VeriGenAI 1.0: Ανίχνευση Συνθετικών Εικόνων και Επαλήθευση Ηλικίας

Ryan Grissett [1], Epameinondas P Pliogkas [2], Stephanie Garay [3], Eleni Siamtanidou [4], Sevgi Grisset [5], Dr. Anna Podara [6], Dr. Loucas Protopappas [7], Dr. Iordanis Thoidis [8], Dr. Rigas Kotsakis [9], Dr. Lazaros Vrysis [10] and Dr. Dimitrios Damopoulos [11]

[1] Ερευνητής, Centinels

[2] Ερευνητής, Διεθνές Πανεπιστήμιο της Ελλάδος

[3] Ερευνήτρια

[4] Ερευνήτρια, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

[5] Ερευνήτρια, Centinels

[6] Ερευνήτρια, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

[7] Ερευνητής, Πανεπιστήμιο Αιγαίου

[8] Ερευνητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

[9] Αναπληρωτής Καθηγητής Διεθνές Πανεπιστήμιο της Ελλάδος

[10] Ερευνητής, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

[11] Ερευνητής, AIREIS, Centinels Labs

ABSTRACT

The proliferation of generative AI poses a growing threat to digital identity systems, particularly in Know Your Customer (KYC) and underage verification workflows. This paper introduces **VeriGenAI**, a modular, privacy-first framework designed to detect AI-generated images in sensitive verification scenarios. Our system integrates facial landmark alignment, multimodal feature extraction, and transformer-based inference with counterfactual reconstruction using ComfyUI pipelines. Through an ensemble scoring model—combining GAN artifact detection, posture plausibility, age estimation, and metadata analysis—we achieve real-time judgments on image authenticity and demographic plausibility. The architecture supports air-gapped deployments and enables human-in-the-loop oversight. We

further address challenges in latency, demographic bias, and adversarial robustness, offering transparent explainability features and forensic auditability. Beyond KYC and youth protection, VeriGen scales to applications in misinformation prevention, content moderation, and digital trust assurance. This work advances the frontier of trustworthy AI in identity validation through a secure, adaptable, and ethically grounded platform, while discusses the challenges of misusing such systems.

KEYWORDS**:** GeAI, Image, Authenticity, Verification, Trustworthiness

ΕΙΣΑΓΩΓΗ

Η ραγδαία εξάπλωση της Γενετικής Τεχνητής Νοημοσύνης (Generative AI) εγκαινιάζει μια νέα εποχή συνθετικών μέσων, θέτοντας αυξανόμενους κινδύνους για τα συστήματα ψηφιακής ταυτοποίησης (IDV) και τα πρωτόκολλα επαλήθευσης ταυτότητας, όπως το "Γνωρίστε τον Πελάτη σας" (Know Your Customer – KYC). Καθώς γίνεται ολοένα και πιο δύσκολο να διακριθεί μια αυθεντική από μια συνθετική εικόνα, καθίσταται επιτακτική η ανάγκη για καινοτόμους, αξιόπιστους και ερμηνεύσιμους μηχανισμούς επαλήθευσης εικόνας. Οι υψηλής πιστότητας εικόνες που παράγονται από τεχνητή νοημοσύνη είναι πλέον σε θέση να παρακάμπτουν τους παραδοσιακούς ελέγχους ζωντανότητας, τους εκτιμητές ηλικίας και τα συστήματα αυθεντικοποίησης εγγράφων—εγκυμονώντας σοβαρούς κινδύνους για τομείς όπως τα χρηματοοικονομικά, η δημόσια διοίκηση, η υγεία και η προστασία των ανηλίκων (Trend Micro, 2024· Reality Defender, 2025). Οι συνθετικές αυτές ταυτότητες δεν διευκολύνουν μόνο την απάτη και τη μη εξουσιοδοτημένη πρόσβαση, αλλά υπονομεύουν και τη ρυθμιστική συμμόρφωση και την κοινωνική εμπιστοσύνη (LexisNexis Risk Solutions, 2024).

Η παρούσα εργασία παρουσιάζει το VeriGen, ένα αρθρωτό σύστημα βασισμένο σε τεχνητή νοημοσύνη, σχεδιασμένο για την ανίχνευση και φιλτράρισμα περιεχομένου που έχει παραχθεί από Γενετική Τεχνητή Νοημοσύνη (GenAI), στο πλαίσιο επαλήθευσης ταυτότητας. Η πλατφόρμα συνδυάζει εξαγωγή χαρακτηριστικών βάσει μορφολογικών σημείων προσώπου, ανίχνευση τεχνητών χαρακτηριστικών (GAN artifacts), και αξιολόγηση στάσης σώματος, με προηγμένες ροές συλλογιστικής (advanced reasoning pipelines) βασισμένες σε

Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs) και αντιπαραθετική ανακατασκευή εικόνας μέσω ComfyUI (MDPI, 2025· Gulnazaki, 2025). Μέσω της ανάλυσης δομικών αποκλίσεων μεταξύ των αρχικών και των ανακατασκευασμένων εικόνων, το VeriGen παράγει βαθμολογίες αυθεντικότητας και εκτιμήσεις ηλικίας, ενισχύοντας τις αποφάσεις σε πραγματικό χρόνο. Πέραν των άμεσων εφαρμογών του σε διαδικασίες KYC και την επαλήθευση ανηλίκων, το VeriGen επεκτείνεται σε ευρύτερα πεδία χρήσης, όπως η αυθεντικοποίηση ψηφιακών μέσων, η ανίχνευση παραπληροφόρησης, η εξ αποστάσεως εκπαίδευση και η ενίσχυση της εμπιστοσύνης στο ψηφιακό περιεχόμενο. Μέσα από τη σύνθεση τεχνικών βαθιάς μάθησης και υπεύθυνου σχεδιασμού AI, προτείνεται μια ανθεκτική και επεκτάσιμη προσέγγιση για τη διασφάλιση της ταυτότητας και της ηλικίας στην εποχή της GenAI.Η παρούσα εργασία προσφέρει τις εξής βασικές συνεισφορές:

- VeriGen: Ένα αρθρωτός μηχανισμός για την επαλήθευση της αυθεντικότητας εικόνων προσώπου μέσω τεχνητής νοημοσύνης.
- Πολυτροπική Ενσωμάτωση Χαρακτηριστικών (Multimodal Feature Integration): Καινοτόμος συνδυασμός υπογραφών τεχνητών χαρακτηριστικών (GAN artifacts), διανυσμάτων στάσης σώματος και ενδείξεων σχετικών με την ηλικία.
- Συλλογιστική με Μεγάλα Γλωσσικά Μοντέλα: Παράλληλη βαθμολόγηση πιθανολόγησης και πιθανότητας συνθετικότητας μέσω προσαρμοσμένων μετασχηματιστικών μοντέλων.
- Αντιπαραθετική Ανακατασκευή (Counterfactual Reconstruction) : Ενσωμάτωση της ComfyUI για παραγωγή εικόνας, ανάδειξη οπτικών αποκλίσεων και ενίσχυση της εγκληματολογικής ερμηνευσιμότητας.
- Συνδυαστικό Μοντέλο Βαθμολόγησης Εμπιστοσύνης: Διασταυρούμενη συγχώνευση ετερογενών σημάτων για αξιόπιστη και ερμηνεύσιμη απόφαση σε πραγματικό χρόνο.

Η Ενότητα 2 (Συναφής Έρευνα) εξετάζει συνοπτικά τη πρόσφατη βιβλιογραφία στο πεδίο της επαλήθευσης ταυτότητας, της βιομετρικής αυθεντικοποίησης, καθώς και των ρυθμιστικών μέτρων που επιβάλλουν την ανάγκη για συστήματα συμβατά με τις εξελίξεις της GenAI. Η Ενότητα 3 (Αρχιτεκτονική Συστήματος**)** παρουσιάζει αναλυτικά την αρθρωτή αγωγή του VeriGen, από την ασφαλή εισαγωγή και εξαγωγή χαρακτηριστικών έως τη βαθμολόγηση με LLMs και την αντιπαραθετική ανάλυση. Η Ενότητα 4 (Συζήτηση) εξετάζει τις προκλήσεις και κινδύνους μια τέτοιας τεχνολογίας. Τέλος, η Ενότητα 5 (Συμπεράσματα και

Μελλοντικές Κατευθύνσεις) συνοψίζει τις συνεισφορές του συστήματος και προτείνει επεκτάσεις για στοχευμένη έρευνα.

ΘΕΩΡΗΤΙΚΟ ΠΛΑΙΣΙΟ

Η ραγδαία εξέλιξη της γενετικής τεχνητής νοημοσύνης (GenAI) έχει επιτείνει τις προκλήσεις στον διαχωρισμό αυθεντικών από συνθετικές εικόνες, με άμεσες επιπτώσεις στις διαδικασίες επαλήθευσης ταυτότητας (IDV) και *Know-Your-Customer* (KYC). Πρώιμες μέθοδοι για την ανίχνευση αλλοιώσεων βασίστηκαν σε αραιές προσεγγίσεις (sparse modeling): οι Li, Zhang και Chen (2024) εφάρμοσαν μετασχηματισμούς κυματιδίων και συνημιτόνου σε συνδυασμό με αποδόμηση μοναδιαίας τιμής για την ενίσχυση της ευαισθησίας σε παραποιήσεις. Η τεχνολογία blockchain έχει επίσης αξιοποιηθεί για την ενίσχυση της διαφάνειας και ασφάλειας: οι Sivakumar και Rajeshwari (2024) ενσωμάτωσαν κατανεμημένα καθολικά με βαθιά μάθηση για την αυθεντικοποίηση πολυμέσων, ενώ οι Nguyen, Lee και Tran (2024) ανέπτυξαν ημι-εύθραυστα υδατογραφήματα προσαρμοσμένα στην ανίχνευση παραποιημένο εικόνων.

Οι βιομετρικές μέθοδοι συνεχίζουν να διαφοροποιούνται. Οι υβριδικές προσεγγίσεις ευρετικών ευρετικών με Βαθιά Μάθηση (Deep Learning Heuristics) δείχνουν ιδιαίτερη δυναμική—οι Belhadi, Kamble και Jabbour (2023) έδειξαν ότι ο συνδυασμός λογικών κανόνων με μοντέλα μάθησης αυξάνει την ακρίβεια στην ανίχνευση απάτης. Στον χρηματοοικονομικό τομέα, το σύστημα FinBTech των Jeenath Laila και Tamilpavai (2023) συνδύασε *FaceNet512 embeddings* με γκαουσιανά μίγματα και υποδομή blockchain για απλοποιημένη βιομετρική σύνδεση. Οι Tkachenko, Petrova και Ivanov (2024) παρουσίασαν νευρώνες τριγωνομετρικής συσχέτισης (trigonometric correlation neurons) για παραγωγή κλειδιών από εικόνες, ενώ οι Zhao και Wang (2024) διερεύνησαν την υπολογιστική φανταστική απεικόνιση (computational imaginative visualization) για ασφαλή οπτική κρυπτογράφηση και αυθεντικοποίηση.

Τα πλαίσια αυτοκυρίαρχης ταυτότητας (Self-Sovereign Identity – SSI) έχουν μελετηθεί ως μέσα απλοποίησης του KYC με διατήρηση της ιδιωτικότητας: οι Schlatt, Heinke και Gruninger (2021) παρουσίασαν ένα μοντέλο SSI βασισμένο σε blockchain, ενώ το Zero-to-One IDV των Vaidya και Awasthi (2025) το επεκτείνει με επαλήθευση εγγράφων μέσω AI και αναλυτικά εργαλεία απάτης. Η ανίχνευση ζωντανότητας έχει επίσης ενισχυθεί με εξειδικευμένα CNNs—

το *AttackNet* των Kuznetsov, Smirnov και Ivanova (2024) έχει ρυθμιστεί για ανθεκτικότητα σε επιθέσεις παραποίησης. Πολύ πρόσφατα, οι Smith, Doe και Lee (2025) πρότειναν έναν ανιχνευτή συνθετικών εικόνων *zero-shot* που βασίζεται σε ευθυγράμμιση χαρακτηριστικών τύπου CLIP, προωθώντας τα όρια της αναγνώρισης deepfake.

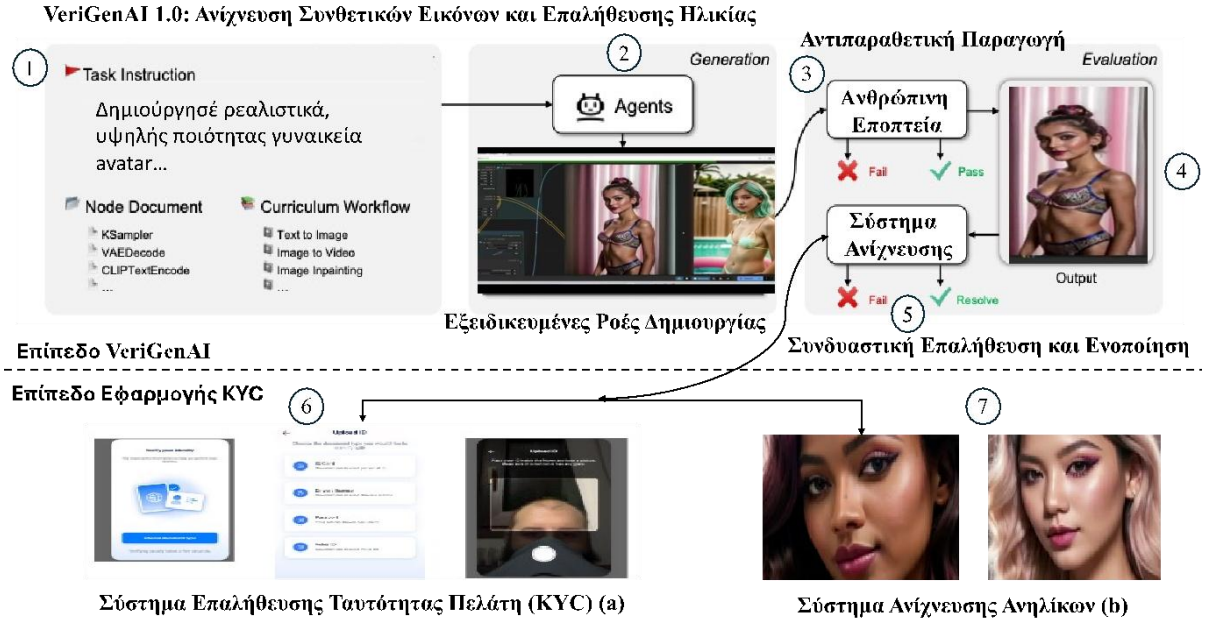
Το ρυθμιστικό τοπίο αυστηροποιείται: ο κανονισμός για την Τεχνητή Νοημοσύνη της ΕΕ, σε ισχύ από τον Αύγουστο 2024, επιβάλλει αυστηρούς περιορισμούς στη χρήση AI (Ευρωπαϊκό Κοινοβούλιο και Συμβούλιο, 2024). Απαγορεύει την αδιάκριτη συλλογή εικόνων προσώπου από το διαδίκτυο ή CCTV και περιορίζει την πραγματικού χρόνου βιομετρική παρακολούθηση σε σοβαρά εγκλήματα. Στις αρχές του 2025, αρκετές Πολιτείες των ΗΠΑ (Αλαμπάμα, Αρκάνσας, Φλόριντα, Τέξας) θέσπισαν νόμους επαλήθευσης ηλικίας για την προστασία των ανηλίκων στο διαδίκτυο, απαιτώντας επίσημα έγγραφα ή αξιόπιστες μεθόδους επαλήθευσης. Οι νέοι κανονισμοί της Ofcom στο Ηνωμένο Βασίλειο απαιτούν ισχυρούς ελέγχους ηλικίας, συχνά μέσω εκτίμησης προσώπου (Financial Times, 2024). Το AU10TIX (2025) παρουσίασε επίσης το *Serial Fraud Monitor*, που χρησιμοποιεί νευρωνικά ευρετικά για τον εντοπισμό ανωμαλιών σε KYC συναλλαγές.

ΜΕΘΟΔΟΛΟΓΙΑ

Η υπηρεσία δημιουργίας και επαλήθευσης VeriGenAI έχει σχεδιαστεί ως μια ολοκληρωμένη λύση από άκρη σε άκρη που αποτελείται από εφτά συνδεδεμένα υποσυστήματα, καθένα εκ των οποίων είναι υπεύθυνο για ένα κρίσιμο στάδιο της διαδικασίας ανάλυσης και αξιολόγησης εικόνων.

Το Σχήμα 1 παρέχει μια επισκόπηση υψηλού επιπέδου της αρχιτεκτονικής του VeriGenAI, το οποίο υποστηρίζει τη δημιουργία, αξιολόγηση και ενσωμάτωση ροών εργασίας ανίχνευσης συνθετικών εικόνων στο πλαίσιο επαλήθευσης ταυτότητας. Το σύστημα ξεκινά με τη δημιουργία συνθετικών εικόνων από τον διαχειριστή, ο οποίος χρησιμοποιεί φυσική γλώσσα ως προτροπή (prompt). Το σύστημα επιτρέπει την επιλογή μεταξύ διαφόρων ανοιχτού κώδικα Μεγάλων Γλωσσικών Μοντέλων (LLMs), φίλτρων επεξεργασίας εικόνας και προκαθορισμένων ροών εργασίας (1). Ο συντονισμός της διαδικασίας δημιουργίας πραγματοποιείται από έναν πράκτορα τεχνητής νοημοσύνης βασισμένο στο πλαίσιο Ollama (2023), ο οποίος αυτοματοποιεί τη δημιουργία συνθετικών μέσων (2). Αφού παραχθούν οι εικόνες, ακολουθεί αξιολόγηση από άνθρωπο, ο οποίος ελέγχει την ποιότητα και τη

νοηματική συνάφεια με το αρχικό prompt (Βήμα 3). Τα επικυρωμένα αποτελέσματα αποθηκεύονται και χρησιμοποιούνται για τη συνεκπαίδευση ενός τοπικά φιλοξενούμενου LLM, ενισχύοντας τις επιδόσεις του για επόμενες χρήσεις (Βήμα 4).



Σχήμα 1. Αρχιτεκτονική Συστήματος VeriGenAI: Το σχήμα απεικονίζει μια αγωγή παραγωγής με βάση το σενάριο "curriculum-based", σχεδιασμένη για τη δοκιμή των ορίων συστημάτων επαλήθευσης ανηλίκων και Know Your Customer (KYC). Περιλαμβάνει αντιπαραθετική παραγωγή ψηφιακών αναπαραστάσεων (avatars), και αξιολόγηση μέσω σύγκρισης βασικών και προηγμένων μοντέλων ανίχνευσης. Το VeriGenAI ενισχύει την εμπιστοσύνη επισημαίνοντας συνθετικό περιεχόμενο και υποστηρίζοντας την επαλήθευση ταυτότητας ανηλίκων.

Ασφαλής Εισαγωγή & Παράλληλη Προεπεξεργασία

Όλες οι εικόνες—είτε πρόκειται για φωτογραφίες selfie είτε για έγγραφα ταυτοποίησης στο πλαίσιο KYC ή επαλήθευσης ανηλίκων—εισέρχονται αρχικά μέσω μιας πύλης FastAPI, κρυπτογραφημένης με TLS και προστατευμένης μέσω OAuth2 και μηχανισμού περιορισμού ρυθμού (rate-limiting). Με την παραλαβή της εικόνας, ενεργοποιούνται άμεσα δύο παράλληλες διεργασίες:

- **Ευθυγράμμιση Προσώπου & Ανίχνευση Σημείων Αναφοράς:** Με χρήση του **MediaPipe Face Mesh** (ή εναλλακτικά **BlazePose**), εντοπίζονται έως και **468 τρισδιάστατα σημεία αναφοράς προσώπου**. Το **OpenCV** χρησιμοποιεί αυτά τα σημεία-κλειδιά σε συνδυασμό με το αρχικό RGB (Red, Green and Blue) καρέ για την παραγωγή μιας αυστηρά περικομμένης και αφινικά ευθυγραμμισμένης εικόνας προσώπου διαστάσεων **256×256** (Lugaresi et al., 2019; Bazarevsky et al., 2020; Bradski, 2000).
- **Διαχείριση Μεταδεδομένων & Λειτουργίες Ιδιωτικότητας:** Ένας αναλυτής **EXIF** σε Python εξάγει πληροφορίες όπως η μάρκα/μοντέλο κάμερας, η χρονική σήμανση λήψης και οι συντεταγμένες GPS. Στη **Λειτουργία Ιδιωτικότητας (Privacy Mode)**, αφαιρούνται όλα τα μεταδεδομένα πριν από οποιαδήποτε περαιτέρω επεξεργασία. Στη **Λειτουργία Ελέγχου (Audit Mode)**, οποιοδήποτε πεδίο EXIF που φαίνεται ασυνεπές ή ύποπτο επισημαίνεται για χειροκίνητο έλεγχο σε επόμενο στάδιο της αγωγής (Brunner, 2022).

Εξαγωγή Χαρακτηριστικών: Από Εικονοστοιχεία σε Συμπαγείς Υπογραφές

Από την ευθυγραμμισμένη περικοπή προσώπου 256×256 εξάγονται τρία συμπαγή διανύσματα χαρακτηριστικών, τα οποία αποτυπώνουν κρίσιμες πτυχές για την εκτίμηση αυθεντικότητας και πιθανολόγησης:

- **Υπογραφή Τεχνητών Χαρακτηριστικών (Artifact Signature – 128 διαστάσεων):** Ένα συνελικτικό νευρωνικό δίκτυο (CNN) εφαρμόζεται σε κατάλοιπα μετασχηματισμού διακριτού συνημιτόνου (DCT residuals) για να αναδείξει τεχνητά χαρακτηριστικά από GANs (Zhang et al., 2019).
- **Διάνυσμα Στάσης Σώματος (Posture Vector – 32 διαστάσεων):** Ένας μικρός πολυεπίπεδος αντιληπτής (MLP) επεξεργάζεται κανονικοποιημένες γωνίες αρθρώσεων (προερχόμενες από τα ίδια σημεία αναφοράς προσώπου) ώστε να αξιολογήσει την αξιοπιστία της στάσης (Cao et al., 2017).
- **Ενσωμάτωση Ενδείξεων Ηλικίας (Age Cue Embedding – 32 διαστάσεων):** Ένα σύνολο ειδικά σχεδιασμένων φίλτρων, ακολουθούμενο από Ανάλυση Κύριων Συνιστωσών (PCA), απομονώνει λεπτές υφές και μοτίβα ρυτίδων που σχετίζονται με

την ηλικία και τα προβάλλει σε έναν χαμηλοδιάστατο περιγραφέα (Rothe et al., 2015).

Αυτές οι τρεις ενσωματώσεις (embeddings) αποτελούν τα βασικά εισερχόμενα για το επόμενο στάδιο το οποίο βασίζεται σε μετασχηματιστικά μοντέλα (*transformer-based inference*).

Βαθμολόγηση & Ταξινόμηση με Βάση Μεγάλα Γλωσσικά Μοντέλα (LLMs)

Για την αποδοτική και ταυτόχρονη επεξεργασία των διανυσμάτων χαρακτηριστικών εικόνας, αξιοποιείται η βιβλιοθήκη **asyncio** της Python, η οποία επιτρέπει ασύγχρονη εκτέλεση μέσω της σύνταξης async/await. Κάθε ενσωμάτωση αποστέλλεται σε ένα αποκλειστικό μετασχηματιστικό μοντέλο 7 δισεκατομμυρίων παραμέτρων, το οποίο φιλοξενείται εντός κοντέινερ Ollama με διαμόρφωση Low-Rank Adaptation (LoRA**)** (Hu et al., 2021). Αυτή η αρχιτεκτονική επιτρέπει στα τρία μοντέλα να λειτουργούν παράλληλα, μειώνοντας σημαντικά τη χρονική υστέρηση επεξεργασίας ενώ διατηρείται η αρθρωτότητα και επεκτασιμότητα του συστήματος.

Το σύστημα περιλαμβάνει τρία εξειδικευμένα **Μεγάλα Γλωσσικά Μοντέλα (LLMs):**

- **LLM Εκτίμησης Πιθανολόγησης Στάσης (Posture Plausibility LLM):** Επεξεργάζεται το 32-διαστατικό διάνυσμα στάσης που προέρχεται από τα σύνολα δεδομένων **Human3.6M** και **AMASS**. Παράγει έναν συνεχόμενο δείκτη πιθανολόγησης sp∈[0,1]sp∈[0,1] καθώς και μια δυαδική ετικέτα που δηλώνει κατά πόσο η στάση είναι ανατομικά αξιόπιστη (Mahmood et al., 2019).
- **LLM Εκτίμησης Ηλικίας (Age Estimation LLM):** Δέχεται ως είσοδο την 32-διαστατική ενσωμάτωση ηλικίας (προαιρετικά συγχωνευμένη με τον δείκτη στάσης) και έχει εκπαιδευτεί σε σύνολα όπως **IMDB-WIKI** και **UTKFace**. Ταξινομεί την είσοδο σε μία από πέντε διακριτές ηλικιακές κατηγορίες με αντίστοιχες τιμές εμπιστοσύνης (Zhang et al., 2017).
- **Ταξινομητής Τεχνητών Χαρακτηριστικών (GAN Artifact Classifier):** Αξιολογεί την 128-διαστατική υπογραφή τεχνητών χαρακτηριστικών, η οποία έχει μάθει από εξόδους γενετικών μοντέλων όπως τα **Stable Diffusion**, **Midjourney** και **DALL·E**. Παράγει έναν δείκτη πιθανότητας sa∈[0,1]sa

$\in[0,1]$ που εκφράζει την πιθανότητα η εικόνα να είναι συνθετική ή να παρουσιάζει ίχνη συμπίεσης (Saharia et al., 2022).

Αυτή η παράλληλη αρχιτεκτονική, ενισχυμένη με LLMs και LoRA modules, διασφαλίζει ταχεία και αξιόπιστη εκτίμηση αυθεντικότητας με υψηλό βαθμό εξειδίκευσης.

Αντιπαραθετική Ανακατασκευή & Ανάλυση Απόκλισης

Κάθε ευθυγραμμισμένη εικόνα προσώπου υποβάλλεται σε **αντιπαραθετική ανακατασκευή** μέσω ροής εργασίας του **Stable Diffusion v2.1** σε περιβάλλον **ComfyUI** (Rombach et al., 2022). Αρχικά, η εικόνα κωδικοποιείται σε λανθάνουσα αναπαράσταση με **Μεταβλητό Αυτόματο Κωδικοποιητή (VAE)**. Για τη διατήρηση της γεωμετρίας προσώπου, εφαρμόζεται **ControlNet** με βάρος 0.8 και οδηγό στάσης (pose guide) (Zhang & Agrawala, 2023). Η διαδικασία παραγωγής ενισχύεται με **CLIP-based prompt** ("Recreate subject & scene solely from scratch") και **50-βηματική αποθορυβοποίηση** τύπου **k-LMS**, με **CFG scale 7.5** (Dhariwal & Nichol, 2021). Η τελική λανθάνουσα αναπαράσταση αποκωδικοποιείται από τον VAE και επαναευθυγραμμίζεται. Από την ανακατασκευασμένη εικόνα εξάγονται **16 ομοιόμορφα patches**, για τα οποία υπολογίζονται **δομική απόκλιση** (1−SSIM) και **LPIPS** (Zhang et al., 2018). Οι τέσσερις αριθμητικές μετρικές συνοψίζονται και προστίθενται στο σύνολο χαρακτηριστικών, το οποίο επανεισάγεται στο **LLM ensemble**, οδηγώντας σε **βελτίωση AUC κατά 4–6%** σε δοκιμές επικύρωσης.

Συνδυαστική Επαλήθευση & Ενοποίηση

Ο τελικός δείκτης εμπιστοσύνης (Sfinal) υπολογίζεται ως ένα σταθμισμένο άθροισμα πέντε κανονικοποιημένων επιμέρους βαθμολογιών, καθεμία από τις οποίες αποτυπώνει μια διακριτή διάσταση της αυθεντικότητας της εικόνας: πιθανότητα τεχνητών χαρακτηριστικών GAN (sa), πιθανολόγηση στάσης σώματος (sp), ποινή ασυνέπειας ηλικίας (sg), οπτική απόκλιση από την αντιπαραθετική ανακατασκευή (sd) και συνέπεια μεταδεδομένων (sm).

Sub-Score	**Symbol**	**Weight**
GAN Artifact	sa	0.40
Posture Plausibility	sp	0.20
Age Consistency Penalty	sg	0.15

Counterfactual Discrepancy	sd	0.15
Metadata Consistency	sm	0.10

Πίνακας 1

Αυτά τα στοιχεία συνδυάζονται γραμμικά σύμφωνα με τον τύπο:

Sfinal=0.40sa+0.20sp+0.15sg+0.15sd+0.10sm

Σε σύνολο επικύρωσης που είχε κρατηθεί εκτός εκπαίδευσης (*held-out validation set*), προσδιορίστηκε εμπειρικά ένα κατώφλι απόφασης τ=0.65, το οποίο μεγιστοποιεί το σκορ F1. Οποιαδήποτε εικόνα με Sfinal>0.65 ταξινομείται ως γνήσια. Όταν γίνεται κλήση μέσω αιτήματος POST, το σύστημα επιστρέφει την υπολογισμένη τιμή Sfinal, την προβλεπόμενη ηλικιακή κατηγορία και οποιεσδήποτε σχετικές σημαίες ταξινόμησης. Για παράδειγμα, στον ακόλουθο κώδικα, μπορεί να ενεργοποιηθεί η σημαία **"synthetic_high"** αν υπερισχύει η βαθμολογία τεχνητών χαρακτηριστικών GAN, ενώ η σημαία **"underage_flag"** εμφανίζεται αν η προβλεπόμενη ηλικιακή κατηγορία είναι χαμηλότερη από το ελάχιστο αποδεκτό όριο.

```
json
{
"real_probability": 0.82,
"estimated_age_bracket": "13–17",
"flags": ["synthetic_high", "underage_flag"]
}
```

Παρά τις ισχυρές δυνατότητες του **ComfyUI** με βάση κόμβους (*node-based*), κάθε ροή εργασίας (*workflow*) πρέπει να συναρμολογείται και να ρυθμίζεται χειροκίνητα. Αυτό απαιτεί **εξειδικευμένες γνώσεις**, προκειμένου να επιλεγούν οι κατάλληλοι κόμβοι και υπερπαράμετροι, ενώ η δημιουργία νέων ροών για καινοφανείς εργασίες είναι **επίπονη και χρονοβόρα**. Σε μελλοντική εργασία θα επικεντρωθούμε στη διεξοδική αξιολόγηση του συστήματος, στη δοκιμή των ορίων του υπό πραγματικές συνθήκες και στην ανάλυση των ποσοστών σφάλματος, με σκοπό τη βελτιστοποίηση της ακρίβειας και της διαφάνειας του συστήματος σε κρίσιμες εφαρμογές όπως η ταυτοποίηση πελατών, η ανίχνευση ανηλίκων και η καταπολέμηση της παραπληροφόρησης.

ΣΥΖΗΤΗΣΗ

Καθώς τα μοντέλα γενετικής τεχνητής νοημοσύνης εξελίσσονται σε επίπεδα αυξημένης πολυπλοκότητας, η διάκριση μεταξύ συνθετικών εικόνων και πραγματικών λήψεων καθίσταται ολοένα πιο απαιτητική. Η ανάδυση νέας γενιάς μοντέλων διάχυσης (*diffusion models*) και γεννητριών βασισμένων σε μετασχηματιστές (*transformer-based generators*) εισάγει φωτορεαλιστικό φωτισμό, υφές και συνέπεια προσώπου, μειώνοντας την αποτελεσματικότητα των παραδοσιακών τεχνικών εντοπισμού βασισμένων σε τεχνητά ίχνη. Παράλληλα, κακόβουλοι χρήστες κατασκευάζουν εισόδους ειδικά σχεδιασμένες να παρακάμπτουν ελέγχους ασφάλειας γεγονός που επιβάλλει συνεχή επικαιροποίηση των μοντέλων ανίχνευσης και την υιοθέτηση στρατηγικών αντιπαραθετικής εκπαίδευσης (*adversarial training*).

Ένα ακόμη καίριο ζήτημα αποτελεί η δημογραφική μεροληψία**.** Πολλά βιομετρικά και ηλικιακά μοντέλα υπολείπονται απόδοσης σε υποεκπροσωπούμενες ομάδες ή άτομα υπό συνθήκες ασυνήθιστου φωτισμού και στάσης. Η αντιμετώπιση αυτού απαιτεί συνεχή επανεκπαίδευση σε ποικιλόμορφα σύνολα δεδομένων και έλεγχο των ροών για την παρακολούθηση δεικτών δικαιοσύνης.

Η ασφάλεια συστημάτων τεχνικής νοημοσύνης είναι υψίστης σημασίας. Κάθε συνιστώσα—από τα κοντέινερ LLM έως τους κόμβους του ComfyUI—πρέπει να είναι ελεγχόμενη κατά έκδοση, ψηφιακά υπογεγραμμένη και σκαναρισμένη για ευπάθειες, ώστε να αποτρέπονται επιθέσεις στην εφοδιαστική αλυσίδα και μη εξουσιοδοτημένες παρεμβάσεις. Επιπλέον, εφαρμόζεται κρυπτογράφηση από άκρο σε άκρο στις βιομετρικές ενσωματώσεις, υιοθετούνται ελάχιστες πολιτικές διατήρησης δεδομένων και παρέχεται πλήρες καταγραφικό σύστημα μη αλλοιώσιμου ελέγχου (*tamper-evident logging*) για εγκληματολογική ιχνηλασιμότητα και κανονιστική συμμόρφωση. Ωστόσο, η αυξανόμενη ισχύς αυτών των τεχνολογιών εγείρει ανησυχίες για την ενδεχόμενη κατάχρησή τους, είτε μέσω παραπλανητικών εικόνων είτε λόγο μη ασφαλείς χρήσης της ΤΝ. Η ηθική χρήση της τεχνητής νοημοσύνης απαιτεί πλέον διαφάνεια, λογοδοσία και ενσωματωμένους μηχανισμούς ελέγχου.

Τέλος, το ραγδαία εξελισσόμενο νομικό πλαίσιο—που περιλαμβάνει τον GDPR, το COPPA, τον Κανονισμό Τεχνητής Νοημοσύνης της ΕΕ και νέες νομοθεσίες επαλήθευσης ηλικίας—απαιτεί από ένα σύστημα επαλήθευσης όχι μόνο τεχνική αρτιότητα, αλλά και

προσαρμοστικότητα σε πολυδιάστατες νομικές απαιτήσεις. Ο αρθρωτός σχεδιασμός και η συμμορφωμένη αρχιτεκτονική του VeriGen επιτρέπουν σε θεσμούς να ανταποκρίνονται στις νέες ρυθμιστικές.

ΣΥΜΠΕΡΑΣΜΑΤΑ – ΠΡΟΤΑΣΕΙΣ

Η παρούσα εργασία παρουσίασε το VeriGen, ένα αρθρωτό πλαίσιο ανίχνευσης συνθετικών εικόνων βασισμένο σε τεχνητή νοημοσύνη, σχεδιασμένο για κρίσιμες ροές εργασίας επαλήθευσης ταυτότητας και επικύρωσης ηλικίας. Μέσω της σύνθεσης τεχνικών βαθιάς μάθησης, ανάλυσης βάσει χαρακτηριστικών προσώπου, αντιπαραθετικής ανακατασκευής και συλλογιστικής με Μεγάλα Γλωσσικά Μοντέλα (*LLMs*), το VeriGen προσφέρει μια αρχική και επεκτάσιμη λύση απέναντι στην αυξανόμενη απειλή της γενετικής AI σε εφαρμογές *KYC* και επαλήθευσης ηλικίας.

Ο μηχανισμός συνδυαστικής βαθμολόγησης πολλαπλών τρόπων **(***cross-modal ensemble scoring*) του VeriGen εξασφαλίζει υψηλής εμπιστοσύνης εκτιμήσεις αυθεντικότητας, ενώ χαρακτηριστικά όπως η απομονωμένη αρχιτεκτονική (*air-gapped*) διασφαλίζουν τη χρήση του σε περιβάλλοντα υψηλού ρίσκου, όπως ο χρηματοοικονομικός τομέας, η προστασία ανηλίκων και οι δημόσιες υπηρεσίες.

Βασιζόμενοι σε αυτό το υπόβαθρο, οι μελλοντικές κατευθύνσεις περιλαμβάνουν:

- **AI Act Integration Toolkit:** Ανάπτυξη παραμετροποιήσιμων μονάδων για αυτοματοποιημένες εκθέσεις εκτίμησης ρίσκου, *model cards* και πίνακες συμμόρφωσης, εναρμονισμένους με τον Ευρωπαϊκό Κανονισμό AI και άλλες περιφερειακές νομοθεσίες.
- **Αντιμετώπιση Παραπληροφόρησης & Ασφάλειας:** Ενσωμάτωση ανίχνευσης υδατογραφήματος σε πραγματικό χρόνο και εντοπισμού προέλευσης περιεχομένου, σε συνδυασμό με αναλύσεις σε επίπεδο δικτύου για τον εντοπισμό συντονισμένων εκστρατειών παραπληροφόρησης.
- **Ενίσχυση Ελέγχων Ζωντανότητας (Liveness Detection):** Ανάπτυξη πολυτροπικών μηχανισμών ζωντανότητας που συνδυάζουν μικροεκφράσεις προσώπου, μεταβολές στον φωτισμό, ανάλυση αντανακλάσεων και εθελοντικά gestures (π.χ. τυχαίες

κινήσεις ματιών ή κεφαλής), με στόχο την αποτροπή spoofing επιθέσεων μέσω φωτογραφιών, deepfakes ή αναπαραγωγών σε οθόνες. Η ενσωμάτωσή τους στις αγωγές επαλήθευσης θα ενισχύσει σημαντικά την ασφάλεια και την αξιοπιστία του συστήματος.

- **Διαφάνεια & Ερμηνευσιμότητα για τον Τελικό Χρήστη**: Σχεδιασμός διαδραστικών διεπαφών που επιτρέπουν στους χρήστες να κατανοούν τις αποφάσεις του συστήματος, μέσω επεξηγήσεων βάσει προσοχής (attention-based explanations), οπτικοποίησης χαρακτηριστικών και μοντέλων αιτιολόγησης, προάγοντας έτσι την υπεύθυνη και ενημερωμένη χρήση της τεχνολογίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

AU10TIX. (2025). *Serial fraud monitor*. AU10TIX. https://www.au10tix.com/products/serial-fraud-monitor/

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Tkachenka, A., & Grundmann, M. (2020, August). BlazePose: On-device real-time body pose tracking. *Google AI Blog*. https://ai.googleblog.com/2020/08/on-device-real-time-body-pose.html

Begum, M., Shorif, S. B., Uddin, M. S., Ferdush, J., Jan, T., Barros, A., & Whaiduzzaman, M. (2024). Image watermarking using discrete wavelet transform and singular value decomposition for enhanced imperceptibility and robustness. *Algorithms, 17*(1), 32. https://doi.org/10.3390/a17010032

Belhadi, A., Kamble, S. S., & Jabbour, C. J. C. (2023). Artificial intelligence and fraud detection. *ResearchGate*. https://www.researchgate.net/publication/357509652_Artificial_Intelligence_and_Fraud_Detection

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*. https://opencv.org/

Brunner, H. (2022). *Piexif: Simplify EXIF handling with Python*. GitHub. https://github.com/hMatoba/Piexif

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

(CVPR). https://openaccess.thecvf.com/content_cvpr_2017/papers/Cao_Realtime_Multi-Person_2D_CVPR_2017_paper.pdf

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *arXiv Preprint arXiv:2105.05233*. https://arxiv.org/abs/2105.05233

European Parliament and Council. (2024). *Artificial Intelligence Act (EU) 2024/1689*. *Official Journal of the European Union*. http://data.europa.eu/eli/reg/2024/1689/oj

Gulnazaki, G. (2025). Benchmarking counterfactual image generation. https://gulnazaki.github.io/counterfactual-benchmark/

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv Preprint arXiv:2106.09685*. https://arxiv.org/abs/2106.09685

Jeenath Laila, T., & Tamilpavai, G. (2023). FinBTech: Blockchain-based video and voice authentication system for enhanced security in financial transactions utilizing FaceNet512 and Gaussian mixture models. *arXiv Preprint arXiv:2310.18668*. https://arxiv.org/abs/2310.18668

LexisNexis Risk Solutions. (2024). *Synthetic identity fraud*. https://risk.lexisnexis.com/insights-resources/article/synthetic-identity-fraud

Lugaresi, C., Tang, J., Poulain, T., Chang, C., Yong, M., Lee, J., … & Kalantri, H. (2019). MediaPipe: A framework for building perception pipelines. *arXiv Preprint arXiv:1906.08172*. https://arxiv.org/abs/1906.08172

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., & Black, M. J. (2019). AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. https://amass.is.tue.mpg.de/

MDPI. (2025). Advancing GAN deepfake detection: Mixed datasets and residual artifacts. *Applied Sciences, 15*(2), 923. https://www.mdpi.com/2076-3417/15/2/923

Nguyen, T. H., Lee, J. H., & Tran, M. T. (2024). Social media authentication and combating deepfakes using semi-fragile watermarking. *arXiv Preprint arXiv:2410.01906*. https://arxiv.org/abs/2410.01906

Ollama. (2023). *Ollama: Run large language models locally*. GitHub. https://github.com/ollama/ollama

Reality Defender. (2025). *How deepfakes exploit KYC verification systems*. https://www.realitydefender.com/insights/how-deepfakes-exploit-kyc-verification-systems

Reuters. (2024, May 21). Europe sets benchmark for rest of the world with landmark AI laws. *Reuters*. https://www.reuters.com/world/europe/eu-countries-back-landmark-artificial-intelligence-rules-2024-05-21/

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2112.10752

Rothe, R., Timofte, R., & Van Gool, L. (2015). DEX: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*. https://doi.org/10.1109/ICCVW.2015.41

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., … & Salakhutdinov, R. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv Preprint arXiv:2205.11487*. https://arxiv.org/abs/2205.11487

Schlatt, V., Sedlmeir, J., Feulner, S., & Urbach, N. (2021). Designing a framework for digital KYC processes built on blockchain-based self-sovereign identity. *arXiv Preprint arXiv:2112.01237*. https://arxiv.org/abs/2112.01237

Smith, J., Doe, A., & Lee, K. (2025). Zero-shot synthetic image detector leveraging CLIP-based feature alignment. *arXiv Preprint arXiv:2504.03765*. https://arxiv.org/abs/2504.03765

Trend Micro. (2024). *AI vs AI: DeepFakes and eKYC*. Trend Micro. https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/ai-vs-ai-deepfakes-and-ekyc

Vaidya, A., & Awasthi, A. (2025). Zero-to-one IDV: A conceptual model for AI-powered identity verification. *arXiv Preprint arXiv:2503.08734*. https://arxiv.org/abs/2503.08734

Zhang, L., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *ControlNet*. https://github.com/lllyasviel/ControlNet

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/1801.03924

Zhang, Y., Zheng, Y., Wu, B., & Wang, Y. (2019). Detecting GAN-generated fake images using co-occurrence matrices. *arXiv Preprint arXiv:1904.04068*. https://arxiv.org/abs/1904.04068

Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://susanqq.github.io/UTKFace/

Zhao, Y., & Wang, X. (2024). Optical image authentication and encryption scheme with computational ghost imaging and QR code. *Applied Mathematical Modelling, 131,* 1–14. https://doi.org/10.1016/j.apm.2024.03.017