN (Ψευδές Αρνητικό): Η αρνητική ανίχνευση η οποία ανιχνεύθηκε λάθος σύμφωνα με το dataset.

mean Average Precision - (mAP): Η μέση ακρίβεια των συνολικών ανιχνεύσεων του μοντέλου σύμφωνα με το dataset. Για τον υπολογισμό της ακρίβειας αξιοποιείτε η ορθότητα των ανιχνεύσεων.
Ο τύπος της μέσης ακρίβειας είναι ο παρακάτω:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

[1]**Training Convergence:** Η διαδικασία κατά την οποία η ακρίβεια του μοντέλου κατά την εκπαίδευση θεωρείται ότι έχει φτάσει σε ένα σημείο πλατό, δηλαδή δεν μπορεί να βελτιωθεί άλλο.

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \tag{2}$$

Όπου $N = 1, 2, ..k$ το σύνολο εικόνων, Το AP_i (Average Precision) η μέση ακρίβεια κάθε εικόνας.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

$$\text{mR} = \frac{1}{N}\sum_{i=1}^{N} \text{R}_i \tag{4}$$

Όπου $N = 1, 2, ..k$ το σύνολο εικόνων, Το R_i (Mean Recall) η μέση ανάκληση κάθε εικόνας.
Στην ουσία με την μέτρηση της ανάκλησης (Recall), μπορούμε να δούμε πόσο συχνά το μοντέλο ανιχνεύει σύμφωνα **Intersection over Union** - IoU: Η διασταύρωση πάνω από την ένωση είναι μια τεχνική αξιολογησείς η οποία συχνά χρησιμοποιείται στον τομέα ανίχνευσης αντικειμένων. Στη πράξη, αυτή η μέτρηση συγκρίνει την ποσότητα εμβαδού του κουτιού το οποίο έχει ετικετοποιηθεί σε ένα dataset με την ίδια την ανίχνευση του μοντέλου.

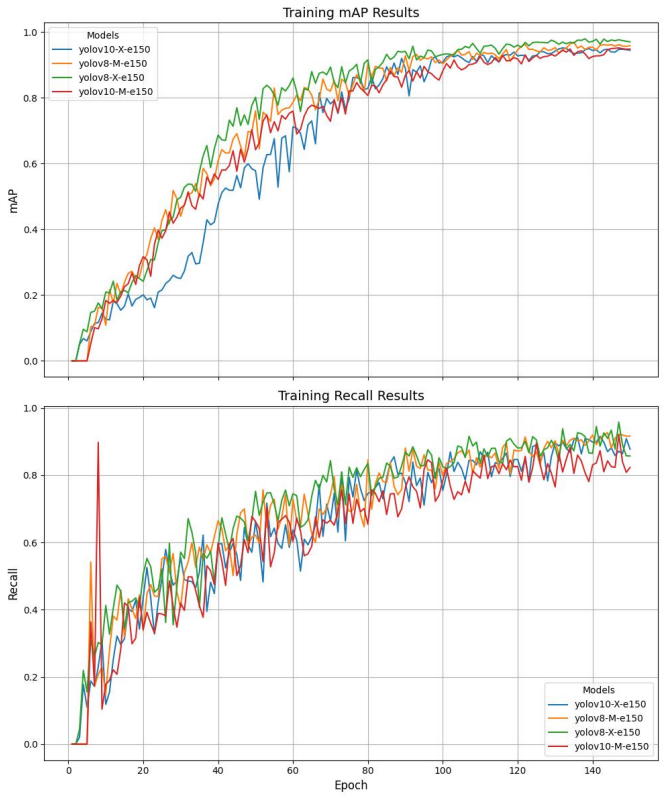
$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{5}$$

$$\begin{aligned}\text{Area of Union} = &+\text{Area of Detection} \\ &+ \text{Area of Groundtruth} \\ &- \text{Area of Overlap}\end{aligned} \tag{6}$$

Όπου Area of Overlap το εμβαδόν της ανίχνευσης με το εμβαδό της ετικέτας του dataset, Area of Detection το εμβαδόν της ανίχνευσης, Area of Groundtruth το εμβαδόν της ετικέτας.

B. Εκπαίδευση Μοντέλων

Κατά την εκπαίδευση εκτελέστηκαν διάφορα πειράματα στα μοντέλα ανίχνευσης. Όπως αναφέρθηκε προηγουμένως, τα μοντέλα που χρησιμοποιήθηκαν και εκπαιδεύτηκαν είναι το YOLOv8 και YOLOv10 με την βασική διαφορά μεταξύ τους είναι πως ότι το v10 δεν αξιοποιεί NMS συνάρτηση. Τα μοντέλα εκπαιδεύτηκαν με την κατανεμημένη μορφή του dataset συγκεκριμένα με την χρήση του training dataset και τεσταρίστηκαν με την χρήση του Validation Dataset. Το κάθε μοντέλο εκπαιδεύτηκε στις 150 εποχές, με σκοπό η ακρίβεια των μοντέλων να φτάσει ένα στάδιο "Convergence". Το Σχήμα 10 εμφανίζει τις μετρήσεις αξιολόγησης mAP και Recall. Στο mAP γράφημα, μπορούμε να παρατηρήσουμε ότι το μοντέλο YOLOv10 απαιτεί περισσότερο χρόνο να βελτιώσει την ακρίβεια σε σύγκριση με το YOLOv8.



Σχήμα 10: Αποτελέσματα Εκπαίδευσης Μοντέλων

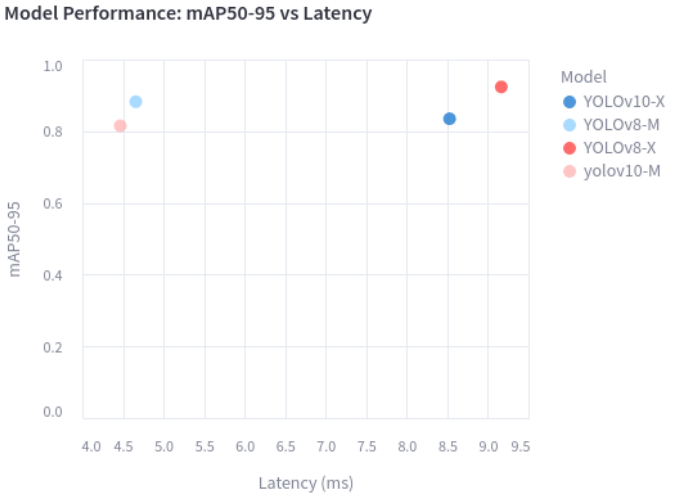
Γ. Αξιολόγηση Μοντέλων

Ο πίνακας III εμφανίζει τα αποτελέσματα των αξιολογήσεων για το κάθε μοντέλο του πειράματος στο Validation dataset. Μπορούμε να παρατηρήσουμε πως ότι ακόμα και αν το version του μοντέλου είναι μεγαλύτερο, δεν σημαίνει πως ότι και η απόδοση των μοντέλων είναι καλύτερη όσον αφορά την ακρίβεια. Ωστόσο, το μοντέλο με την καλύτερη επίδοση από όλα τα υπόλοιπα είναι το YOLOv8-X. Οι προηγούμενες έρευνες ανέφεραν πως ότι το μοντέλο YOLOv10 εμφάνισε καλύτερη απόδοση από ότι το v8, όπως αναφέρεται στον πίνακα II. Αυτό μπορεί να σχετίζεται για διάφορους λόγους, όπως το v10 να συμπεριφέρεται διαφορετικά με βάση το dataset, οι 150 εποχές να μην ήταν αρκετές για το μοντέλο, ή τα δεδομένα που αναφέρονται στον πίνακα II είναι χειραγωγημένα, το οποίο είναι αρκετά συχνό στον ανταγωνισμό απόδοσης προς ταχύτητας των YOLO μοντέλων.

Μοντέλο	**Αρχιτεκτονική**	**mAP**^{val} 50	**mAP**^{val} 50-95	**Recall**^{val}
YOLOv8	'M'	0.995	0.882	0.993
YOLOv8	'X'	0.995	0.924	1
YOLOv10	'M'	0.971	0.816	0.903
YOLOv10	'X'	0.978	0.84	0.92

Πίνακας III: Αποτελέσματα με τα καλύτερα βάρη των μοντέλων YOLOv8 και YOLOv10. Τα δεδομένα περιλαμβάνουν τις επιδόσεις στο validation dataset.

Στο τελικό πείραμα, απεικονίζονται οι αποδώσεις των μοντέλων ως προς την ταχύτητα τους, Εικόνα 11. Μπορούμε να παρατηρήσουμε πως ότι το YOLOv10 όσον αφορά την ταχύτητα αποδίδει με σχετικά καλύτερο latency σε σύγκριση με τα μοντέλο YOLOv8.



Σχήμα 11: Αποτελέσματα μοντέλων ακρίβειας ως προς ταχύτητα (ms).

Αναφορές

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016. [Online]. Available: https://arxiv.org/abs/1506.02640

[2] M. Khoshdeli, R. Cong, and B. Parvin, "Detection of nuclei in h&e stained sections using convolutional neural networks," in *IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, 2017, pp. 105–108. [Online]. Available: https://doi.org/10.1109/BHI.2017.7897216

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[4] D. Mpouziotas, E. Mastrapas, N. Dimokas, P. Karvelis, and E. Glavas, "Object detection for low light images," in *2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, 2022, pp. 1–6.

[5] D. Mpouziotas, P. Karvelis, I. Tsoulos, and C. Stylios, "Automated wildlife bird detection from drone footage using computer vision techniques," *Applied Sciences*, vol. 13, no. 13, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/13/7787

[6] D. Mpouziotas, P. Karvelis, and C. Stylios, "Advanced computer vision methods for tracking wild birds from drone footage," *Drones*, vol. 8, no. 6, 2024. [Online]. Available: https://www.mdpi.com/2504-446X/8/6/259

[7] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Neural Information Processing Systems*, 31st of May, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235254713

[8] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, pp. 302 – 321, 18th of August, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:11371972

[9] B. Dwyer, J. Nelson, T. Hansen *et al.*, "Roboflow (version 1.0)," 2024, computer vision software. [Online]. Available: https://roboflow.com

[10] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[11] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," 2024. [Online]. Available: https://arxiv.org/abs/2405.14458

[12] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14113767

[13] J. H. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6469–6477, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:7211062

[14] Z. Ma, Y. Dong, Y. Xia, D. Xu, F. Xu, and F. Chen, "Wildlife real-time detection in complex forest scenes based on yolov5s deep learning network," *Remote Sensing*, vol. 16, no. 8, 2024. [Online]. Available: https://www.mdpi.com/2072-4292/16/8/1350