



Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών  
Τομέας Ηλεκτρονικής και Υπολογιστών

---

Εύρωση σε συμπίεση ανίχνευση Deepfake μέσω  
Βαθιάς Μάθησης

---

Διπλωματική Εργασία

Τερζόγλου Βασίλειος

AEM 9737

**Επιθετικός:** Πετραντωνάκης Παναγιώτης  
Επίκουρος Καθηγητής Α.Π.Θ.

11 Οκτωβρίου 2023

## **Περίληψη**

Οι επίκαιρες εξελίξεις στον χώρο της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης έχουν αναμφισβήτητα βοηθήσει στην επίλυση καθημερινών προβλημάτων. Παρ' όλα αυτά, όμως, έχουν χρησιμοποιηθεί και για κακόβουλους σκοπούς, όπως η δημιουργία συνθετικού περιεχομένου εικόνων και βίντεο που απεικονίζουν πρόσωπα. Τα Deepfakes, όπως είναι ο όρος για τέτοιου είδους περιεχόμενο, κυκλοφορούν πλέον στο Διαδίκτυο υπονομεύοντας την εμπιστοσύνη που εμπνέουν τα ψηφιακά μέσα και έχουν το δυναμικό να επηρεάσουν την κοινωνία περαιτέρω, μέσω της διάδοσης ψευδούς πληροφορίας.

Σε αυτό το πλαίσιο, η όλο εξελισσόμενη διαμάχη μεταξύ των μεθόδων παραγωγής και ανίχνευσης υποδεικνύει την ανάγκη για συστήματα που εμφανίζουν ισχυρές ικανότητες γενίκευσης. Δεδομένης της ταχείας μετάδοσης των Deepfakes μέσω των μέσων κοινωνικής δικτύωσης, όπου υπόκεινται σε συμπίεση, η υλοποίηση συστημάτων που είναι εύρωστα σε αυτές τις διαδικασίες εμφανίζουν μεγάλο ενδιαφέρον.

Σε αυτή την εργασία, δομούνται πολλαπλά μοντέλα ανίχνευσης Βαθιάς Μάθησης στη βάση ενός backbone δικτύου τύπου EfficientNet. Με σκοπό να αξιολογηθεί τόσο η ακρίβεια ανίχνευσης τους όσο και η ευρωστία σε συμπίεση, ελέγχονται υπό διαφορετικές συνθήκες, λαμβάνοντας υπ' όψιν και απαιτητικά σενάρια όπως έλεγχο σε επίπεδα συμπίεσης που δεν έχουν ιδωθεί στην φάση της εκπαίδευσης.

Βασίλειος Τερζόγλου  
vterzoglou@gmail.com

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονικής, Ελλάδα,  
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών  
Οκτώβριος, 2023

## Title

# Compression-robust Deepfake detection via Deep Learning

### Abstract

Recent advancements in the field of Artificial Intelligence and Machine Learning have undeniably enabled remarkable solutions for everyday problems. However, they have been leveraged for malicious purposes too, such as creating photo-realistic synthetic facial images and videos. Deepfakes, which is the term for such content, have been circulating the Internet eroding trust to digital media and have the potential to further influence society, through the spreading of false information.

In this context, the ever evolving arms race between production and detection methods calls for detection systems that exhibit strong generalization capabilities. Given the rapid dissemination of Deepfakes through social media, where they undergo compression, implementing systems that are robust to such processes is of great interest.

In this work, multiple Deep Learning detection models are built on the basis of an EfficientNet backbone network. In order to evaluate their detection accuracy as well as their robustness to compression, they are tested under various settings, considering challenging scenarios such as testing on compression levels that are unseen on the training phase.

Vasileios Terzoglou  
vterzoglou@gmail.com  
Aristotle University of Thessaloniki, Greece,  
Department of Electrical & Computer Engineering  
October, 2023

## **Ευχαριστίες**

Με την παρούσα εργασία ολοκληρώνονται οι προπτυχιακές μου σπουδές στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης.

Θα ήθελα να ευχαριστήσω, τον καθηγητή κ. Παναγιώτη Πετραντωνάκη για την ανάθεση της εργασίας και τον Δρ. Συμεών Παπαδόπουλο, ο οποίος ήταν ο επιστημονικός υπεύθυνος της εργασίας εκτός του πανεπιστημίου. Ευχαριστώ ιδιαίτερα τον Νικόλαο Γιάτσογλου, ερευνητή του ΙΠΤΗΛ/ΕΚΕΤΑ, για την εξαιρετική συνεργασία που είχαμε, την καθοδήγηση και την υποστήριξη που μου παρείχε για την εκπόνηση αυτής της εργασίας.

Ευχαριστώ τους φίλους και συμφοιτητές για την βοήθεια και τις μοναδικές εμπειρίες που μοιράστηκα μαζί τους κατά την διάρκεια των σπουδών μου και, τέλος, ευχαριστώ την οικογένεια μου για την συνεχή υποστήριξη τους.

# Περιεχόμενα

<b>1 Εισαγωγή</b>	<b>8</b>
1.1 Σκοπός της εργασίας . . . . .	10
<b>2 Επισκόπηση της Ερευνητικής Περιοχής και Θεωρητικό Υπόβαθρο</b>	<b>12</b>
2.1 Μέθοδοι δημιουργίας Deepfake . . . . .	12
2.1.1 Autoencoders, Αρχιτεκτονική Encoder-Decoder . . . . .	12
2.1.2 Generative Adversarial Networks (GANs) . . . . .	13
2.2 Μέθοδοι ανίχνευσης Deepfake . . . . .	14
2.2.1 Non-Machine Learning προσεγγίσεις . . . . .	14
2.2.2 Machine και Deep Learning προσεγγίσεις . . . . .	15
2.2.3 Μοντέλα που χρησιμοποιούν χρονικά (temporal) χαρακτηριστικά	17
2.3 Αντιμετώπιση του προβλήματος της ευρωσίας σε συμπίεση . . . . .	18
2.3.1 Domain Generalization - Domain Adaptation . . . . .	18
2.3.2 Μοντέλα ανίχνευσης που αντιμετωπίζουν το πρόβλημα της ευρωσίας σε συμπίεση . . . . .	19
<b>3 Υλοποιήσεις</b>	<b>23</b>
3.1 Dataset και preprocessing pipeline . . . . .	23
3.1.1 FaceForensics++ . . . . .	23
3.1.2 Preprocessing Pipeline . . . . .	24
3.2 Μοντέλα που χρησιμοποιήθηκαν . . . . .	25
3.2.1 Backbone - Feature Generator: EfficientNet . . . . .	25
3.2.2 Classifier και Baseline μοντέλο . . . . .	26
3.2.3 Adversarial και Similarity μοντέλα . . . . .	27
3.2.4 Ensemble μοντέλα . . . . .	29
<b>4 Πειράματα και Αποτελέσματα</b>	<b>30</b>
4.1 Μετρική αξιολόγησης και βάρη ανά κατηγορία . . . . .	30
4.2 Εκπαίδευση μοντέλων . . . . .	31
4.3 Πείραμα 1: Έλεγχος προσαρμογής σε συμπίεση και γενίκευσης σε άλλο σύνολο δεδομένων . . . . .	32

4.4 Πείραμα 2: Χρήση ασυμπίεστων και χαμηλού βαθμού συμπίεσης δεδομένων . . . . .	34
4.5 Πείραμα 3: Εξειδίκευση σε έναν τύπο παραποίησης . . . . .	35
4.6 Πείραμα 4: έλεγχος ακρίβειας σε unseen domains . . . . .	38
<b>5 Συμπεράσματα και Μελλοντική Εργασία</b>	<b>41</b>
5.1 Συμπεράσματα . . . . .	41
5.2 Μελλοντική εργασία . . . . .	42
<b>Βιβλιογραφία</b>	<b>44</b>
<b>Γλωσσάρι</b>	<b>52</b>
<b>Ακρωνύμια</b>	<b>54</b>

# Κατάλογος Σχημάτων

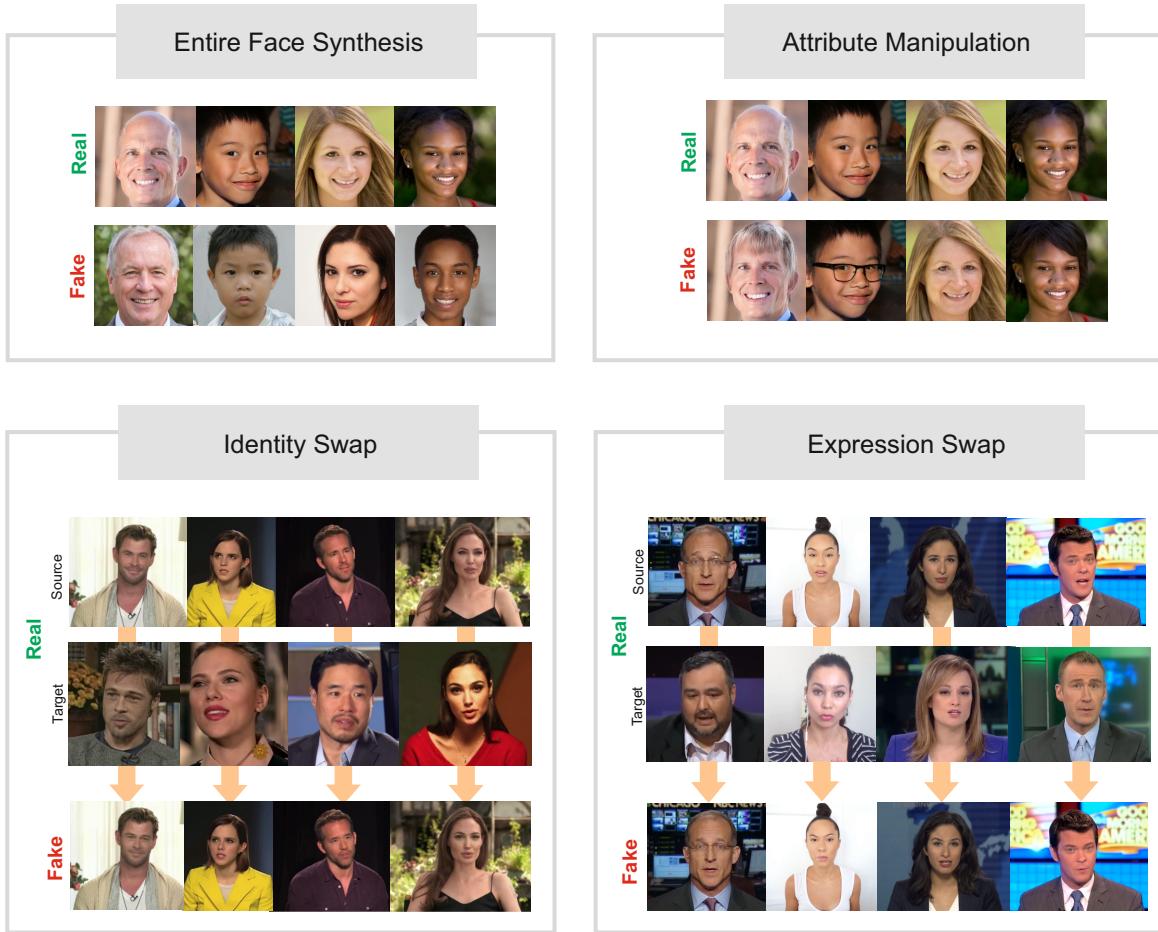
1.1 Κατηγορίες παραποίησης Deepfake . . . . .	9
2.1 Η αρχιτεκτονική Encoder-Decoder για δημιουργία Deepfake . . . . .	13
2.2 Η αρχιτεκτονική ενός GAN . . . . .	14
2.3 Παράδειγμα εικόνας με εμφανείς οπτικές ατέλειες . . . . .	15
2.4 Εικόνες με εμφανή artifacts από ανάμειξη εικόνων . . . . .	16
2.5 Σχηματική απεικόνιση του προβλήματος του Domain Generalization . .	19
2.6 Συνδυασμός πολλαπλών ασθενών ταξινομητών σε ένα μοντέλο ensemble .	20
2.7 Ένα μοντέλο για εξαγωγή Domain-Invariant χαρακτηριστικών . . . . .	21
3.1 Παράδειγμα ενός frame που έχει υποστεί συμπίεση . . . . .	24
3.2 Σύγκριση μοντέλων EfficientNet με άλλα μοντέλα . . . . .	25
3.3 Η γενική αρχιτεκτονική του μοντέλου EfficientNet-B0 . . . . .	26
3.4 Η αρχιτεκτονική του baseline μοντέλου και του ταξινομητή . . . . .	26
3.5 Η αρχιτεκτονική του Adversarial μοντέλου . . . . .	27
3.6 Η αρχιτεκτονική του μοντέλου Similarity . . . . .	28
3.7 Η αρχιτεκτονική του μοντέλου Similarity - Adversarial . . . . .	28
3.8 Η δομή του ταξινομητή επιπέδου συμπίεσης (domain) για το μοντέλο Ensemble + CNN . . . . .	29
4.1 Αποτελέσματα από το Πείραμα 1 . . . . .	32
4.2 Αποτελέσματα για το μοντέλο Baseline όταν χρησιμοποιούνται επαυξήσεις	33
4.3 Αποτελέσματα από το Πείραμα 2 . . . . .	34
4.4 Αποτελέσματα για το Πείραμα 3, κατηγορία DeepFakes . . . . .	35
4.5 Αποτελέσματα για το Πείραμα 3, κατηγορία Face2Face . . . . .	36
4.6 Αποτελέσματα για την κατηγορία DeepFakes με backbone EfficientNet-B4	37
4.7 Αποτελέσματα για την κατηγορία Face2Face με backbone EfficientNet-B4	37
4.8 Η δομή του video inferencing pipeline . . . . .	38
4.9 Αποτελέσματα για το Πείραμα 4 σε επίπεδο βίντεο . . . . .	39
4.10 Αποτελέσματα για το Πείραμα 4 σε επίπεδο frame . . . . .	39

# 1

## Εισαγωγή

Η ψηφιοποίηση της καθημερινότητας και η επακόλουθη ανάδειξη και ανάπτυξη της τεχνητής νοημοσύνης έχει μετασχηματίσει πολλές πτυχές της ζωής, δημιουργώντας παράλληλα νέες λειτουργίες και εφαρμογές. Η σχετικά πρόσφατη επινόηση των Παραγωγικών Αντιπαραθετικών Δικτύων (Generative Adversarial Networks - GANs) [1] και η χρήση των Autoencoders (AEs) έχουν συμβάλλει στην δημιουργία συνθετικού πολυμεσικού περιεχομένου εντυπωσιακής αληθοφάνειας. Η κυριότερη κατηγορία τέτοιου περιεχομένου είναι τα **Deepfakes**: ψευδείς εικόνες ή βίντεο που απεικονίζουν πρόσωπα τα οποία έχουν παραχθεί ή τροποποιηθεί με μεθόδους Βαθιάς Μάθησης (Deep Learning - DL). Οι κυριότεροι τύποι παραποίησης (manipulation) περιλαμβάνουν:

- την τροποποίηση των χαρακτηριστικών των προσώπων (Facial attribute editing), όπως για παράδειγμα η προσθήκη ή αφαίρεση γυαλιών ή η αλλαγή της ηλικίας ενός προσώπου,
- την μίμηση-μεταφορά των εκφράσεων (Facial reenactment) από ένα προσώπου-πηγή (source) σε ένα πρόσωπο-στόχο (target),
- την μεταφορά της ταυτότητας (Facial identity swap) του προσώπου-πηγή στο πρόσωπο-στόχο
- και την δημιουργία πλήρως συνθετικών προσώπων (Fully synthetic face generation).



Σχήμα 1.1: Κατηγορίες παραποίησης Deepfake [2]

Οι πρώτες εμφανίσεις αυτών των παραποιήσεων χρονολογούνται περίπου στο 2017, όπου χρησιμοποιήθηκαν για την δημιουργία ψευδούς πορνογραφικού περιεχομένου και έκτοτε το μεγαλύτερο μέρος τους χρησιμοποιείται κακόβουλα. Ιδιαίτερος κίνδυνος εμφανίζεται όταν αφορούν διάσημους και πολιτικούς, αφού η αλλοίωση πληροφοριών σχετικά με αυτά τα πρόσωπα μπορεί να έχει σημαντική επίδραση σε μεγάλη μάζα πληθυσμού. Τα συγκεκριμένα άτομα είναι μάλιστα πιο πιθανό να υποστούν παραποίηση του προσώπου τους με χρήση Deepfake, δεδομένου του όγκου πληροφορίας που υπάρχει για αυτά [3]. Υπάρχουν βέβαια και άλλες περιπτώσεις όπου Deepfakes στοχοποιούν πιο καθημερινούς ανθρώπους, όταν χρησιμοποιούνται ως μέσα για εκβιασμό και revenge porn ή για παραποίηση αποδεικτικών στοιχείων σε δικαστικές υποθέσεις [4, 5].

Είναι αλήθεια ότι υπάρχει και ένα μέρος των Deepfakes που δημιουργείται και χρησιμοποιείται καλοπροσαίρετα, τόσο για ψυχαγωγικούς, όσο και για ενημερωτικούς και εκπαιδευτικούς σκοπούς. Υπάρχουν εφαρμογές δημιουργίας Deepfake τις οποίες δημοσιογραφικοί σταθμοί έχουν χρησιμοποιήσει για να παρουσιάσουν ειδησεογραφικά γεγονότα<sup>1</sup>, ενώ έχουν χρησιμοποιηθεί και για την εμψυχοποίηση (animation) έργων

<sup>1</sup><https://www.reuters.com/article/rpb-synthesia-prototype-idUSKBN201103>

τέχνης.

Δεδομένης της αύξησης των ψηφιακής πληροφορίας, της ύπαρξης πληθώρας ελεύθερων σε πρόσθαση βάσεων δεδομένων προσώπων και της ανάπτυξης φιλικών προς χρήση εφαρμογών για δημιουργία παραπομήσεων<sup>2,3,4,5</sup>, τα Deepfakes αναμένεται να αποτελέσουν ένα νέο μέσο παραπληροφόρησης. Δυσφήμιση, στρέβλωση της κοινής γνώμης, αλλοίωση πολιτικών δηλώσεων και εκλογικών αποτελεσμάτων είναι μόνο μερικά παραδείγματα στα οποία είναι πιθανό να παίξουν σημαντικό ρόλο [6]. Αν στα προηγούμενα προστεθεί και η συνεχής βελτίωση της οπτικής ποιότητας των παραπομήσεων, και επομένως η αύξηση της πιθανότητας εξαπάτησης, γίνεται προφανής η ισχύς αυτού του μέσου. Σε ένα περιθάλλον όπου δεν είναι εύκολο κανείς να διακρίνει το τι είναι αλήθεια και τι ψέμα, είναι εύκολο να καλλιεργηθεί η δυσπιστία προς τα μέσα ενημέρωσης και να εδραιωθούν κινήματα προπαγάνδας [7].

### 1.1 ΣΚΟΠΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ

---

Για τους παραπάνω λόγους, έχει δοθεί μεγάλη έμφαση στο πρόβλημα της ανίχνευσης των Deepfakes, που παραδοσιακά εντάσσεται στην γενικότερη κατηγορία της ανίχνευσης χαλκεύσεων (Forgery Detection) [8] και εγκληματολογίας μέσων ενημέρωσης (Media forensics) [9].

Ετσι, οι πρώιμες τεχνικές βασίστηκαν στην ανίχνευση "αποτυπωμάτων" και ιχνών τα οποία δημιουργήθηκαν κατά την παραπομή των πολυμέσων, με τη δημιουργία αυτοσχέδιων χαρακτηριστικών (handcrafted features) που τα αναδεικνύουν. Ωστόσο, όσο οι μέθοδοι δημιουργίας Deepfake αναπτύσσονται, οι παραδοσιακές μέθοδοι ανίχνευσης δείχνουν να υστερούν και το ενδιαφέρον πλέον στρέφεται σε αυτοματοποιημένα συστήματα ανίχνευσης με μεθόδους Deep Learning. Τα συστήματα αυτά εξάγουν αυτοματοποιημένα χωρικά χαρακτηριστικά από εικόνες, ή/και χρονικά και ηχητικά χαρακτηριστικά στην περίπτωση των βίντεο, που παρ' ότι δεν είναι άμεσα ερμηνεύσιμα, επιτρέπουν ανίχνευση των Deepfake με υψηλή ακρίβεια.

Εκτός του ζητήματος αυτού - της έλλειψης δυνατότητας εξήγησης των μοντέλων μηχανικής μάθησης και των χαρακτηριστικών που χρησιμοποιούν (explainability issue)- κοινό πρόβλημα που συνήθως αντιμετωπίζεται είναι αυτό της γενίκευσης, δηλαδή της καλής ακρίβειας σε δεδομένα τα οποία δεν είναι διαθέσιμα στη διαδικασία της εκπαίδευσης. Σε αυτή την κατεύθυνση, ένα ειδικότερο πρόβλημα που έχει αναγνωριστεί από τη βιβλιογραφία και σχετίζεται με το ζήτημα της γενίκευσης είναι η συμπίεση. Συγκεκριμένα, ένα σημαντικό ποσοστό των Deepfakes μεταδίδονται μέσω των μέσων κοινωνικής δικτύωσης, όπου κατά κανόνα συμπιέζονται για εξοικονόμηση χώρου και

<sup>2</sup><https://github.com/iperov/DeepFaceLab>

<sup>3</sup><https://deepfakesweb.com/>

<sup>4</sup><https://reface.ai/>

<sup>5</sup><https://www.faceapp.com/>

ταχύτερη μετάδοση. Αναπόφευκτα, η διαδικασία αυτή αλλοιώνει τα χαρακτηριστικά των πολυμέσων που βρίσκονται σε χαμηλό επίπεδο, πχ σε επίπεδο pixel, ώστε τα αυτοματοποιημένα συστήματα ανίχνευσης που εκπαιδεύονται και ελέγχονται σε δεδομένα με διαφορετικό βαθμό συμπίεσης να εμφανίζουν χαμηλή ακρίβεια [10, 11]. Το πρόβλημα αυτό εντάσσεται στην κατηγορία του Domain Generalization, η οποία παρουσιάζεται στη συνέχεια.

Σκοπός της παρούσας εργασίας είναι να δημιουργηθούν και να συγκριθούν διαφορετικές υλοποιήσεις, αναζητώντας ένα σύστημα ανίχνευσης Deepfake το οποίο θα έχει δυνατότητα γενίκευσης και θα είναι εύρωστο (robust) ως προς την συμπίεση.

# 2

## Επισκόπηση της Ερευνητικής Περιοχής και Θεωρητικό Υπόβαθρο

### 2.1 ΜΕΘΟΔΟΙ ΔΗΜΙΟΥΡΓΙΑΣ DEEPFAKE

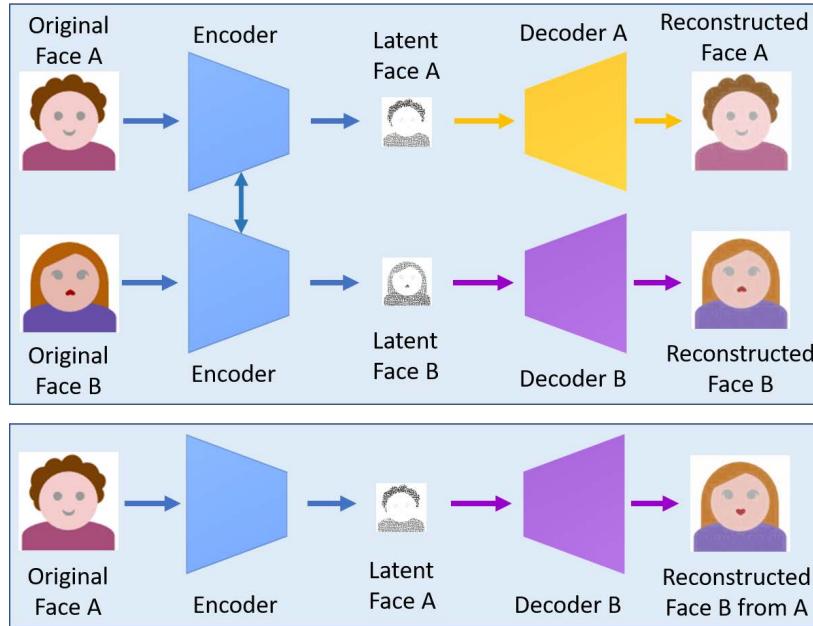
---

#### 2.1.1 Autoencoders, Αρχιτεκτονική Encoder-Decoder

Οι Autoencoders είναι ένα είδος Νευρωνικού Δικτύου (Neural Network - NN) που χρησιμοποιείται για σκοπούς μη επιθλεπόμενης (Unsupervised) εκμάθησης αναπαράστασης (Representation Learning). Αποτελούν μια ειδική περίπτωση των δικτύων Encoder-Decoder και βάση για τα περισσότερα συστήματα δημιουργίας Deepfake.

Στόχος του Autoencoder είναι ανακατασκευάσει την είσοδο που δέχεται στην έξοδο του. Ειδικότερα, μαθαίνει δύο συναρτήσεις - μία για τον κωδικοποιητή (Encoder)  $E$  και μία για τον απόκωδικοποιητή (Decoder)  $D$  - δεδομένης μιας εισόδου  $x$  (από κάποια κατανομή  $X$ ) ώστε η έξοδος  $D(E(x)) = x_g$  να ελαχιστοποιεί κάποια συνάρτηση απώλειας (Loss function)  $L(x, x_g)$  με βάση την ομοιότητα των  $x$  και  $x_g$ . Τα δύο τμήματα που απαρτίζουν τον Autoencoder και συνήθως είναι συμμετρικά συνδέονται σε ένα μοντέλο κλεψύδρας (stacked hourglass), όπου τα ενδιάμεσα στρώματα είναι πιο στενά ώστε να ενθαρρύνεται το μοντέλο να συνοψίσει τα κύρια χαρακτηριστικά της εισόδου, χωρίς να μάθει την ταυτοτική συνάρτηση, ενώ ενδέχεται να επιβάλλονται και επιπλέον περιορισμοί σε αυτή την κατεύθυνση [12, 13].

Για την δημιουργία Deepfakes της κατηγορίας Facial identity swap χρησιμοποιο-



Σχήμα 2.1: Η βασική αρχιτεκτονική Encoder-Decoder για δημιουργία Deepfake [14]

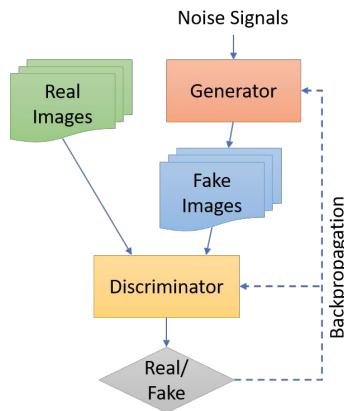
ύνται δύο Autoencoders, οι οποίοι εκπαιδεύονται αντίστοιχα σε εικόνες από δύο διαφορετικά πρόσωπα  $A$  και  $B$ . Για να είναι συμβατοί μεταξύ τους οι Autoencoders, το τμήμα του Encoder είναι κοινό (χρησιμοποιεί τα ίδια βάρη). Αυτό επιτρέπει την εκμάθηση των κοινών χαρακτηριστικών των προσώπων από τον Encoder και των ιδιαίτερων χαρακτηριστικών της ταυτότητας από τους Decoders. Έτσι, αφού έχει ολοκληρωθεί η εκπαίδευση, μπορεί να δημιουργηθεί το *Faceswap* με πηγή το πρόσωπο  $A$  και στόχο το  $B$  χρησιμοποιώντας τον Decoder του προσώπου  $B$  και ως είσοδο κάποια εικόνα του  $A$ , όπως παρουσιάζεται στο Σχήμα 2.1.

Τα αποτελέσματα της μεθόδου σε οπτικό επίπεδο πουκίλλουν αναλόγως της υλοποίησης και της ομοιότητας των δεδομένων για τα δύο πρόσωπα. Πρόσωπα που εμφανίζονται σε μερική απόκρυψη (partial facial occlusion), σε διαφορετικές πόζες ή υπό διαφορετικές συνθήκες φωτισμού συνήθως οδηγούν σε χαμηλότερης οπτικής ποιότητας αποτελέσματα. Παρ' όλα αυτά, εφαρμόζεται σε αρκετά συστήματα και νεότερες υλοποιήσεις [15, 16] επιλαμβάνονται αυτών των ζητημάτων.

### 2.1.2 Generative Adversarial Networks (GANs)

Τα Παραγωγικά Αντιπαραθετικά Δίκτυα (GANs) αποτελούνται από δύο Νευρωνικά Δίκτυα, τον Generator  $G$  και τον Discriminator  $D$  που λειτουργούν αντιπαραθετικά μεταξύ τους σε ένα παίγνιο μηδενικού αθροίσματος (zero-sum game).

Δεδομένου ενός συνόλου δεδομένων από εικόνες  $x$  με κατανομή  $p_{data}(x)$ , ο  $G$  λαμβάνει ως είσοδο κάποιο δείγμα  $z$  από θόρυβο με κατανομή  $p_z(z)$  και δημιουργεί εικόνες  $G(z)$  παρόμοιες με τις πραγματικές (αυτές που προέρχονται από το σύνολο δεδομένων).



Σχήμα 2.2: Η βασική αρχιτεκτονική GAN. Το δίκτυο μπορεί να εκπαιδευτεί με την μέθοδο της οπισθοδιάδοσης [14].

Σκοπός του  $D$  είναι να μεγιστοποιήσει την πιθανότητα να αποδώσει την σωστή ετικέτα τόσο στις πραγματικές όσο και στις συνθετικές εικόνες, ενώ του  $G$  να παράγει εικόνες που θα μεγιστοποιήσουν την πιθανότητα να κάνει λάθος ο  $D$ . Μαθηματικά το παίγνιο μοντελοποιείται από την παρακάτω συνάρτηση τιμής:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (2.1)$$

όπου ως  $D(x)$  συμβολίζεται η πιθανότητα η εικόνα  $x$  να είναι πραγματική εικόνα παρά να προέρχεται από τον  $G$  [1].

Τα δύο τμήματα του δικτύου εκπαιδεύονται παράλληλα με την μέθοδο της οπισθοδιάδοσης (backpropagation) μέχρι μία κατάσταση ισορροπίας όπου οι παραγόμενες εικόνες δεν μπορούν να διακριθούν από τις πραγματικές. Στη συνέχεια ο Generator μπορεί να χρησιμοποιηθεί για την δημιουργία φωτορεαλιστικών εικόνων. Μερικά παραδείγματα υλοποίησεων Generative Adversarial Network (GAN) τα οποία χρησιμοποιούνται και για παραγωγή Deepfake είναι τα StyleGAN [17-19], PG-GAN [20], StarGAN [21], InterFaceGAN [22].

## 2.2 ΜΕΘΟΔΟΙ ΑΝΙΧΝΕΥΣΗΣ DEEPFAKE

---

### 2.2.1 Non-Machine Learning προσεγγίσεις

Οι πρώτες προτάσεις της ερευνητικής κοινότητας για ανίχνευση Deepfake προέρχονται από την περιοχή των Media Forensics και αφορούν την αναζήτηση αυθεντικών μέσων χρησιμοποιώντας συγκεκριμένα οπτικά ευρήματα (visual artifacts). Τα στοιχεία αυτά δημιουργούνται από την καταγραφή και την επεξεργασία των μέσων με μηχανισμούς τόσο in-camera - αισθητήρες, φακοί, φίλτρα, color filter arrays (CFAs) - όσο και out-camera όταν πρόκειται για post-processing επεξεργασία ή σύνθεση με χρήση



Σχήμα 2.3: Αριστερά: Αληθινή εικόνα, δεξιά: παραποιημένη εικόνα με εμφανείς ατέλειες [24].

GAN. Ενδεικτικό παράδειγμα είναι το [23] όπου ως χαρακτηριστικό για την ταξινόμηση σε αληθινά και παραποιημένα μέσα χρησιμοποιείται ένα είδος θορύβου που οφείλεται στην ανομοιομορφία κατασκευής των αισθητήρων που χρησιμοποιούνται σε κάμερες, το Photo Response Non-Uniformity (PRNU). Οι υλοποιήσεις που εντάσσονται σε αυτή την περιοχή είναι περιορισμένες, αφού είναι σαφές πως τα αποτελέσματα που προκύπτουν δεν είναι γενικεύσιμα και το ενδιαφέρον στρέφεται σε μοντέλα Μηχανικής Μάθησης (Machine Learning - ML).

### 2.2.2 Machine και Deep Learning προσεγγίσεις

Οι αρχικές ML μέθοδοι ανίχνευσης συνήθως βασίζονται σε εμφανείς ατέλειες των Deepfakes και στοχεύουν να κατασκευάσουν και να χρησιμοποιήσουν χαρακτηριστικά που θα τις αναδεικνύουν, ώστε να διακρίνουν αληθινές και παραποιημένες εικόνες.

Υπό αυτή τη λογική, στο [24] παρουσιάζεται ένα σύστημα ανίχνευσης που βασίζεται σε σχετικά απλά αυτοσχέδια χαρακτηριστικά που εκμεταλλεύονται artifacts, όπως αυτά στο Σχήμα 2.3. Οι ταξινομητές που χρησιμοποιούνται είναι δύο μοντέλα λογιστικής παλινδρόμησης (Logistic Regression) και Multi-Layer Perceptron (MLP). Το καλύτερο μοντέλο εμφανίζει 0.851 Area Under Curve (AUC), χρησιμοποιώντας όμως σύνολο δεδομένων (dataset) το οποίο δεν έχει δημοσιευθεί.

Στο [25] γίνεται η υπόθεση ότι οι αλγόριθμοι Deepfake μπορούν να δημιουργήσουν εικόνες περιορισμένης ανάλυσης, οι οποίες χρειάζεται να παραμορφωθούν για να ταιριάζουν με τις πραγματικές εικόνες στα πηγαία βίντεο. Τέτοιοι μετασχηματισμοί δημιουργούν artifacts που βοηθούν στην ανίχνευση. Επομένως προτείνεται ένα σύστημα ανίχνευσης που βασίζεται σε Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs). Τέσσερα μοντέλα εκπαιδεύτηκαν: VGG16 [26], ResNet50, ResNet101 και ResNet152 [27]. Η προσέγγιση αυτή ελέγχθηκε χρησιμοποιώντας τις βάσεις δεδομένων UADFV [28] και DeepfakeTIMIT [29] ξεπερνώντας τις σύγχρονες επιδόσεις σε αυτές.

Το μοντέλο Face X-Ray που παρουσιάζεται στο [30] στοχεύει να αναγνωρίσει παραποιημένες εικόνες που έχουν παραχθεί με την ανάμειξη (blending) εικόνων - μια



Σχήμα 2.4: Εικόνες με εμφανή artifacts από ανάμειξη (blending) εικόνων [31]

διαδικασία που επίσης μπορεί να δημιουργήσει artifacts, όπως αυτά στο Σχήμα 2.4. Συγκεκριμένα στόχος είναι να εντοπιστούν τα όρια (boundaries) εντός των οποίων πραγματοποιείται η ανάμειξη, εκπαιδεύοντας ένα CNN. Χρησιμοποιούνται δεδομένα από το FaceForensics++ (FF++) [11], ελέγχονται σενάρια cross-manipulation, όπου το μοντέλο εκπαιδεύεται σε έναν τύπο παραποίησης και ελέγχεται σε κάποιον διαφορετικό, εμφανίζοντας καλές δυνατότητες γενίκευσης.

Μια προσέγγιση Deep Learning παρουσιάζεται στο [32], όπου όμως ο αριθμός των στρωμάτων και των παραμέτρων που χρησιμοποιούνται είναι μικρός, ώστε να αναδεικνύονται τα μεσαίου επιπέδου χαρακτηριστικά των εικόνων. Τα δύο μοντέλα που προτείνονται - το MesoNet και το MesoInception-4 (μια παραλλαγή του πρώτου βασισμένη στα δίκτυα Inception [33]) - αποτελούνται από συνελικτικά στρώματα. Τα μοντέλα παρουσιάζουν υψηλή ακρίβεια, που όμως μειώνεται σημαντικά όταν πρόκειται για καρέ (frame) από βίντεο που έχουν υποστεί συμπίεση.

Στο [11] εκτός της βάσης δεδομένων FF++ και ενός benchmark<sup>1</sup> που δημοσιεύονται, συγκρίνονται και οι επιδόσεις πολλών μοντέλων ανίχνευσης. Συγκεκριμένα, εξετάζονται πέντε μοντέλα: (i) ένα μοντέλο CNN εκπαιδευμένο μέσω αυτοσχέδιων χαρακτηριστικών [34], (ii) ένα μοντέλο CNN με στρώματα ειδικά εκπαιδευμένα να καταπιέζουν το περιεχόμενο υψηλού επιπέδου της εικόνας, ώστε να χρησιμοποιούνται χαρακτηριστικά χαμηλότερου επιπέδου [35], (iii) ένα μοντέλο CNN με ένα στρώμα global pooling που υπολογίζει τέσσερις εκτιμήσεις για στατιστικά (μέση τιμή, διακύμανση, μέγιστο, ελάχιστο), (iv) το MesoInception-4 [32] και (v) το XceptionNet [36], ένα προεκπαιδευμένο στο ImageNet [37] CNN το οποίο επανεκπαιδεύεται. Τα μοντέλα αποδίδουν με ακρίβεια πάνω από 95% όσον αφορά εικόνες που δεν έχουν υποστεί συμπίεση (raw), όμως η ακρίβεια τους μειώνεται σημαντικά όταν εμφανίζεται υψηλή συμπίεση. Σε αυτό το σενάριο φαίνεται η σχετική υπεροχή των μεθόδων Deep Learning έναντι των μεθόδων Machine Learning που χρησιμοποιούν αυτοσχέδια χαρακτηριστικά, η οποία επαληθε-

---

<sup>1</sup>[https://kaldir.vc.in.tum.de/faceforensics\\_benchmark/](https://kaldir.vc.in.tum.de/faceforensics_benchmark/)

ύεται και σε άλλες συγκριτικές αναλύσεις [38, 39] - μια βασική διαπίστωση στην οποία στηρίζεται και η παρούσα εργασία.

### 2.2.3 Μοντέλα που χρησιμοποιούν χρονικά (temporal) χαρακτηριστικά

Οι προσεγγίσεις που αναλύθηκαν προηγουμένως αφορούν αποκλειστικά μοντέλα που χρησιμοποιούν χωρικά (spatial) χαρακτηριστικά που εξάγονται από εικόνες. Όταν το μέσο που εξετάζεται είναι βίντεο, αυτές λειτουργούν με είσοδο δείγματά του (frames) και μπορούν να δώσουν αποτελέσματα είτε ανά δείγμα, είτε συνολικά σε επίπεδο βίντεο.

Μια άλλη κατηγορία μεθόδων που μπορεί να δώσει μία επιπλέον διάσταση στο πρόβλημα της ανίχνευσης Deepfake είναι αυτές που αξιοποιούν χρονικά (temporal) χαρακτηριστικά. Χρησιμοποιώντας τα frames όχι ως ανεξάρτητα δείγματα, αλλά ως ακολουθία δειγμάτων και εξάγοντας χαρακτηριστικά μέσω 3D-CNN και Επαναληπτικών Νευρωνικών Δικτύων (RNN) - συνήθως με τη μορφή δικτύων Long Short-Term Memory (LSTM) - έχουν την δυνατότητα να αναλύσουν διαφορετικές πτυχές όπως η χρονική συνοχή (temporal coherence). Γενικά εμφανίζουν καλύτερη ακρίβεια συγκριτικά με τις προσεγγίσεις που χρησιμοποιούν αποκλειστικά χωρικά χαρακτηριστικά, σε βάρος, όμως, του απαιτούμενου υπολογιστικού φόρτου και χρόνου για εκπαίδευση.

Οι υλοποιήσεις που δημιουργούν Deepfakes χρησιμοποιώντας Autoencoders, λειτουργούν καρέ-καρέ (frame-by-frame) αγνοώντας πλήρως τις προηγούμενες εικόνες προσώπου που μπορεί να έχουν δημιουργηθεί. Η έλλειψη αυτής της χρονικής διάστασης μπορεί να δημιουργήσει διάφορες ασυνέπειες, όπως για παράδειγμα διαφορετικό φωτισμό των προσώπων μεταξύ των καρέ. Στο [40] προτείνεται ένα μοντέλο που συνδυάζει ένα CNN με ένα LSTM δίκτυο, με σκοπό να αποτυπωθούν αυτές ακριβώς οι αστοχίες και να διακριθούν πραγματικά και παραποτημένα βίντεο. Η ακρίβεια που πετυχαίνει το μοντέλο τους είναι άνω του 95%, χρησιμοποιώντας όμως ένα ιδιωτικό dataset.

Μια διαφορετική, σχετική, όμως, διαπίστωση είναι πως λόγω αυτής της χρονικής άγνοιας των μεθόδων δημιουργίας Deepfake, αυτά αποτυγχάνουν να αναπαράγουν βιομετρικά χαρακτηριστικά αληθινών προσώπων σε βίντεο. Το [41] βασίζεται στον βλεφαρισμό (blinking), χρησιμοποιώντας ένα μοντέλο Long Recurrent Convolutional Network (LRCN) [42] για εξαγωγή χαρακτηριστικών από ακολουθίες εικόνων οι οποίες περικόπτονται στην περιοχή των ματιών. Στο [28] προτείνεται η εκτίμηση της τρισδιάστατης πόζας των προσώπων μέσω εξαγωγής σημείων αναφοράς (landmarks) και η κατηγοριοποίηση με βάση την ασυνέπεια των διαδοχικών καρέ μέσω ενός ταξινομητή Support Vector Machine (SVM). Και οι δύο μέθοδοι πετυχαίνουν υψηλά επίπεδα AUC στα dataset που ελέγχονται - 0.99 και 0.974 αντίστοιχα, επόμενες έρευνες, όμως, δείχνουν ότι τα αποτελέσματα των μεθόδων δεν είναι γενικεύσιμα σε άλλα σύνολα δεδομένων [2, 43, 44].

Στα [45, 46] προτείνεται η ανίχνευση ασυνήθιστων οπτικών ευρημάτων λόγω κίνησης (motion artifacts) σε βίντεο Deepfakes τα οποία αποτυπώνονται μέσω της οπτικής ροής (optical flow, κίνηση ενός αντικειμένου σε διαδοχικά καρέ). Χρησιμοποιείται το dataset του FaceForensics++ [11], εξάγονται χαρακτηριστικά οπτικής ροής μέσω του PWC-Net [47] και τροφοδοτούνται σε ένα δίκτυο ResNet50 [27]. Εξετάζονται σενάρια εκπαίδευσης του μοντέλου και για τις τρεις κατηγορίες συμπίεσης του συνόλου δεδομένων ξεχωριστά, αλλά και σενάρια cross-forgery, όπου το μοντέλο εκπαιδεύεται σε μία μέθοδο παραποίησης και ελέγχεται σε διαφορετική. Ο συνδυασμός αυτής της προσέγγισης με ένα μοντέλο που χρησιμοποιεί χωρικά χαρακτηριστικά, επίσης δοκιμάζεται και οδηγεί σε καλύτερη ακρίβεια σε cross-forgery σενάρια, ενώ η πτώση της ακρίβειας με βάση την ποιότητα των βίντεο είναι εμφανής (95.75% και 80.56% για τις HQ και LQ ποιότητες του dataset αντίστοιχα).

## 2.3 ΑΝΤΙΜΕΤΩΠΙΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ ΤΗΣ ΕΥΡΩΣΤΙΑΣ ΣΕ ΣΥΜΠΙΕΣΗ

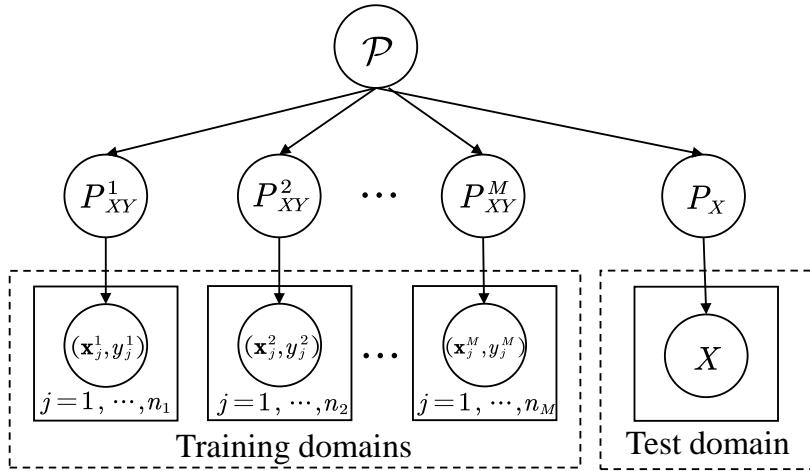
---

Παρά τις πολυάριθμες προσεγγίσεις για ανίχνευση Deepfakes, ένα σενάριο που εμφανίζει ενδιαφέρον και εξακολουθεί να είναι απαιτητικό είναι αυτό στο οποίο τα εξεταζόμενα βίντεο έχουν υποστεί συμπίεση, δεδομένης της απώλειας ενός μέρους πληροφορίας. Αναγνωρίζονται από την βιβλιογραφία [2, 9, 43, 44] τόσο το γεγονός ότι τα μοντέλα ανίχνευσης, γενικά, εμφανίζουν χαμηλότερη ακρίβεια σε αυτά τα σενάρια, αλλά και η ανάγκη για μοντέλα που θα αποδίδουν καλά σε συνθήκες ελέγχου out-of-distribution, δηλαδή σε βίντεο σε συνθήκες συμπίεσης που δεν έχουν συμπεριληφθεί στα σύνολα εκπαίδευσης των μοντέλων αναγνώρισης. Τα δύο αυτά προβλήματα που σχετίζονται με την ευρωστία στην συμπίεση, εντάσσονται στην κατηγορία προβλημάτων Domain Adaptation και Domain Generalization αντίστοιχα [48, 49].

### 2.3.1 Domain Generalization - Domain Adaptation

Ένα πεδίο (domain) αποτελείται από δεδομένα που δειγματοληπτούνται από κάποια κατανομή:  $D = \{(x_i, y_i)\}_{i=1}^n \sim P_{XY}$ , όπου  $x \in \mathcal{X} \subset \mathbb{R}^3$  είναι μία εικόνα εισόδου,  $y \in \mathcal{Y} \subset \mathbb{R}$  είναι η έξοδος - ετικέτα ταξινόμησης (classification label), με χώρους εισόδου και έξόδου  $\mathcal{X}$  και  $\mathcal{Y}$ , τυχαίες μεταβλητές  $X$  και  $Y$  αντίστοιχα και από κοινού κατανομή  $P_{XY}$ .

Όπως παρουσιάζεται και στο Σχήμα 2.5, στο πρόβλημα του Domain Generalization θεωρούνται δεδομένα  $M$  train domains  $D_{train} = \{D^i \mid i = 1, \dots, M\}$  με διαφορετικές από κοινού κατανομές  $P_{XY}^i \neq P_{XY}^j$ ,  $1 \leq i \neq j \leq M$ . Ζητούμενο είναι να εκπαιδευτεί ένα μοντέλο πρόβλεψης  $f : \mathcal{X} \rightarrow \mathcal{Y}$  χρησιμοποιώντας τα  $M$  train domains ώστε να ελαχιστοποιείται το σφάλμα πρόβλεψης σε ένα test domain  $D_{test}$  (το οποίο δεν είναι



Σχήμα 2.5: Σχηματική απεικόνιση του προβλήματος του Domain Generalization [50]

διαθέσιμο κατά την εκπαίδευση και για το οποίο ισχύει  $P_{XY}^{test} \neq P_{XY}^i \forall i \in \{1, \dots, M\}$ ):

$$\min_f \mathbb{E}_{(\mathbf{x}, y) \in D_{test}} [\ell(f(\mathbf{x}), y)] \quad (2.2)$$

όπου  $\ell$  είναι η συνάρτηση απώλειας (loss function). Η διαφορά με το πρόβλημα του Domain Adaptation έγκειται στο ότι σε αυτό είναι διαθέσιμα δεδομένα - ενδεχομένως χωρίς labels - από το test domain.

### 2.3.2 Μοντέλα ανίχνευσης που αντιμετωπίζουν το πρόβλημα της ευρωστίας σε συμπίεση

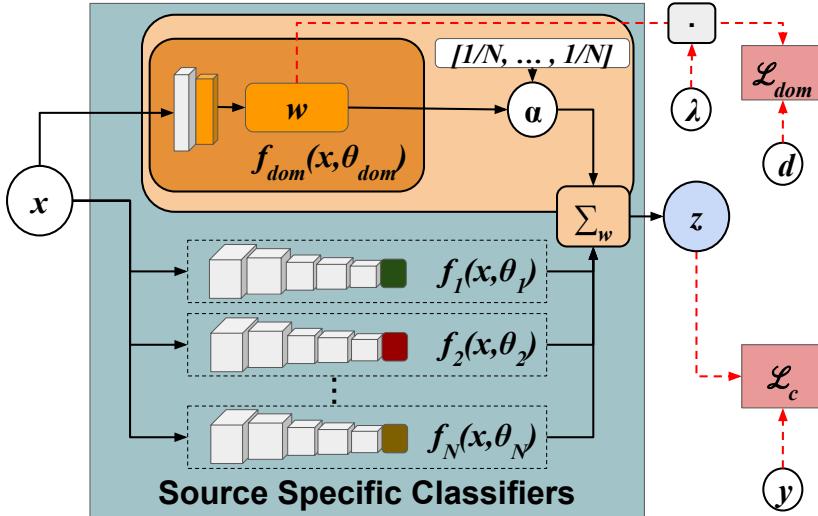
#### Fusing domain-specific classifiers

Μια κατηγορία μεθόδων που δίνουν λύση στα προβλήματα Domain Adaptation - Domain Generalization, χωρίς όμως να έχει βρει εφαρμογή, εξ όσων είναι γνωστά στη βιβλιογραφία, στο πρόβλημα της ευρωστίας σε συμπίεση, είναι αυτή του Ensemble Learning [51, 52]. Αυτή η κατηγορία μεθόδων βασίζεται στην χρήση πολλαπλών ασθενών μοντέλων - ταξινομητών για κάθε διαθέσιμο domain και τον κατάλληλο συνδυασμό τους για την δημιουργία ενός συνολικού μοντέλου.

Στο [53] παρουσιάζεται μια σχετική προσέγγιση, όπου προτείνεται η εκπαίδευση ενός ταξινομητή  $f_j$ , με παραμέτρους  $\theta_j$  για κάθε domain  $j = \{1, \dots, N\}$  και ο γραμμικός συνδυασμός των επί μέρους εξόδων ως

$$z_i = f(\mathbf{x}_i; \Theta) = (1 - a) \sum_{j=1}^N w_{ij} f_j(\mathbf{x}_i; \theta_j) + \frac{a}{N} \sum_{j=1}^N f_j(\mathbf{x}_i; \theta_j) \quad (2.3)$$

για κάθε εικόνα εισόδου  $\mathbf{x}_i$ ,  $i = \{1, \dots, M\}$ . Το διάνυσμα των βαρών  $w_i$  προκύπτει



Σχήμα 2.6: Συνδυασμός πολλαπλών ασθενών ταξινομητών (weak classifiers) σε ένα μοντέλο ensemble [53]

από ένα δίκτυο πρόβλεψης domain  $f_{dom}$  με παραμέτρους  $\theta_{dom}$  και το συνολικό δίκτυο εκπαιδεύεται με βάση την σύνθετη loss function

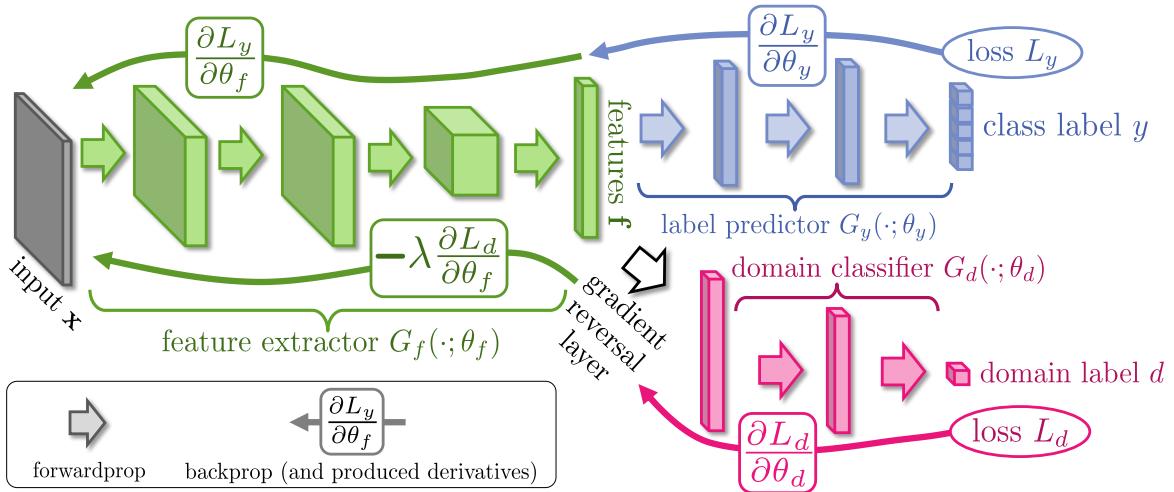
$$\mathcal{L} = \frac{1}{M} \sum_{i=1}^M (\mathcal{L}_c(z_i, y_i) + \lambda \mathcal{L}_{dom}(w_i, d_i)) \quad (2.4)$$

όπου  $M$  το πλήθος των εικόνων,  $\mathcal{L}_c$ ,  $\mathcal{L}_{dom}$  οι επιμέρους loss functions για ταξινόμηση σε αληθινές και παραποιημένες εικόνες (semantic classification) και πρόβλεψη domain αντίστοιχα,  $y_i$ ,  $d_i$  οι σχετικές ground-truth ετικέτες. Η υπερπαράμετρος  $\alpha$  χρησιμοποιείται για να προστεθεί ένα domain-agnostic στοιχείο, σταθμίζοντας την εκτίμηση των βαρών και μια ομοιόμορφη κατανομή τους, ενώ η υπερπαράμετρος  $\lambda$  σταθμίζει το βάρος της ταξινόμησης domain. Στην παρούσα εργασία εξετάζεται μία υλοποίηση που εφαρμόζει αυτή την προσέγγιση στο πρόβλημα της ευρωστίας στην συμπίεση για ανίχνευση Deepfakes.

## Learning Domain-Invariant features

Μία διαφορετική κατηγορία μεθόδων που χρησιμοποιούνται για τον ίδιο σκοπό είναι αυτές που προσπαθούν να δημιουργήσουν χαρακτηριστικά τα οποία θα είναι αναλλοίωτα ως προς το domain. Στο [54] παρουσιάζεται μια προσέγγιση για την επίλυση του προβλήματος του Domain Adaptation όταν μερικά από τα διαθέσιμα για εκπαίδευση δεδομένα έχουν ετικέτες (partially labeled).

Έχοντας πρόσβαση σε εικόνες από ένα domain πηγή (source)  $S$  και τις σχετικές ετικέτες (semantic labels) και εικόνες από ένα domain στόχο (target)  $T$  για τις οποίες δεν είναι διαθέσιμες ετικέτες, το μοντέλο που παρουσιάζεται αποτελείται από τρία κύρια



Σχήμα 2.7: Η αρχιτεκτονική που προτείνεται στο [54]

εξαρτήματα (components):

- μία γεννήτρια χαρακτηριστικών  $G_f(x; \theta_f)$ , που δέχεται εικόνες  $x \in X$  και παράγει ενδιάμεσα χαρακτηριστικά  $f$ ,
- έναν label predictor (semantic classifier)  $G_y(f; \theta_y)$  που προβλέπει αν οι εικόνες είναι αληθινές ή παραπομένες με βάση τα ενδιάμεσα χαρακτηριστικά τους και
- έναν domain classifier  $G_d(f; \theta_d)$  που προβλέπει αν οι εικόνες ανήκουν στο  $S$  ή στο  $T$ .

Κατά τη φάση της εκπαίδευσης στόχος είναι να ελαχιστοποιηθεί το σφάλμα πρόβλεψης εικέτας για τις εικόνες που ανήκουν στο  $S$ , ώστε τα χαρακτηριστικά  $f$  και ο  $G_y$  να έχουν διακριτικές ιδιότητες και ικανότητα αντίστοιχα. Παράλληλα στόχος είναι και τα χαρακτηριστικά  $f$  να είναι αναλλοίωτα ως προς το domain, δηλαδή οι κατανομές  $S(f) = \{G_f(x; \theta_f) | x \sim S\}$  και  $T(f) = \{G_f(x; \theta_f) | x \sim T\}$  να είναι όμοιες, ώστε η ακρίβεια στο  $T$  να είναι ίδια με αυτή στο  $S$  - covariate shift assumption [55]. Δεδομένου ότι ο domain classifier  $G_d$  εκπαιδεύεται για να ξεχωρίζει τις δύο κατανομές ελαχιστοποιώντας την συνάρτηση απώλειας του, η λύση που προτείνεται είναι η γεννήτρια χαρακτηριστικών  $G_f$  να εκπαιδεύεται με τον αντίθετο ακριβώς στόχο, μεγιστοποιώντας την συνάρτηση απώλειας του domain classifier.

Η λύση αυτή, η οποία λειτουργεί παρόμοια με ένα GAN, μαθηματικά περιγράφεται από την συνάρτηση

$$E(\theta_f, \theta_y, \theta_d) = \sum_{\substack{i=1..N \\ d_i=0}} \mathcal{L}_y^i(\theta_f, \theta_y) - \lambda \sum_{i=1..N} \mathcal{L}_d^i(\theta_f, \theta_d) \quad (2.5)$$

η οποία μπορεί να βελτιστοποιηθεί με τη μέθοδο backpropagation, ελαχιστοποιούμενη από την γεννήτρια χαρακτηριστικών και τον label predictor και μεγιστοποιούμενη από

τον domain classifier:

$$(\hat{\theta}_f, \hat{\theta}_y) = \arg \min_{\theta_f, \theta_y} E(\theta_f, \theta_y, \hat{\theta}_d) \quad (2.6)$$

$$\hat{\theta}_d = \arg \max_{\theta_d} E(\hat{\theta}_f, \hat{\theta}_y, \theta_d) \quad (2.7)$$

Στο [56] η παραπάνω γενική προσέγγιση βρίσκει εφαρμογή στο πρόβλημα της ευρωστίας σε συμπίεση για ανίχνευση Deepfake. Προτείνεται η επιθολή ενός επιπλέον περιορισμού στην γεννήτρια χαρακτηριστικών που σχετίζεται με την ευκλείδεια απόσταση των χαρακτηριστικών που παράγονται από frames σε διαφορετικές συνθήκες συμπίεσης (διαφορετικά domains), ενώ τα αποτελέσματα της μεθόδου ξεπερνούν σε ακρίβεια πολλές προηγούμενες προσεγγίσεις. Στην παρούσα εργασία εξετάζονται υλοποιήσεις που βασίζονται στις δύο αυτές προσεγγίσεις [54, 56].

# 3

## Υλοποιήσεις

### 3.1 DATASET KAI PREPROCESSING PIPELINE

---

#### 3.1.1 FaceForensics++

To FaceForensics++ [11] (FF++) είναι ένα από τα πιο δημοφιλή σύνολα δεδομένων που περιέχουν παραποιήσεις προσώπων. Αποτελείται από 1000 πραγματικά βίντεο που προέρχονται είτε απευθείας από το YouTube, είτε από το Youtube-8m dataset [57] και τα οποία χρησιμοποιούνται για την δημιουργία παραποιημένων βίντεο μέσω μεθόδων Deep Learning (DL) αλλά και Computer Graphics (CG):

- DeepFakes <sup>1,2</sup> (DL, Facial identity swap),
- Face2Face [58] (CG, Facial reenactment),
- FaceShifter [16] (DL, Facial identity swap),
- FaceSwap <sup>3</sup> (CG, Facial identity swap) και
- NeuralTextures [59] (DL, Facial reenactment)

---

<sup>1</sup><https://github.com/deepfakes/faceswap/>

<sup>2</sup>Λόγω της συνωνυμίας με την γενική κατηγορία μεθόδων παραποίησης, η ειδική μέθοδος αναφέρεται ως DeepFakes (διαφοροποιείται στο κεφαλαίο γράμμα "F")

<sup>3</sup><https://github.com/MarekKowalski/FaceSwap/>



Σχήμα 3.1: Frame που δεν έχει υποστεί συμπίεση (αριστερά), σε ελαφριά συμπίεση (μέση) και σε έντονη συμπίεση (δεξιά) [61]

οπότε τελικά προκύπτουν 6000 βίντεο (1000 πραγματικά και 5000 αντίστοιχα παραπομένα).

Με σκοπό να αξιολογηθεί η ευρωστία ως προς την συμπίεση, τα βίντεο συμπιέζονται χρησιμοποιώντας το H.264 codec<sup>4</sup> του framework FFmpeg<sup>5</sup> με διαφορετικές τιμές της παραμέτρου Constant Rate Factor (CRF) που αντιστοιχούν σε διαφορετικό βαθμό συμπίεσης και διαφορετική ποιότητα: (i) c0 (Raw Quality - Raw), (ii) c23 (High Quality - HQ) και (iii) c40 (Low Quality - LQ). Ένα ενδεικτικό frame στις τρεις παραπάνω συνθήκες συμπίεσης παρουσιάζεται στο Σχήμα 3.1. Το τμήμα αυτό του συνόλου δεδομένων FF++ ονομάζεται FF (στην παρούσα εργασία) και συμπληρώνεται από ένα δεύτερο τμήμα, το Deepfake Detection dataset (DFD) [60], που αποτελείται από 363 πραγματικά βίντεο και 3086 παραπομένα που υπόκεινται στην ίδια διεργασία συμπίεσης.

Στην παρούσα εργασία το σύνολο δεδομένων FF++ προεπεξεργάζεται και χρησιμοποιείται για εκπαίδευση και έλεγχο. Συγκεκριμένα, το τμήμα του DFD (ποιότητες LQ και HQ) χρησιμοποιείται αποκλειστικά για έλεγχο, ενώ το τμήμα του FF χρησιμοποιείται για εκπαίδευση και έλεγχο των υλοποιήσεων που εξετάζονται.

### 3.1.2 Preprocessing Pipeline

Για την εξαγωγή των frames από τα βίντεο, γίνεται δειγματοληψία με ρυθμό 1 frame-per-second και χρησιμοποιείται μια υλοποίηση<sup>6</sup> ενός προεκπαιδευμένου συνελικτικού νευρωνικού δικτύου Multi-Task Cascaded Convolutional Neural Networks (MTCNN) [62] για την ανίχνευση των προσώπων σε κάθε δείγμα. Στην συνέχεια γίνεται περικοπή στην περιοχή του προσώπου χρησιμοποιώντας επεκτείνοντας το περιθώριο της

<sup>4</sup><https://www.itu.int/rec/T-REC-H.264/>

<sup>5</sup><https://ffmpeg.org/>

<sup>6</sup><https://github.com/timesler/facenet-pytorch/>

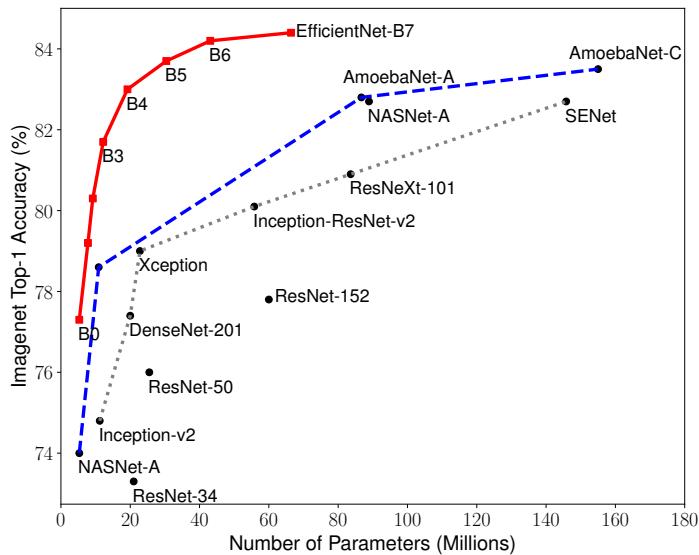
ανίχνευσης κατά έναν παράγοντα 1.3. Ο διαχωρισμός των δειγμάτων σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου (train/validation/test split, ποσοστά 72%/14%/14% αντίστοιχα) γίνεται βάσει των βίντεο από όπου εξήχθησαν και ακολουθεί τις οδηγίες του dataset ώστε να μην υπάρχουν επικαλύψεις προσώπων στα σύνολα εκπαίδευσης και ελέγχου.

## 3.2 ΜΟΝΤΕΛΑ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ

### 3.2.1 Backbone - Feature Generator: EfficientNet

Ως backbone των μοντέλων που αναπτύσσονται επιλέγεται η εκδοχή B0 από την οικογένεια συνελικτικών νευρωνικών δικτύων EfficientNet [63], αφού παρουσιάζει καλή αποδοτικότητα και ισορροπία μεταξύ της ακρίβειας και του αριθμού παραμέτρων. Αρχιτεκτονικές που έχουν ως backbone μοντέλα EfficientNet έχουν χρησιμοποιηθεί στη βιβλιογραφία για ανίχνευση Deepfake [64–66], ενώ η νικητήρια υλοποίηση<sup>7</sup> στον διαγωνισμό Deepfake Detection Challenge<sup>8</sup> βασίζεται επίσης σε αυτά.

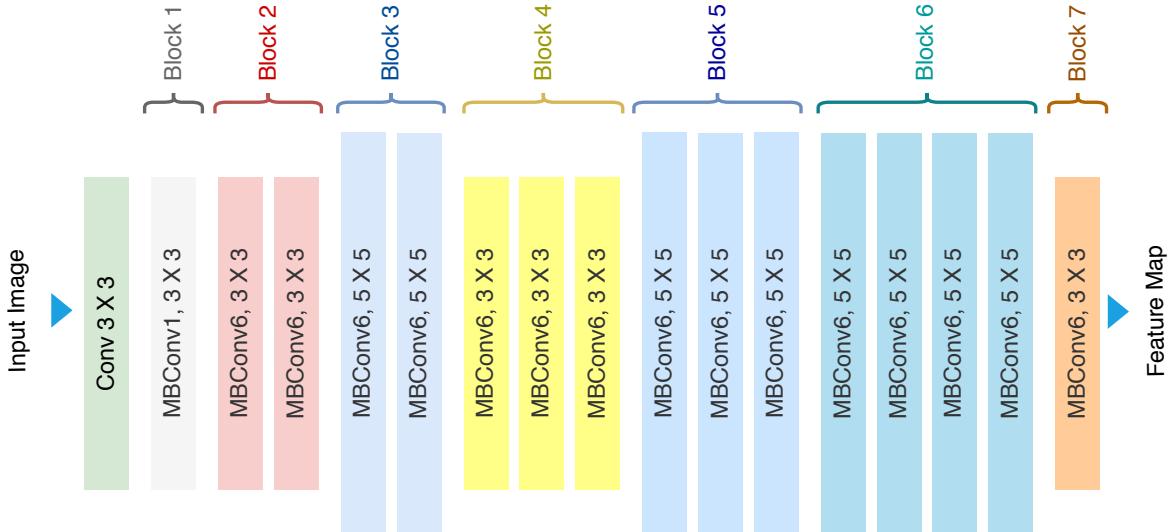
Σε πειράματα που παρουσιάζονται στη συνέχεια εξετάζεται, ακόμη, η χρήση της πιο ισχυρής εκδοχής B4, ενώ τα μοντέλα EfficientNet που χρησιμοποιούνται είναι προεκπαίδευμένα στο ImageNet [37] και επανεκπαιδεύονται σε μια διαδικασία fine-tuning.



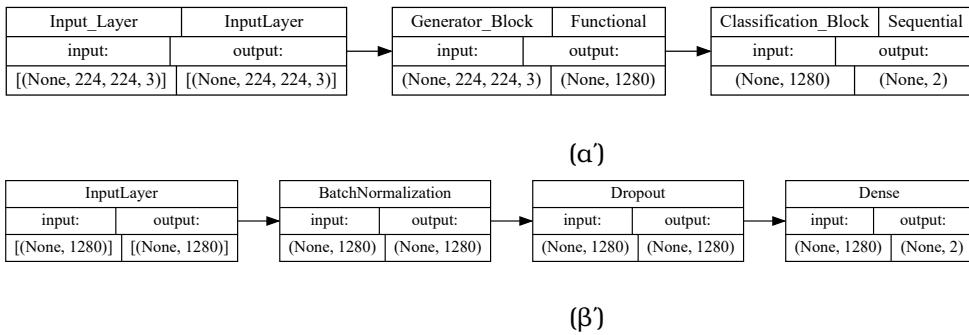
Σχήμα 3.2: Σύγκριση των μοντέλων EfficientNet με άλλα state-of-the-art μοντέλα με βάση την ακρίβεια και τον αριθμό παραμέτρων [63]

<sup>7</sup>[https://github.com/selimsef/dfdc\\_deepfake\\_challenge/](https://github.com/selimsef/dfdc_deepfake_challenge/)

<sup>8</sup><https://www.kaggle.com/competitions/deepfake-detection-challenge/>



Σχήμα 3.3: Η γενική αρχιτεκτονική του μοντέλου EfficientNet-B0 [67]

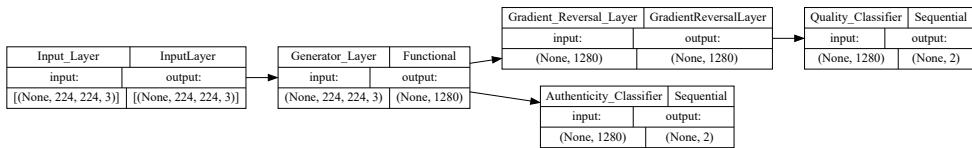


Σχήμα 3.4: (a') Η αρχιτεκτονική του baseline μοντέλου. (β') Η αρχιτεκτονική του ταξινομητή

### 3.2.2 Classifier και Baseline μοντέλο

Το μοντέλο που χρησιμοποιείται ως Baseline παρουσιάζεται στο Σχήμα 3.4α' και αποτελείται από το EfficientNet-B0 ως γεννήτρια χαρακτηριστικών (feature generator) και έναν ταξινομητή.

Ο ταξινομητής είναι το κοινό classification head που ενσωματώνεται σε όλα μοντέλα EfficientNet, ο οποίος ανακατασκευάζεται και εκπαιδεύεται εκ νέου. Είναι ένα σειρακό μοντέλο: Input layer → Batch-Normalization → Dropout layer → Dense output layer, όπως φαίνεται στο Σχήμα 3.4β', που δέχεται διάνυσμα χαρακτηριστικών και ταξινομεί τις εικόνες εισόδου σε παραπομένες και αληθινές. Όλοι οι ταξινομητές που χρησιμοποιούνται στην παρούσα εργασία έχουν αυτή τη δομή και χρησιμοποιούν ως συνάρτηση απώλειας την Cross-Entropy, εκτός αν αναφέρεται διαφορετικά.



Σχήμα 3.5: Η αρχιτεκτονική του Adversarial μοντέλου

### 3.2.3 Adversarial και Similarity μοντέλα

#### Adversarial μοντέλο

Το μοντέλο που βασίζεται στην προσέγγιση που ακολουθείται στο [54] ονομάζεται Adversarial, λόγω της ομοιότητας που εμφανίζει η εκπαίδευση του δικτύου με αυτή ενός Generative Adversarial Network. Η γεννήτρια χαρακτηριστικών (EfficientNet-B0) συνδέεται με έναν semantic (authenticity) classifier και μέσω ενός στρώματος Gradient Reversal Layer με έναν domain (quality) classifier, όπως φαίνεται στο Σχήμα 3.5.

Το δίκτυο εκπαιδεύεται όπως στο [54], ελαχιστοποιώντας την συνάρτηση

$$\mathcal{L}_{adv-model} = \mathcal{L}_y + \mathcal{L}_d \quad (3.1)$$

όπου  $\mathcal{L}_y$  και  $\mathcal{L}_d$  οι απώλειες που προκύπτουν από τον authenticity classifier και τον domain classifier αντίστοιχα και χρησιμοποιώντας την μέθοδο της οπισθοδιάδοσης.

Για την ενσωμάτωση του παράγοντα  $-\lambda$ , κατ' αντίστοιχία με την Εξίσωση 2.5, χρησιμοποιείται ένα στρώμα Gradient Reversal Layer, το οποίο κατά την φάση της διάδοσης προς τα εμπρός (forward propagation), δεν έχει καμία επίδραση (λειτουργεί όπως η ταυτοτική συνάρτηση), ενώ κατά την φάση της οπισθοδιάδοσης (back propagation) πολλαπλασιάζει την παραγώγου της απώλειας του quality classifier με τον παράγοντα  $-\lambda$ . Έτσι μπορεί οριστεί ως μία ψευδοσυνάρτηση  $R_\lambda(x)$  που αποτελείται από δύο ασύμβατες εξισώσεις για τις δύο αυτές φάσεις αντίστοιχα:

$$R_\lambda(x) = x \quad (3.2)$$

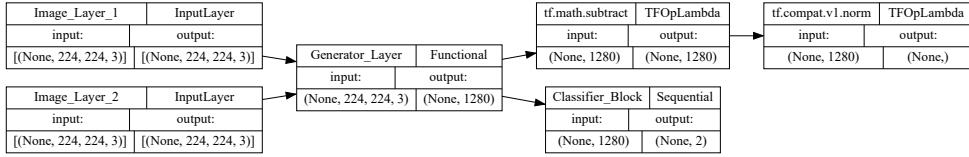
$$R_\lambda(x) = -\lambda I \quad (3.3)$$

όπου  $I$  είναι ο μοναδιαίος πίνακας.

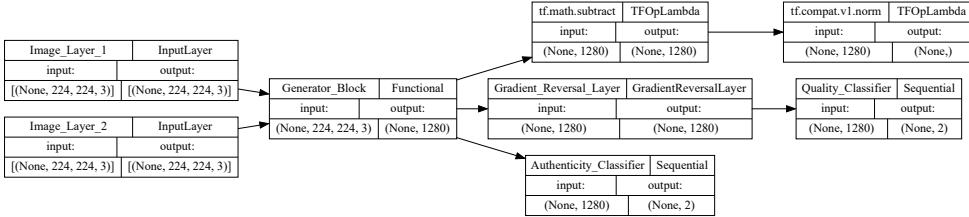
#### Similarity μοντέλο

Το μοντέλο που βασίζεται στην προσέγγιση του [56] ονομάζεται Similarity, λόγω της ομοιότητας (similarity) που προσπαθεί να πετύχει μεταξύ των χαρακτηριστικών από δύο domains.

Όπως φαίνεται στο Σχήμα 3.6, δέχεται ως είσοδο ένα ζεύγος από εικόνες ( $x_1, x_2$ ) οι



Σχήμα 3.6: Η αρχιτεκτονική του μοντέλου Similarity



Σχήμα 3.7: Η αρχιτεκτονική του μοντέλου Similarity - Adversarial

οποίες απεικονίζουν την ίδια εικόνα, ανήκουν, όμως, σε διαφορετικά επίπεδα συμπίεσης και για τις οποίες η γεννήτρια χαρακτηριστικών παράγει ενδιάμεσα χαρακτηριστικά  $f_1$  και  $f_2$  αντίστοιχα. Βάσει αυτών των χαρακτηριστικών και μέσω του ταξινομητή, παράγονται οι προβλέψεις αυθεντικότητας των εικόνων, ενώ παράλληλα δημιουργείται ως συνάρτηση απώλειας η ευκλείδεια απόσταση των  $f_1$  και  $f_2$ :

$$\mathcal{L}_{sim} = \|f_1 - f_2\|_2 \quad (3.4)$$

οπότε η συνολική συνάρτηση απώλειας που ελαχιστοποιείται είναι η

$$\mathcal{L}_{sim-model} = \mathcal{L}_{sim} + \mathcal{L}_y \quad (3.5)$$

όπου  $\mathcal{L}_y$  είναι η συνάρτηση απώλειας που προκύπτει από την έξοδο του ταξινομητή.

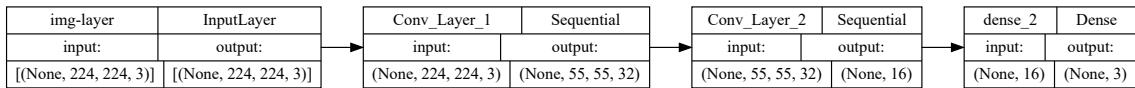
## Similarity - Adversarial μοντέλο

Βάσει των δύο προηγούμενων μοντέλων δημιουργείται το μοντέλο που ονομάζεται Similarity - Adversarial, το οποίο συνδυάζει τις δύο προηγούμενες προσεγγίσεις.

Λαμβάνει ως είσοδο ένα ζεύγος όμοιων εικόνων σε διαφορετικά επίπεδα συμπίεσης, παράγει εκτίμηση για την αυθεντικότητα τους και την απόσταση των ενδιάμεσων χαρακτηριστικών - όπως στο μοντέλο Adversarial - και παράγει επίσης εκτιμήσεις για το domain που ανήκει η κάθε εικόνα - όπως στο μοντέλο Similarity.

Η αρχιτεκτονική του μοντέλου παρουσιάζεται στο Σχήμα 3.7, ενώ η συνάρτηση απώλειας που ελαχιστοποιείται είναι η

$$\mathcal{L}_{sim-adv-model} = \mathcal{L}_y + \mathcal{L}_d + \mathcal{L}_{sim} \quad (3.6)$$



Σχήμα 3.8: Η δομή του ταξινομητή επιπέδου συμπίεσης (domain) για το μοντέλο Ensemble + CNN

κατά αντιστοιχία με τις εξισώσεις (3.1) και (3.5).

### 3.2.4 Ensemble μοντέλα

Με βάση την προσέγγιση που ακολουθείται στο [53] δημιουργούνται δύο μοντέλα Ensemble, καθένα από τα οποία χρησιμοποιούν επί μέρους μοντέλα που έχουν εξειδικευτεί σε ένα συγκεκριμένο επίπεδο συμπίεσης (domain) και έναν ταξινομητή επιπέδου συμπίεσης για την εξαγωγή βαρών και την στάθμιση της τελικής εξόδου, παρόμοια με την Εξίσωση 2.3.

Η διαφορά των δύο μοντέλων Ensemble εντοπίζεται στον ανιχνευτή επιπέδου συμπίεσης. Συγκεκριμένα, στο πρώτο μοντέλο (Ensemble + CNN) χρησιμοποιείται ένα νευρωνικό δίκτυο που έχει την γενική δομή που παρουσιάζεται στο Σχήμα 3.8. Αποτελέεται από δύο συνελικτικά στρώματα (με  $32 \times 3 \times 3$  και  $16 \times 7 \times 7$  φίλτρα αντίστοιχα και  $stride = 2$ ) με ενεργοποίηση ReLU και average pooling που ακολουθούνται από ένα dense στρώμα από όπου προκύπτει το διάνυσμα  $w$  των βαρών. Ο ανιχνευτής αυτός εκπαιδεύεται ξεχωριστά από τα εξειδικευμένα επί μέρους μοντέλα, χρησιμοποιώντας την cross-entropy loss function και είναι προφανές ότι λειτουργεί σε επίπεδο frame.

Αντίθετα, στο δεύτερο μοντέλο (Ensemble + Centroid ακολουθείται μια προσέγγιση για πρόβλεψη επιπέδου συμπίεσης η οποία είναι σε video level. Ως χαρακτηριστικό για την πρόβλεψη του επιπέδου συμπίεσης του δείγματος βίντεο χρησιμοποιείται ο λογάριθμος του κανονικοποιημένου ως προς την ανάλυση μέσου bitrate (log normalized average bitrate), ενώ ως ταξινομητής χρησιμοποιείται ένας Nearest Centroid classifier. Ο τελευταίος αναθέτει κάποιο δείγμα ελέγχου σε ένα επίπεδο συμπίεσης με βάση την απόσταση του χαρακτηριστικού από το μέσο χαρακτηριστικό για κάθε επίπεδο συμπίεσης. Τελικά, όλα τα δείγματα (frames) που προέρχονται από το ίδιο βίντεο αντιστοιχούν σε μία κοινή πρόβλεψη επιπέδου συμπίεσης.

# 4

## Πειράματα και Αποτελέσματα

Σε αυτό το κεφάλαιο περιγράφονται τα πειράματα που διεξήχθησαν και τα αποτελέσματα που προέκυψαν, καθώς και η εκπαίδευση και αξιολόγηση των μοντέλων.

Όλα τα μοντέλα και τα πειράματα υλοποιούνται χρησιμοποιώντας τις βιβλιοθήκες TensorFlow [68] και Keras [69] (έκδοση 2.13) στην υπερυπολογιστική συστοιχία "Αριστοτέλης" του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης.

### 4.1 ΜΕΤΡΙΚΗ ΑΞΙΟΛΟΓΗΣΗΣ ΚΑΙ ΒΑΡΗ ΑΝΑ ΚΑΤΗΓΟΡΙΑ

---

Όπως ήδη αναφέρθηκε το σύνολο δεδομένων που χρησιμοποιείται για εκπαίδευση είναι το τμήμα FF του FaceForensics++ [11], όπου ο λόγος των πραγματικών δεδομένων προς τα παραποιημένα είναι περίπου 1 : 5. Δεδομένης αυτής της δυσαναλογίας που εμφανίζεται μεταξύ των δύο κλάσεων, αντί της κοινής ακρίβειας (micro - average accuracy) χρησιμοποιείται η σταθμισμένη ακρίβεια (balanced accuracy ή macro - average accuracy).

Θεωρώντας ως θετικά (positive - P) τα παραποιημένα δεδομένα και αρνητικά (negative - N) τα πραγματικά, η μετρική balanced accuracy  $BA$  ορίζεται ως:

$$BA = \frac{TPR + TNR}{2} \tag{4.1}$$

όπου οι μετρικές του λόγου αληθών θετικών (True Positive Rate)  $TPR$  και του λόγου

αληθών αρνητικών (True Negative Rate)  $TNR$  ορίζονται αντίστοιχα ως:

$$TPR = \frac{TP}{TP + FN} \quad (4.2)$$

$$TNR = \frac{TN}{TN + FP} \quad (4.3)$$

Για τον ίδιο ακριβώς λόγο κατά τον υπολογισμό της συνάρτησης απώλειας Cross-Entropy χρησιμοποιούνται βάρη ανά κλάση. Έτσι η class-weighted Cross-Entropy για ένα δείγμα εισόδου  $x$  με αντίστοιχη έξοδο  $\hat{y}$  ορίζεται ως:

$$\mathcal{L}_{wCCE} = w_c \cdot y \cdot \log \hat{y} \quad (4.4)$$

όπου  $y$  η πραγματική ετικέτα (ground truth label) και  $w_c$  το διάνυσμα των βαρών, το οποίο αντίστοιχα ορίζεται ως:

$$w_c = \frac{|P| + |N|}{2} \left( \frac{1}{|N|}, \frac{1}{|P|} \right) \quad (4.5)$$

όπου με  $|P|$  και  $|N|$  συμβολίζεται το πλήθος των θετικών και αρνητικών δειγμάτων αντίστοιχα. Έτσι, όσο μικρότερο είναι το πλήθος των δειγμάτων μίας κλάσης, τόσο σημαντικότερη επίδραση έχει κάθε δείγμα της στον υπολογισμό της συνάρτησης απώλειας.

## 4.2 ΕΚΠΑΙΔΕΥΣΗ ΜΟΝΤΕΛΩΝ

---

Όλα τα μοντέλα που χρησιμοποιούν αρχιτεκτονική τύπου EfficientNet αρχικοποιούνται βάσει προεκπαίδευμένων μοντέλων στη βάση δεδομένων ImageNet [37]. Επανεκπαίδεύονται αρχικά για 10 epochs με την γεννήτρια χαρακτηριστικών ανενεργή (δεν εκπαίδευεται) και στην συνέχεια για 10 ακόμη epochs με 2 blocks<sup>1</sup> της γεννήτριας ενεργά.

Για την εκπαίδευση των μοντέλων χρησιμοποιείται ο Adam optimizer με learning rate =  $10^{-3}$  και προκαθορισμένες τις λοιπές παραμέτρους και batch size = 1024 frames. Στα μοντέλα Adversarial και Similarity - Adversarial χρησιμοποιείται scheduling για την υπερπαράμετρο  $\lambda$ , μεταβάλλοντας την σταδιακά από την τιμή 0 στην τιμή 1, όπως προτείνεται στο [54]:

$$\lambda_p = \frac{2}{1 + exp(-\gamma \cdot p)} - 1 \quad (4.6)$$

όπου  $\gamma = 10$  και  $p$  το ποσοστό της προόδου της εκπαίδευσης (από 0 έως 1).

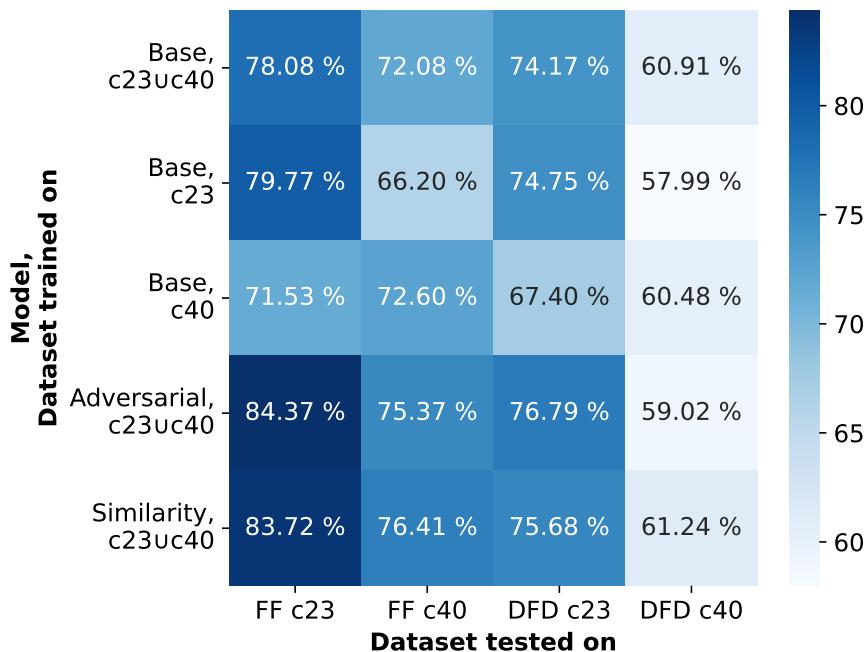
---

<sup>1</sup>Τα μοντέλα τύπου EfficientNet αποτελούνται από στρώματα που ομαδοποιούνται σε 7 τμήματα (blocks), όπως φαίνεται και στο Σχήμα 3.3

### 4.3 ΠΕΙΡΑΜΑ 1: ΕΛΕΓΧΟΣ ΠΡΟΣΑΡΜΟΓΗΣ ΣΕ ΣΥΜΠΙΕΣΗ ΚΑΙ ΓΕΝΙΚΕΥΣΗΣ ΣΕ ΆΛΛΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ

Το παρόν πείραμα διεξάγεται με σκοπό να εκτιμηθεί η ικανότητα των μοντέλων να προσαρμοστούν και ενδεχομένως να γενικεύσουν τόσο σε διαφορετικό βαθμό συμπίεσης, όπως και σε ένα διαφορετικό σύνολο εκπαίδευσης.

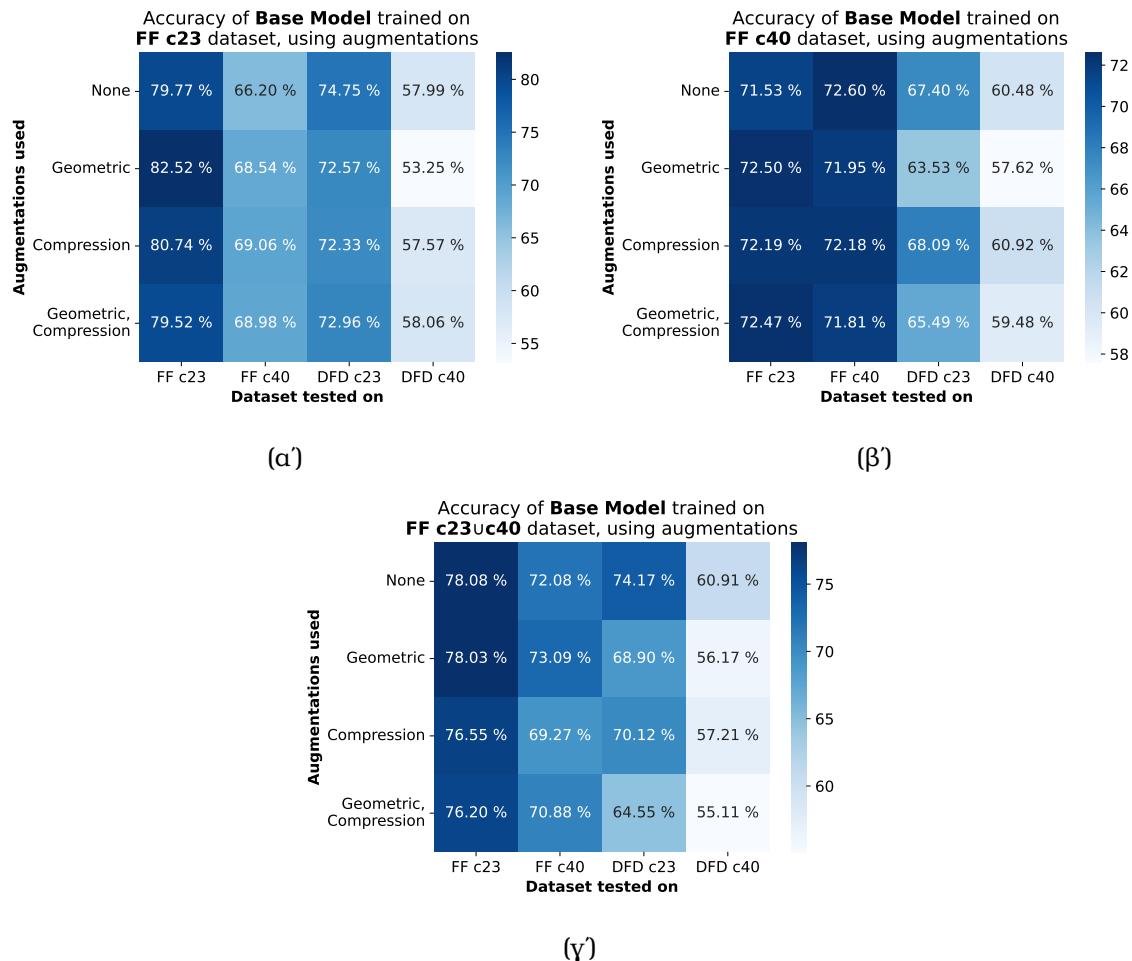
Έτσι, τα μοντέλα Baseline (Base), Similarity και Adversarial εκπαιδεύονται σε ένα σύνολο δεδομένων (συμβ. c23  $\cup$  c40) που περιέχει δεδομένα υψηλής (HQ - c23) και χαμηλής (LQ - c40) ποιότητας από το FF και αξιολογούνται στις ίδιες ποιότητες από το FF και το DFD. Επιπλέον το μοντέλο Baseline εκπαιδεύεται ξεχωριστά και στις δύο συνόλα δεδομένων διαφορετικού βαθμού συμπίεσης.



Σχήμα 4.1: Αποτελέσματα από το Πείραμα 1

Τα αποτελέσματα του πειράματος (balanced accuracy) παρουσιάζονται στο Σχήμα 4.1, όπου είναι εμφανές ότι τα μοντέλα εμφανίζουν καλή προσαρμογή εντός του dataset και του βαθμού συμπίεσης που εκπαιδεύονται, αλλά η ακρίβεια τους μειώνεται σημαντικά σε cross-dataset και cross-compression σενάρια.

Όπως είναι αναμενόμενο το Baseline μοντέλο που εκπαιδεύεται σε δεδομένα χαμηλής και υψηλής ποιότητας (c23  $\cup$  c40) εμφανίζει ακρίβεια παρόμοια με το μοντέλο που έχει εκπαιδευτεί μόνο στην αντίστοιχη κατηγορία. Τα μοντέλα Adversarial και Similarity συνολικά εμφανίζουν βελτιωμένη ακρίβεια σε σχέση με το Baseline μοντέλο, με εξαίρεση την περίπτωση του Adversarial μοντέλου στο σύνολο δεδομένων DFD c40.



Σχήμα 4.2: Αποτελέσματα για το μοντέλο Baseline όταν χρησιμοποιούνται επαυξήσεις και εκπαιδεύεται σε δεδομένα της κατηγορίας: (α') c23, (β') c40, (γ') c23 Υ c40

## Έλεγχος χρήσης επαυξήσεων

Για την βελτίωση ικανότητας γενίκευσης των μοντέλων ανίχνευσης και τη μείωση της υπερπροσαρμογής (overfitting) προτείνεται η χρήση επαυξήσεων (augmentations) τόσο στο γενικό πρόβλημα του Domain Generalization [50], όσο και σε μοντέλα ανίχνευσης Deepfake [64, 70–72].

Έτσι, για το μοντέλο Baseline και για τις κατηγορίες εκπαίδευσης c23, c40, c23 Υ c40 εξετάζεται επίσης η χρήση επαυξήσεων. Συγκεκριμένα, χρησιμοποιώντας την βιβλιοθήκη Albumentations [73] εξετάζεται η χρήση γεωμετρικών επαυξήσεων (Horizontal flip, Random Rotate 90°, Shift-scale-rotate), επαυξήσεων συμπίεσης εικόνας (JPEG compression) και η σύνθεση των δύο προηγούμενων κατηγοριών.

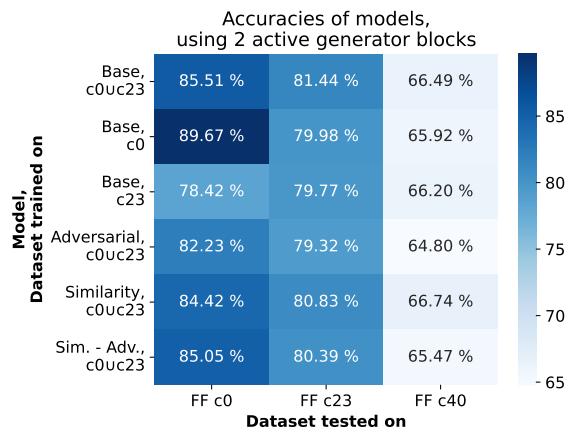
Τα αποτελέσματα του ελέγχου των μοντέλων ανά κατηγορία εκπαίδευσης παρουσιάζονται στα Σχήματα 4.2α', 4.2β' και 4.2γ' αντίστοιχα. Όπως παρατηρείται, η αύξηση της ακρίβειας είναι μικρή σε όλες τις κατηγορίες - σε μερικές μάλιστα εμφανίζεται και πτώση - και δεν είναι συνεπής. Επομένως, επιλέγεται να μην χρησιμοποιηθούν επαυξήσεις στα επόμενα πειράματα.

## 4.4 ΠΕΙΡΑΜΑ 2: ΧΡΗΣΗ ΑΣΥΜΠΙΕΣΤΩΝ ΚΑΙ ΧΑΜΗΛΟΥ ΒΑΘΜΟΥ

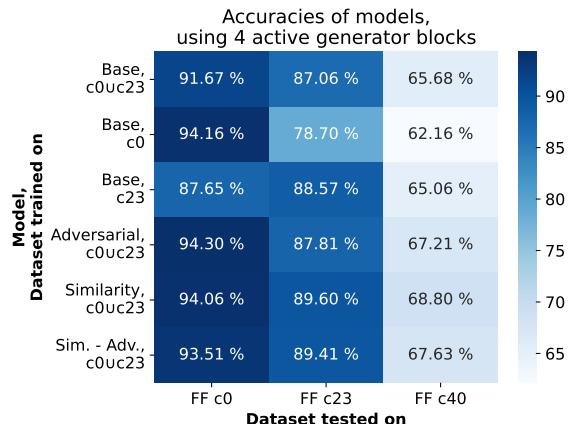
### ΣΥΜΠΙΕΣΗΣ ΔΕΔΟΜΕΝΩΝ

Στο παρόν πείραμα εξετάζεται η χρήση δεδομένων που δεν έχουν υποστεί συμπίεση ή έχουν υποστεί συμπίεση σε μικρό βαθμό. Η χρήση τέτοιων δεδομένων για εκπαίδευση των μοντέλων είναι πιθανό να οδηγεί σε ανάδειξη χαρακτηριστικών που θα είναι αναλλοίωτα ως προς (ή θα εξαρτώνται σε μικρότερο βαθμό από) την συμπίεση.

Για αυτό το σκοπό εκπαιδεύονται τα μοντέλα Baseline, Similarity, Adversarial και Similarity - Adversarial χρησιμοποιώντας ένα σύνολο δεδομένων εκπαίδευσης που περιέχει δεδομένα από τα επίπεδα συμπίεσης c0 και c23 (συμβ. c0 ∪ c23) και αξιολογούνται στις τρεις ποιότητες του τμήματος FF. Αξιολογείται επίσης η ακρίβεια του Baseline μοντέλου όταν εκπαιδεύεται στις κατηγορίες c0 και c23 ξεχωριστά, αλλά και η εκπαίδευση των μοντέλων με μεγαλύτερο αριθμό από blocks της γεννήτριας χαρακτηριστικών ενεργά (από 2 σε 4).



(α')



(β)

Σχήμα 4.3: Αποτελέσματα από το Πείραμα 2, όταν χρησιμοποιούνται: (α') 2 ενεργά blocks της γεννήτριας χαρακτηριστικών, (β') 4 ενεργά blocks

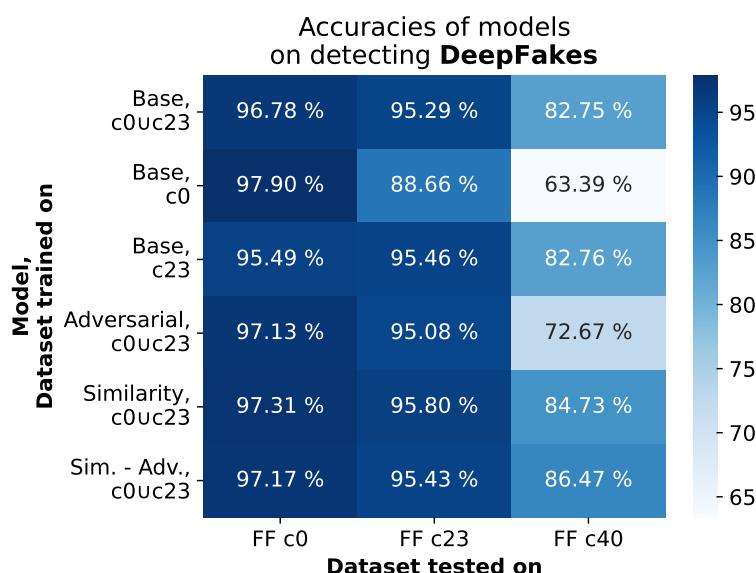
Τα αποτελέσματα για τις δύο επιλογές ενεργών block της γεννήτριας χαρακτηριστικών παρουσιάζοντα στα Σχήματα 4.3α' και 4.3β' αντίστοιχα. Η χρήση μεγαλύτερου αριθμού block οδηγεί σε καλύτερη ακρίβεια, με μικρή διαφορά, όμως, στο επίπεδο συμπίεσης c0 που δεν έχει ιδωθεί στην φάση της εκπαίδευσης, όπου και περιορίζεται κάτω από το 70%. Τα μοντέλα Adversarial, Similarity και Similarity - Adversarial εμφανίζουν καλύτερη ακρίβεια σε σχέση με το Baseline μοντέλο και στα τρία επίπεδα συμπίεσης που εξετάζονται στην περίπτωση της χρήσης τεσσάρων ενεργών block.

Συνολικά, η χρήση δεδομένων από τις κατηγορίες c0 και c23 οδηγεί σε υψηλότερη ακρίβεια συγκριτικά με την περίπτωση που χρησιμοποιούνται δεδομένα μόνο από την κατηγορία c23, με μικρή, όμως, βελτίωση.

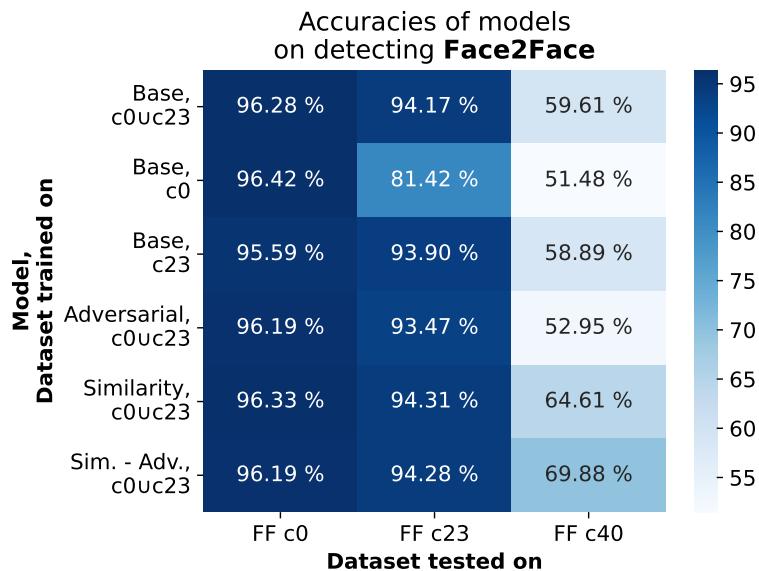
## 4.5 ΠΕΙΡΑΜΑ 3: ΕΞΕΙΔΙΚΕΥΣΗ ΣΕ ΈΝΑΝ ΤΥΠΟ ΠΑΡΑΠΟΙΗΣΗΣ

Δεδομένου ότι οι κατηγορίες παραποίησης που περιλαμβάνονται στο dataset FF++ χρησιμοποιούν διαφορετικές τεχνικές δημιουργίας Deepfake, είναι πιθανό η χρήση εξειδικευμένων μοντέλων ανίχνευσης για κάθε τύπο παραποίησης να οδηγεί σε καλύτερα αποτελέσματα όσον αφορά την ευρωστία σε συμπίεση.

Σε αυτό το πείραμα χρησιμοποιούνται τα μοντέλα Baseline, Adversarial, Similarity και Similarity - Adversarial, τα οποία εξειδικεύονται στην μέθοδο παραποίησης DeepFakes και την μέθοδο Face2Face. Τα δεδομένα εκπαίδευσης είναι από την κατηγορία c0 ∪ c23, ο έλεγχος γίνεται και στις τρεις ποιότητες του FF, ενώ για το μοντέλο Baseline εξετάζεται και η χρήση δεδομένων από τις επί μέρους κατηγορίες c0 και c23. Σε όλα τα μοντέλα χρησιμοποιούνται 4 ενεργά blocks της γεννήτριας χαρακτηριστικών.



Σχήμα 4.4: Αποτελέσματα για το Πείραμα 3, κατηγορία DeepFakes



Σχήμα 4.5: Αποτελέσματα για το Πείραμα 3, κατηγορία Face2Face

Από τα αποτελέσματα που παρουσιάζονται στα Σχήματα 4.4 και 4.5 φαίνεται ότι τα μοντέλα που εξετάζονται έχουν σημαντικά καλύτερη δυνατότητα γενίκευσης στη συμπίεση στην κατηγορία Deepfakes έναντι της κατηγορίας Face2Face.

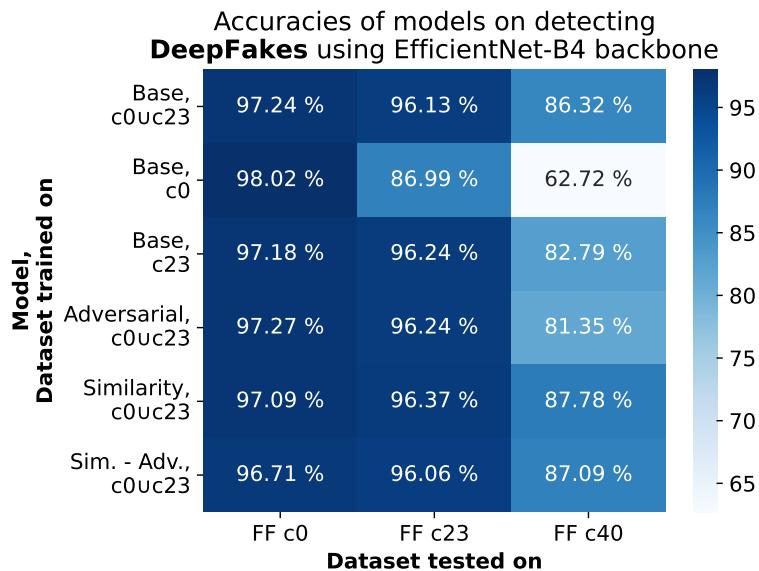
Στην πρώτη εκ των δύο, η ακρίβεια είναι σε υψηλά επίπεδα για τις κατηγορίες συμπίεσης που έχουν ιδωθεί στην εκπαίδευση - περίπου στο 90%, ενώ μπορεί και να ξεπεράσει το 80% στην κατηγορία c40. Στην κατηγορία Face2Face τα ποσοστά ακρίβειας είναι επίσης υψηλά για τους βαθμούς συμπίεσης c0 και c23, όμως μειώνονται δραματικά για τον βαθμό συμπίεσης c40.

Τα μοντέλα Baseline που χρησιμοποιούν μόνο δεδομένα από την κατηγορία c0 δεν επαρκούν για την ανίχνευση δεδομένων από την κατηγορία c40, ενώ συνολικά τα μοντέλα Similarity και Similarity - Adversarial εμφανίζουν καλύτερη ακρίβεια.

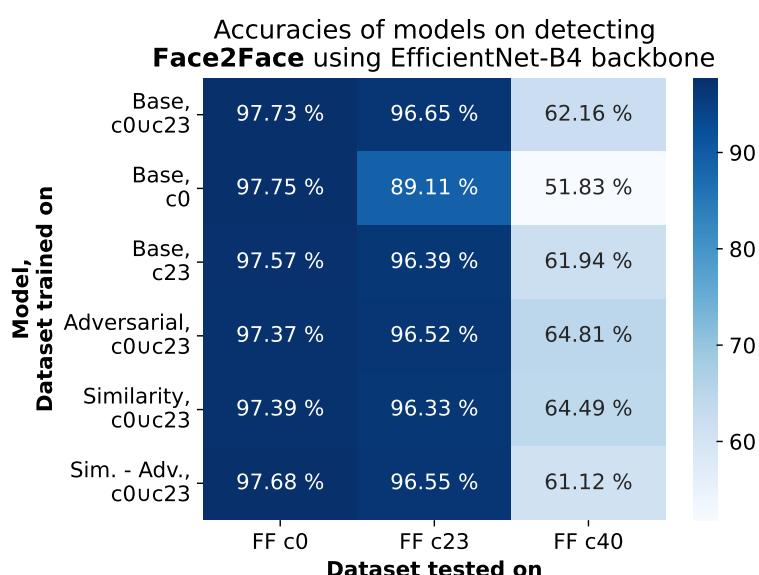
## Έλεγχος χρήσης ισχυρότερου backbone

Με σκοπό βελτίωση των αποτελεσμάτων των εξειδικευμένων μοντέλων, εξετάζεται η χρήση ενός ισχυρότερου μοντέλου ως backbone. Έτσι, επαναλαμβάνεται το Πείραμα 3, χρησιμοποιώντας ένα μοντέλο EfficientNet-B4 (αντί του B0) προεκπαίδευμένο στο ImageNet ως γεννήτρια χαρακτηριστικών των μοντέλων.

Τα αποτελέσματα που παρουσιάζονται στα Σχήματα 4.6 και 4.7 είναι παρόμοια με προηγουμένως, με μικρή βελτίωση της ακρίβειας στην κατηγορία DeepFakes συνολικά και μείωση της ακρίβειας στην κατηγορία Face2Face για τους περισσότερους συνδυασμούς μοντέλου εκπαίδευσης και ελέγχου.



Σχήμα 4.6: Αποτελέσματα για το Πείραμα 3, κατηγορία Deepfakes, όταν χρησιμοποιείται backbone EfficientNet-B4



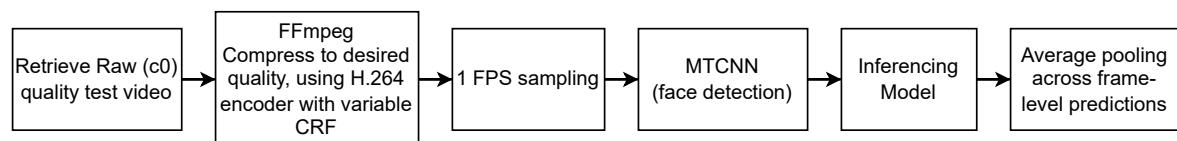
Σχήμα 4.7: Αποτελέσματα για το Πείραμα 3, κατηγορία Face2Face, όταν χρησιμοποιείται backbone EfficientNet-B4

## 4.6 ΠΕΙΡΑΜΑ 4: ΕΛΕΓΧΟΣ ΑΚΡΙΒΕΙΑΣ ΣΕ UNSEEN DOMAINS

---

### Video Inferencing Pipeline

Με σκοπό να ελεγχθεί η ακρίβεια των μοντέλων που χρησιμοποιούνται σε βαθμούς συμπίεσης που δεν έχουν ιδωθεί στην φάση της εκπαίδευσης, πχ. σε βαθμούς που βρίσκονται ενδιάμεσα αυτών που περιλαμβάνονται στο σύνολο εκπαίδευσης (c0, c23, c40) και για τους σκοπούς αυτού του πειράματος γίνονται κάποιες τροποποιήσεις στο inferencing pipeline.



Σχήμα 4.8: Η δομή του video inferencing pipeline

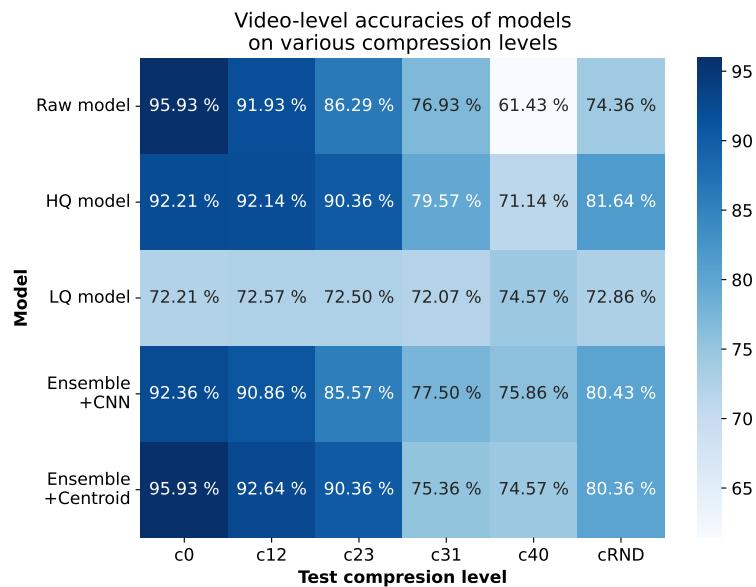
Έτσι, για την δημιουργία βίντεο που ανήκουν στους βαθμούς συμπίεσης c12 και c31 γίνεται χρήση του FFmpeg με κατάλληλη επιλογή της παραμέτρου Constant Rate Factor (CRF, με τιμές 12 και 31) για την για την συμπίεση των ασυμπίεστων (raw) βίντεο από το σύνολο ελέγχου, όπως παρουσιάζεται και στο Σχήμα 4.8. Στην συνέχεια γίνεται δειγματοληψία των βίντεο με ρυθμό 1 frame-per-second και ανίχνευση και περικοπή του μεγαλύτερου προσώπου κάθε frame με την χρήση του προεκπαιδευμένου δικτύου MTCNN. Τέλος, εισάγονται τα frames στο μοντέλο το οποίο χρησιμοποιείται και λαμβάνεται μία απόφαση σχετικά με την αυθεντικότητα σε επίπεδο βίντεο χρησιμοποιώντας average pooling στις αποφάσεις του μοντέλου σε επίπεδο frame.

Εκτός των δύο επιπλέον κατηγοριών c12 και c31, δημιουργείται και μια ακόμη κατηγορία, η οποία ονομάζεται cRND, και για την οποία η παράμετρος CRF είναι μια τυχαία μεταβλητή με ομοιόμορφη κατανομή στο εύρος [0, 51].

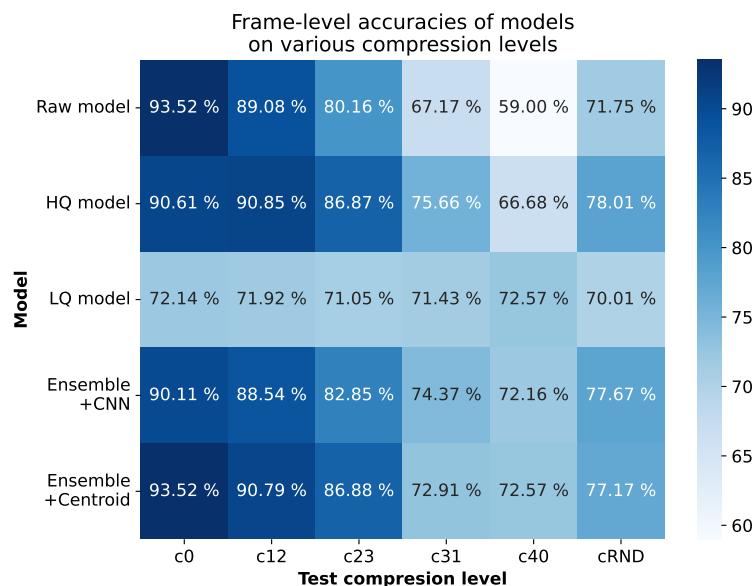
### Πείραμα

Στο παρόν πείραμα χρησιμοποιούνται τα μοντέλα Ensemble + CNN, Ensemble + GNB και τα επί μέρους εξειδικευμένα μοντέλα που τα συνθέτουν, δηλαδή τα μοντέλα Raw model, HQ model και LQ model τα οποία έχουν εκπαιδευτεί σε δεδομένα από τις κατηγορίες c0, c23 και c40 αντίστοιχα. Τα μοντέλα ελέγχονται τόσο σε επίπεδο βίντεο και σε επίπεδο frame στις κατηγορίες που έχει γίνει εκπαίδευση, αλλά και στις κατηγορίες c12, c31, cRND οι οποίες δεν έχουν χρησιμοποιηθεί στη φάση της εκπαίδευσης.

## Αποτελέσματα



Σχήμα 4.9: Αποτελέσματα για το Πείραμα 4 σε επίπεδο βίντεο



Σχήμα 4.10: Αποτελέσματα για το Πείραμα 4 σε επίπεδο frame

Τα αποτελέσματα σε επίπεδο βίντεο και σε επίπεδο frame παρουσιάζονται στα Σχήματα 4.9 και 4.10 αντίστοιχα.

Είναι εμφανές ότι τα εξειδικευμένα μοντέλα υπερέχουν στις αντίστοιχες κατηγορίες στις οποίες έχουν εκπαιδευτεί, όπως είναι και αναμενόμενο, ενώ γενικά παρατηρείται ότι η ακρίβεια αυξάνεται, σε μικρό βαθμό βέβαια, στο επίπεδο ελέγχου των βίντεο σε σύγκριση με το επίπεδο ελέγχου ανά frame. Η ακρίβεια των μοντέλων, πλην του

LQ model, για τις κατηγορίες με  $\text{CRF} \leq 23$  είναι άνω του 85%, ενώ για τις τρεις άλλες κατηγορίες περιορίζεται σημαντικά. Το LQ model παρουσιάζει γενικά σταθερή και συγκριτικά χαμηλή ακρίβεια  $\approx 72\%$  σε όλες τις κατηγορίες, εκτός της c40 (στην οποία έχει εκπαιδευτεί) όπου παρουσιάζει την δεύτερη καλύτερη ακρίβεια σε επίπεδο βίντεο.

Από τα εξειδικευμένα μοντέλα, το HQ model (εκπαιδευμένο στην κατηγορία c23) επιδεικνύει μάλλον την καλύτερη ισορροπία ως προς την ακρίβεια σε όλο το εύρος των εξεταζόμενων επιπέδων συμπίεσης. Τα δύο μοντέλα Ensemble παρουσιάζουν παρόμοια ακρίβεια με το HQ model, με εξαίρεση την κατηγορία c40 στην οποία εμφανίζουν ακρίβεια ίση ή υψηλότερη ακόμη και από το εξειδικευμένο μοντέλο σε επίπεδο βίντεο. Το μοντέλο Ensemble + Centroid επίσης παρουσιάζει την καλύτερη ακρίβεια, με μικρή, όμως, διαφορά στην unseen κατηγορία c12.

# 5

## Συμπεράσματα και Μελλοντική Εργασία

### 5.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

---

Στόχος της παρούσας διπλωματικής εργασίας ήταν η σύγκριση διαφορετικών υλοποιήσεων βασισμένες σε μοντέλα βαθιάς μάθησης για ανίχνευση Deepfake σε διαφορετικούς βαθμούς συμπίεσης. Ιδιαίτερη έμφαση δόθηκε στο πρόβλημα της ανίχνευσης σε βαθμούς συμπίεσης που δεν έχουν ιδωθεί στην φάση της εκπαίδευσης (Domain Generalization), το οποίο είναι ένα απαιτητικό και σημαντικό πρόβλημα, δεδομένης της εκτεταμένης διάδοσης των Deepfakes σε μέσα κοινωνικής δικτύωσης.

Εκτός ενός απλού Baseline μοντέλου, αναπτύχθηκαν μοντέλα με σκοπό την ανάδειξη χαρακτηριστικών που θα είναι αναλλοίωτα ως προς την συμπίεση (Adversarial, Similarity, Similarity - Adversarial μοντέλα), αλλά και δύο σύνθετα μοντέλα που χρησιμοποιούν επί μέρους μοντέλα που είναι εξειδικευμένα σε έναν βαθμό συμπίεσης σε συνδυασμό με ένα τμήμα ανίχνευσης του βαθμού συμπίεσης (Ensemble μοντέλα).

Τα πειράματα που διεξήχθησαν επιβεβαιώνουν συνολικά την αυξημένη δυσκολία που εντοπίζεται σε υψηλούς βαθμούς συμπίεσης, τόσο στην προσαρμογή (Domain Adaptation) όταν χρησιμοποιούνται δεδομένα εκπαίδευσης από αυτούς, όσο και της γενίκευσης (Domain Generalization) όταν δεν υπάρχουν διαθέσιμα δεδομένα εκπαίδευσης.

Πιο συγκεκριμένα, τα συμπεράσματα που προκύπτουν από τα πειράματα συνοψίζονται όπως παρακάτω:

- Μοντέλα, όπως τα Similarity και Adversarial, που αναδεικνύουν χαρακτηριστικά

αναλλοίωτα ως προς την συμπίεση είναι δυνατόν να επιδείξουν καλύτερη ακρίβεια ανίχνευσης Deepfake σε σύγκριση με μοντέλα με πιο απλή δομή, τόσο σε προ-βλήματα Adaptation όσο και Generalization όσον αφορά την συμπίεση.

- Η γενίκευση σε διαφορετικά σύνολα δεδομένων και σε ένα ευρύτερο φάσμα μεθόδων παραποίησης αν και δεν αποτελεί άμεσο στόχο της παρούσας εργασίας, είναι ένα σημαντικό ζήτημα σε πρακτικές εφαρμογές και εξετάστηκε σύντομα στο Πείραμα 1. Τα αποτελέσματα που προέκυψαν επιβεβαιώνουν ότι αποτελεί πρόβλημα το οποίο επίσης χρήζει ειδικής αντιμετώπισης.
- Ο συνδυασμός χρήσης δεδομένων μηδενικής συμπίεσης και ενός χαμηλού βαθμού συμπίεσης βελτίωσαν τα αποτελέσματα γενίκευσης των μοντέλων Similarity και Adversarial σε υψηλότερα επίπεδα συμπίεσης, όχι όμως σε σημαντικό βαθμό.
- Όπως παρουσιάστηκε σύντομα και στο Πείραμα 3, είναι πιθανό η χρήση μοντέλων που εξειδικεύονται σε έναν τύπο παραποίησης να οδηγεί σε ακριβέστερη ανίχνευση των Deepfake, υποθέτει βέβαια την γνώση των κατηγοριών παραποίησης και την πρόσθαση σε δεδομένα εκπαίδευσης από αυτές.
- Τα μοντέλα Ensemble εμφανίζουν παρόμοια ακρίβεια με μοντέλα που εκπαιδεύονται σε ένα μέσο επίπεδο συμπίεσης, εμφανίζουν, όμως, αυξημένη ακρίβεια σε υψηλότερους βαθμούς συμπίεσης, όπου τα τελευταία υστερούν σημαντικά.

## 5.2 ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

---

Η μελέτη που διεξήχθη στην παρούσα εργασία δύναται να αποτελέσει αφορμή για μελλοντικές προεκτάσεις και βελτιώσεις. Μερικές κατευθύνσεις στις οποίες μπορεί να αναζητηθούν νέες λύσεις είναι οι εξής:

- Εφαρμογή μοντέλων Adversarial, Similarity, Ensemble σε μοντέλα άλλης αρχιτεκτονικής. Στα πλαίσια της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκαν μοντέλα που ανήκουν αποκλειστικά στην οικογένεια EfficientNet, για λόγους εξοικονόμησης υπολογιστικών πόρων. Η χρήση διαφορετικών και πιο ισχυρότερων μοντέλων όπως για παράδειγμα του XceptionNet, είναι πιθανό να οδηγήσει σε καλύτερα αποτελέσματα. Πολύτιμη μπορεί να φανεί επίσης αξιοποίηση της χρονικής διάστασης στην ανίχνευση βίντεο Deepfake, η οποία παραβλήθηκε στην παρούσα εργασία για τον ίδιο λόγο.
- Χρήση μεγαλύτερων και νεότερων dataset. Υπάρχουν datasets όπως το DF-DC [38], το DeeperForensics [74] ή το ForgeryNet [75] τα οποία είναι μεγαλύτερα και νεότερα από το FaceForensics++ dataset που χρησιμοποιήθηκε. Η χρήση

νεότερων dataset, τα οποία ενδεχομένως περιέχουν και ένα ευρύτερο φάσμα μεθόδων παραποίησης σε συνδυασμό με τον μεγαλύτερο όγκο δεδομένων μπορεί να οδηγήσει τόσο σε υλοποιήσεις που έχουν καλύτερες δυνατότητες γενίκευσης, αλλά και να εξετάσει σε μεγαλύτερο βάθος υπάρχουσες υλοποιήσεις.

- Αξιολόγηση ανθρώπινης ακρίβειας στην ανίχνευση Deepfake. Ενδιαφέρον θα ήταν επιπλέον να υπάρξει κάποια μελέτη της ανθρώπινης ακρίβειας στο πρόβλημα των Deepfakes και να συγκριθεί με αυτή των αυτοματοποιημένων συστημάτων ανίχνευσης, εξετάζοντας και πάλι το ζήτημα σε ένα ευρύ διάστημα βαθμών συμπίεσης. Με αυτό τον τρόπο θα αξιολογηθεί και το ενδεχόμενο όφελος που προκύπτει από την χρήση των αυτοματοποιημένων συστημάτων, αλλά πιθανώς να υπάρξει και μια συστηματικότερη ανάλυση των δυνατοτήτων τους.

## Βιβλιογραφία

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [3] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, “Protecting world leaders against deep fakes.” in *CVPR workshops*, vol. 1, 2019, p. 38.
- [4] M. Westerlund, “The emergence of deepfake technology: A review,” *Technology innovation management review*, vol. 9, no. 11, 2019.
- [5] M.-H. Maras and A. Alexandrou, “Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos,” *The International Journal of Evidence & Proof*, vol. 23, no. 3, pp. 255–262, 2019.
- [6] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, “Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,” *Applied intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.
- [7] C. Gosse and J. Burkell, “Politics and porn: how news media characterizes problems presented by deepfakes,” *Critical Studies in Media Communication*, vol. 37, no. 5, pp. 497–511, 2020.
- [8] N. B. Abd Warif, A. W. A. Wahab, M. Y. I. Idris, R. Ramli, R. Salleh, S. Shamshirband, and K.-K. R. Choo, “Copy-move forgery detection: survey, challenges and future directions,” *Journal of Network and Computer Applications*, vol. 75, pp. 259–278, 2016.
- [9] L. Verdoliva, “Media forensics and deepfakes: an overview,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

- [10] A. Kumar, A. Bhavsar, and R. Verma, “Detecting deepfakes with metric learning,” in *2020 8th international workshop on biometrics and forensics (IWBF)*. IEEE, 2020, pp. 1–6.
- [11] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.
- [12] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.
- [13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [14] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, “Deep learning for deepfakes creation and detection: A survey,” *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022.
- [15] Y. Nirkin, Y. Keller, and T. Hassner, “Fsgan: Subject agnostic face swapping and reenactment,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7184–7193.
- [16] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, “Faceshifter: Towards high fidelity and occlusion aware face swapping,” *arXiv preprint arXiv:1912.13457*, 2019.
- [17] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.
- [18] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.
- [19] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, “Alias-free generative adversarial networks,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.
- [20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” *arXiv preprint arXiv:1710.10196*, 2017.
- [21] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,”

- in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.
- [22] Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of gans for semantic face editing,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9243–9252.
- [23] M. Koopman, A. M. Rodriguez, and Z. Geraarts, “Detection of deepfake video manipulation,” in *The 20th Irish machine vision and image processing conference (IMVIP)*, 2018, pp. 133–136.
- [24] F. Matern, C. Riess, and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 2019, pp. 83–92.
- [25] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” *arXiv preprint arXiv:1811.00656*, 2018.
- [26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [28] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 8261–8265.
- [29] P. Korshunov and S. Marcel, “Deepfakes: a new threat to face recognition? assessment and detection,” *arXiv preprint arXiv:1812.08685*, 2018.
- [30] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face x-ray for more general face forgery detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.
- [31] S. Lyu, “Deepfake detection: Current challenges and next steps,” in *2020 IEEE international conference on multimedia & expo workshops (ICMEW)*. IEEE, 2020, pp. 1–6.
- [32] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

- [33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [34] D. Cozzolino, G. Poggi, and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection,” in *Proceedings of the 5th ACM workshop on information hiding and multimedia security*, 2017, pp. 159–164.
- [35] B. Bayar and M. C. Stamm, “A deep learning approach to universal image manipulation detection using a new convolutional layer,” in *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016, pp. 5–10.
- [36] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [38] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, “The deepfake detection challenge (dfdc) preview dataset,” *arXiv preprint arXiv:1910.08854*, 2019.
- [39] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, “Deepfake detection: A systematic literature review,” *IEEE access*, vol. 10, pp. 25 494–25 513, 2022.
- [40] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.
- [41] Y. Li, M.-C. Chang, and S. Lyu, “In ictu oculi: Exposing ai created fake videos by detecting eye blinking,” in *2018 IEEE International workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.
- [42] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

- [43] J.-W. Seow, M.-K. Lim, R. C.-W. Phan, and J. K. Liu, “A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities,” *Neurocomputing*, 2022.
- [44] P. Yu, Z. Xia, J. Fei, and Y. Lu, “A survey on deepfake video detection,” *Iet Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.
- [45] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, “Deepfake video detection through optical flow based cnn,” in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.
- [46] R. Caldelli, L. Galteri, I. Amerini, and A. Del Bimbo, “Optical flow based cnn for detection of unlearnt deepfake manipulations,” *Pattern Recognition Letters*, vol. 146, pp. 31–37, 2021.
- [47] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.
- [48] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” *Neurocomputing*, vol. 312, pp. 135–153, 2018.
- [49] G. Wilson and D. J. Cook, “A survey of unsupervised deep domain adaptation,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 5, pp. 1–46, 2020.
- [50] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, “Generalizing to unseen domains: A survey on domain generalization,” *IEEE Transactions on Knowledge and Data Engineering*, 2022.
- [51] R. Polikar, “Ensemble learning,” *Ensemble machine learning: Methods and applications*, pp. 1–34, 2012.
- [52] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, “A survey on ensemble learning,” *Frontiers of Computer Science*, vol. 14, pp. 241–258, 2020.
- [53] M. Mancini, S. R. Bulo, B. Caputo, and E. Ricci, “Best sources forward: domain generalization through source-specific nets,” in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 1353–1357.
- [54] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

- [55] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.
- [56] J. Zhang, J. Ni, and H. Xie, “Deepfake videos detection using self-supervised decoupling network,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.
- [57] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, “Youtube-8m: A large-scale video classification benchmark,” *arXiv preprint arXiv:1609.08675*, 2016.
- [58] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.
- [59] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” *Acm Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.
- [60] N. Dufour and A. Gully, “Contributing data to deepfake detection research,” *Google AI Blog*, vol. 1, no. 2, p. 3, 2019.
- [61] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics: A large-scale video dataset for forgery detection in human faces,” *arXiv preprint arXiv:1803.09179*, 2018.
- [62] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
- [63] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [64] L. Bondi, E. D. Cannas, P. Bestagini, and S. Tubaro, “Training strategies and data augmentations in cnn-based deepfake video detection,” in *2020 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2020, pp. 1–6.
- [65] S. Baxevanakis, G. Kordopatis-Zilos, P. Galopoulos, L. Apostolidis, K. Levacher, I. Baris Schlicht, D. Teyssou, I. Kompatsiaris, and S. Papadopoulos, “The never deepfake detection service: Lessons learnt from developing and deploying in the

- wild,” in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, 2022, pp. 59–68.
- [66] S. Lee, J. An, and S. S. Woo, “Bznet: Unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos,” in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3500–3510.
- [67] T. Ahmed and N. H. N. Sabab, “Classification and understanding of cloud structures via satellite images with efficientunet,” *SN Computer Science*, vol. 3, pp. 1–11, 2022.
- [68] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org. [Online]. Available: <https://www.tensorflow.org/>
- [69] F. Chollet *et al.*, “Keras,” <https://keras.io>, 2015.
- [70] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “Cnn-generated images are surprisingly easy to spot... for now,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8695–8704.
- [71] S. Das, S. Seferbekov, A. Datta, M. S. Islam, and M. R. Amin, “Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3776–3785.
- [72] S. Tariq, S. Lee, and S. Woo, “One detector to rule them all: Towards a general deepfake attack detection framework,” in *Proceedings of the web conference 2021*, 2021, pp. 3625–3637.
- [73] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and flexible image augmentations,” *Information*, vol. 11, no. 2, 2020. [Online]. Available: <https://www.mdpi.com/2078-2489/11/2/125>
- [74] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2889–2898.

- [75] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, “Forgerynet: A versatile benchmark for comprehensive forgery analysis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4360–4369.

# Γλωσσάρι

**Animation** Εμψυχοποίηση

**Augmentations** Επαυξήσεις

**Backpropagation** Οπισθοδιάδοση

**Dataset** Σύνολο δεδομένων

**Decoder** Αποκωδικοποιητής

**Deepfake** Ψευδές ή/και παραπομένο πολυμεσικό περιεχόμενο που απεικονίζει πρόσωπα

**Domain** Πεδίο

**Encoder** Κωδικοποιητής

**Facial attribute editing** Τροποποίηση των χαρακτηριστικών των προσώπων

**Facial identity swap** Μεταφορά της ταυτότητας ενός προσώπου σε ένα άλλο

**Facial reenactment** Μεταφορά των εκφράσεων ενός προσώπου σε ένα άλλο

**Forgery detection** Ανίχνευση χαλκεύσεων/ παραποιήσεων

**Frame** Καρέ από βίντεο

**Fully synthetic face generation** Δημιουργία πλήρως συνθετικού προσώπου

**Handcrafted features** Αυτοσχέδια χαρακτηριστικά

**Label** Ετικέτα ταξινόμησης

**Logistic Regression Model** Μοντέλο λογιστικής παλινδρόμησης

**Loss function** Συνάρτηση απώλειας

**Machine Learning** Μηχανική μάθηση

**Manipulation** Παραποίηση

**Media Forensics** Εγκληματολογία μέσων ενημέρωσης

**Neural Network** Νευρωνικό Δίκτυο

**Overfitting** Υπερπροσαρμογή

**Partial facial occlusion** Μερική απόκρυψη προσώπου

**Representation Learning** Εκμάθηση αναπαράστασης

**Robustness** Ευρωστία

**Unsupervised Learning** Μη επιβλεπόμενη μάθηση

**Visual Artifact** Οπτικό εύρημα, στοιχείο που βοηθά στην ανίχνευση

**Zero-sum game** Παιγνιο μηδενικού αθροίσματος

# Ακρονύμια

**AE** Autoencoder

**AUC** Area Under Curve

**BA** Balanced Accuracy

**CG** Computer Graphics

**CNN** Convolutional Neural Network

**CRF** Constant Rate Factor

**DFD** Deepfake Detection dataset

**DL** Deep Learning

**FF++** FaceForensics++ [11]

**GAN** Generative Adversarial Network

**LRCN** Long Recurrent Convolutional Network

**LSTM** Long Short-Term Memory

**MLP** Multi-Layer Perceptron

**NN** Neural Network

**RNN** Recurrent Neural Network

**SVM** Support Vector Machine