Σύστημα Υποστήριξης Κλινικών Αποφάσεων
για τη Νόσο των Ανευρυσμάτων Κοιλιακής Αορτής
Βασισμένο σε Μοντέλα Τεχνητής Νοημοσύνης

Παραδοτέο Π3.3

Υποδομή Δεδομένων Νέφους και Ομοσπονδιακές Βάσεις: Εργαλείο Προτυποποίησης και Εναρμόνισης Ιατρικών Δεδομένων

Όνομα Αρχείου:	Safe-Aorta-Π3.3-Υποδομή Νέφους και Ομοσπονδιακές Βάσεις Δεδομένων (Τμήμα 3)	**Επίπεδο Διάδοσης:**	Δημόσιο
Ημερομηνία Υποβολής:	Οκτώβριος 2025 (M27)	**Κωδικός Έργου:**	TAEDR-0535983
Κοινοπραξία:	ΕΜΠ, ΠΔΜ, ΠΚ, ΕΛΜΕΠΑ, ΠΑΔΑ, ΙΙΒΕΑΑ, ΠΑΠΕΛ	**Υπεύθυνος Παραδοτέου:**	Πανεπιστήμιο Ιωαννίνων (ανάδοχος)
Διάρκεια:	28 μήνες	**Κατάσταση:**	Τελικό


Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ
ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ


ΓΓΕΚ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ


Με τη χρηματοδότηση
της Ευρωπαϊκής Ένωσης
NextGenerationEU

ΛΙΣΤΑ ΣΥΓΓΡΑΦΕΩΝ

Συγγραφείς				
#	**Επίθετο**	**Όνομα**	**Φορέας**	**Email Επικοινωνίας**
1	Πεζούλας	Βασίλειος	ΕΛΚΕ ΠΙ	bpezoulas@gmail.com
2	Πλέουρας	Δημήτριος	ΕΛΚΕ ΠΙ	dipleouras@gmail.com
3	Σιόγκας	Παναγιώτης	ΕΛΚΕ ΠΙ	psiogkas@uoi.gr
4	Γκόης	Γεώργιος	ΕΛΚΕ ΠΙ	gkois@yahoo.com
Συν-συγγραφείς				
#	**Επίθετο**	**Όνομα**	**Φορέας**	**Email Επικοινωνίας**
1	Κούρου	Κωνσταντίνα	ΕΛΚΕ ΠΙ	konstadina.kourou@gmail.com
2	Καλατζής	Θεοφάνης	ΕΛΚΕ ΠΙ	tkalatz@gmail.com
3	Φωτιάδης	Δημήτριος	ΕΛΚΕ ΠΙ	fotiadis@uoi.gr

ΛΙΣΤΑ ΚΡΙΤΩΝ

Κριτές				
#	**Επίθετο**	**Όνομα**	**Φορέας**	**Email Επικοινωνίας**
1	Μανόπουλος	Χρήστος	ΕΜΠ	manopoul@central.ntua.gr
2	Ράπτης	Αναστάσιος	ΕΜΠ	raptistasos@mail.ntua.gr
3	Μανόπουλος	Χρήστος	ΕΜΠ	manopoul@central.ntua.gr

ΕΛΕΓΧΟΣ ΑΝΑΘΕΩΡΗΣΗΣ

Έκδοση	**Συγγραφέας**	**Ημερομηνία**	**Κατάσταση**
0.1			Προσχέδιο
0.2			
0.3			
0.4			Τελικό

Περιεχόμενα

Πίνακας Σχημάτων

Πίνακας Πινάκων

Λίστα Συντομογραφιών

Συντομογραφία	Ορισμός (English)	Ελληνική Περιγραφή
AAA / AKA	Abdominal Aortic Aneurysm	Ανεύρυσμα Κοιλιακής Αορτής
AI	Artificial Intelligence	Τεχνητή Νοημοσύνη
API	Application Programming Interface	Διεπαφή Προγραμματισμού Εφαρμογών
CSV	Comma-Separated Values	Μορφότυπος αρχείων δεδομένων διαχωρισμένων με κόμμα
CIA	Confidentiality, Integrity, Availability	Τρεις θεμελιώδεις αρχές ασφάλειας πληροφοριών
DICOM	Digital Imaging and Communications in Medicine	Ψηφιακή Απεικόνιση και Επικοινωνία στην Ιατρική
EHDS	European Health Data Space	Ευρωπαϊκός Χώρος Δεδομένων Υγείας
FHIR	Fast Healthcare Interoperability Resources	Πρότυπο διαλειτουργικότητας δεδομένων υγείας
GDPR	General Data Protection Regulation	Γενικός Κανονισμός Προστασίας Δεδομένων (Κανονισμός 2016/679)
GUI	Graphical User Interface	Γραφικό Περιβάλλον Χρήστη
HL7	Health Level Seven	Διεθνές πρότυπο ανταλλαγής δεδομένων υγείας
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission	Διεθνής Οργανισμός Τυποποίησης και Ηλεκτροτεχνικής Επιτροπής
JSON	JavaScript Object Notation	Ελαφρύς μορφότυπος ανταλλαγής δεδομένων
LoC	Level of Compliance	Επίπεδο Συμμόρφωσης
ML	Machine Learning	Μηχανική Μάθηση
MySQL FEDERATED	MySQL Federated Storage Engine	Μηχανισμός ομοσπονδιακής βάσης δεδομένων MySQL
NIST	National Institute of Standards and Technology	Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας
PACS	Picture Archiving and Communication System	Σύστημα Αρχειοθέτησης και Επικοινωνίας Ιατρικών Εικόνων

PyDICOM	Python DICOM Library	Βιβλιοθήκη Python για επεξεργασία DICOM αρχείων
QA	Quality Assurance	Διασφάλιση Ποιότητας
RBAC	Role-Based Access Control	Έλεγχος πρόσβασης βάσει ρόλων
SQL	Structured Query Language	Δομημένη Γλώσσα Ερωτημάτων
TLS	Transport Layer Security	Πρωτόκολλο Ασφαλείας Μεταφοράς Δεδομένων
TPM	Trusted Platform Module	Μονάδα Αξιόπιστης Πλατφόρμας για ασφαλή αποθήκευση κλειδιών
UEFI	Unified Extensible Firmware Interface	Ενοποιημένη Επεκτάσιμη Διεπαφή Υλικολογισμικού
UUID	Universally Unique Identifier	Καθολικά Μοναδικός Αναγνωριστικός Αριθμός
VM	Virtual Machine	Εικονική Μηχανή
VPN	Virtual Private Network	Εικονικό Ιδιωτικό Δίκτυο
vSAN	Virtual Storage Area Network	Εικονικό Δίκτυο Αποθήκευσης
XML	Extensible Markup Language	Επεκτάσιμη Γλώσσα Σήμανσης
ΨηφιΔΑ	Ψηφιακός Δίδυμος Αορτής	Εξατομικευμένο υπολογιστικό μοντέλο της αορτής στο SAFEAORTA

Περίληψη

Το παρόν παραδοτέο τεκμηριώνει τη διαδικασία εναρμόνισης των ιατρικών δεδομένων του έργου **SAFEAORTA**, με έμφαση τόσο στις απεικονιστικές όσο και στις πινακοποιημένες πληροφορίες που συλλέγονται από τους συνεργαζόμενους φορείς. Η εναρμόνιση περιλαμβάνει τη μεθοδολογία προτυποποίησης, τον καθορισμό κοινών σχημάτων δεδομένων και τη διασφάλιση της διαλειτουργικότητας των πληροφοριών στο πλαίσιο της ομοσπονδιακής υποδομής που έχει ήδη αναπτυχθεί στο έργο. Παράλληλα, ενσωματώνει μηχανισμούς αυτοματοποιημένης ανωνυμοποίησης, ελέγχου ποιότητας, και αντιστοίχισης μεταδεδομένων σε κοινό πρότυπο, εξασφαλίζοντας τη συμμόρφωση με το κανονιστικό πλαίσιο του **GDPR** και τις τεχνικές προδιαγραφές ασφάλειας και διαλειτουργικότητας (DICOM, HL7/FHIR, ISO/IEC 27001).

Η ενοποιημένη αυτή προσέγγιση επιτρέπει την ασφαλή ενσωμάτωση και αξιοποίηση δεδομένων από πολλαπλές πηγές, υποστηρίζοντας τη δημιουργία αξιόπιστων βάσεων για ανάλυση, μοντελοποίηση και εκπαίδευση αλγορίθμων τεχνητής νοημοσύνης. Το αποτέλεσμα είναι ένα πλήρως τεκμηριωμένο, επεκτάσιμο και αναπαραγώγιμο πλαίσιο εναρμόνισης δεδομένων, θεμελιώδες για την ανάπτυξη του **Ψηφιακού Διδύμου της Αορτής (ΨηφιΔΑ)** και για την υλοποίηση της πλατφόρμας **SAFEAORTA**.

1. Executive summary

Το παρόν παραδοτέο αποτελεί το τρίτο και τελικό στάδιο του ΠΕ3, το οποίο επικεντρώνεται στην εναρμόνιση και προτυποποίηση των δεδομένων που συλλέγονται στο πλαίσιο του έργου **SAFEAORTA**. Ακολουθώντας τη σειρά των προηγούμενων παραδοτέων που αφορούσαν στην ανάπτυξη του εργαλείου ανωνυμοποίησης DICOM, και **την τεκμηρίωση** της ανάπτυξης της ομοσπονδιακής βάσης δεδομένων και της ασφαλούς υποδομής αποθήκευσης — η παρούσα φάση εστιάζει στη δημιουργία των μηχανισμών ομογενοποίησης και διαλειτουργικότητας των δεδομένων.

Η εναρμόνιση των δεδομένων περιλαμβάνει δύο συμπληρωματικές διαστάσεις:

1. **Εναρμόνιση ιατρικών εικόνων και μεταδεδομένων DICOM**, μέσω διαδικασιών που εξασφαλίζουν την τυποποίηση των πεδίων, τη συνεπή απεικόνιση των ανατομικών περιοχών και την ευθυγράμμιση των συνόλων δεδομένων με διεθνή πρότυπα. Οι εικόνες που προέρχονται από διαφορετικά μηχανήματα, πρωτόκολλα ή ιδρύματα υπόκεινται σε διαδικασίες κανονικοποίησης, ελέγχου ποιότητας και αντιστοίχισης σταθερών ετικετών (UIDs, Modality, Series/Study IDs).

2. **Εναρμόνιση πινακοποιημένων δεδομένων (tabular data)**, όπως δημογραφικά, αιματολογικά, κλινικά και βιοχημικά στοιχεία, μέσω καθορισμού κοινών λεξιλογίων, μορφότυπων και μεταδεδομένων. Η διαδικασία περιλαμβάνει τον εντοπισμό αποκλίσεων, τη μετατροπή μονάδων, την κατηγοριοποίηση μεταβλητών και τη σύνδεση με τα αντίστοιχα identifiers των απεικονιστικών δεδομένων, εξασφαλίζοντας συνοχή στο σύνολο του οικοσυστήματος δεδομένων.

Η μεθοδολογία που ακολουθείται βασίζεται στη συνδυασμένη χρήση εργαλείων **Python** (pandas, pydicom, openpyxl, numpy), προτύπων **FHIR** και **HL7**, καθώς και μοντέλων μεταδεδομένων που καθορίζουν τον ενιαίο χάρτη των δεδομένων προς ένταξη στην ομοσπονδιακή βάση. Κάθε αρχείο δεδομένων υποβάλλεται σε βήματα ελέγχου, συμπεριλαμβανομένων των **consistency checks**, **format validation**, και **semantic mapping**.

Η τελική έξοδος της διαδικασίας είναι ένα **ομοιογενές, εναρμονισμένο dataset**, το οποίο μπορεί να ενσωματωθεί στην υποδομή του **SAFEAORTA Federated Cloud**, επιτρέποντας ασφαλή πρόσβαση μέσω εικονικών μηχανών (VMs), VPN/TLS, και ελεγχόμενων μηχανισμών RBAC. Η συνολική διαδικασία εναρμόνισης θέτει τις βάσεις για την ανάπτυξη προγνωστικών μοντέλων τεχνητής νοημοσύνης και για τη μελλοντική χρήση των δεδομένων σε έργα **federated learning**.

2. Εισαγωγή

2.1 Σκοπός και στόχοι του παραδοτέου

Ο σκοπός του παρόντος παραδοτέου είναι η τεκμηρίωση της διαδικασίας εναρμόνισης των ετερογενών δεδομένων που παράγονται και συγκεντρώνονται από τους φορείς του έργου SAFEAORTA, στο πλαίσιο της ανάπτυξης μιας ολοκληρωμένης και ασφαλούς υποδομής διαχείρισης ιατρικών πληροφοριών.

Η εναρμόνιση αποσκοπεί στην εξάλειψη των διαφορών μεταξύ διαφορετικών πηγών δεδομένων, εξασφαλίζοντας κοινή δομή, ενιαία ονοματολογία, σταθερή μονάδα μέτρησης και πλήρη ιχνηλασιμότητα.

Οι επιμέρους στόχοι περιλαμβάνουν:

- Την καθιέρωση κοινών προτύπων μορφοποίησης και μεταδεδομένων για τα DICOM και tabular datasets.
- Την ανάπτυξη αυτοματοποιημένων διαδικασιών ελέγχου ποιότητας, κανονικοποίησης και μετατροπής.
- Την ενσωμάτωση των εναρμονισμένων δεδομένων στην ομοσπονδιακή βάση του έργου.
- Τη διασφάλιση της συμμόρφωσης με τον GDPR, τις προδιαγραφές ασφαλείας του ISO/IEC 27001, και τα διεθνή πρότυπα FHIR/HL7.

Με την ολοκλήρωση του παραδοτέου, το σύνολο των δεδομένων καθίσταται έτοιμο για χρήση στις επόμενες ενότητες του έργου, υποστηρίζοντας αναλύσεις, υπολογιστικά μοντέλα και προγνωστικούς αλγορίθμους με διαλειτουργικό και ασφαλή τρόπο.

2.2 Σχέση με τα υπόλοιπα παραδοτέα του Π.Ε.3

Το παρόν παραδοτέο **D3.3 (Εναρμόνιση)** αποτελεί φυσική συνέχεια των προηγούμενων φάσεων του ΠΕ3. Συγκεκριμένα:

- Το «**D3.3- Εργαλείο ψευδοταυτοποίησης ή ανωνυμοποίησης ιατρικών δεδομένων»** εισήγαγε το **εργαλείο ανωνυμοποίησης DICOM**, το οποίο εξασφαλίζει την προστασία των προσωπικών δεδομένων πριν τη διαδικασία εναρμόνισης.
- Το «**D3.3- Ανάπτυξη ομοσπονδιακών βάσεων δεδομένων καθώς και η παροχή εξουσιοδοτημένης πρόσβασης σε αυτές»** ανέπτυξε την **τεχνική υποδομή των ομοσπονδιακών βάσεων δεδομένων**, προσφέροντας ασφαλή και αποκεντρωμένη αποθήκευση και πρόσβαση.

- Το **παρόν D3.3 (Εναρμόνιση)** έρχεται να συμπληρώσει τη λειτουργικότητα της υποδομής, εξασφαλίζοντας την **ποιοτική ομογενοποίηση** και **διαλειτουργικότητα** των δεδομένων που θα εισαχθούν σε αυτήν.

Έτσι, το σύνολο των παραδοτέων του ΠΕ3 συγκροτεί μια πλήρη και συνεχόμενη αλυσίδα αξίας:
Ανωνυμοποίηση → Εναρμόνιση → Ασφαλής Αποθήκευση → Ομοσπονδιακή Πρόσβαση → Ανάλυση και Μοντελοποίηση.

2.3 Σύνδεση με τους συνολικούς στόχους του SAFEAORTA

Η παρούσα εργασία συνδέεται άμεσα με τους γενικούς στόχους του έργου SAFEAORTA, που περιλαμβάνουν την ανάπτυξη μιας **διαλειτουργικής, ασφαλούς και προγνωστικής πλατφόρμας** για τη διαχείριση της νόσου των **Ανευρυσμάτων Κοιλιακής Αορτής (ΑΚΑ)**. Η εναρμόνιση των δεδομένων αποτελεί θεμελιώδες στάδιο για:

- Τη δημιουργία **ενοποιημένων βάσεων δεδομένων** υψηλής ακρίβειας και αξιοπιστίας.
- Την υποστήριξη της **ανάπτυξης Ψηφιακών Διδύμων της Αορτής (ΨηφιΔΑ)**.
- Την εκπαίδευση **μοντέλων τεχνητής νοημοσύνης** σε πραγματικά, ομοιογενή δεδομένα.
- Την ενίσχυση της **αξιοπιστίας των αποτελεσμάτων** και της **προγνωστικής ακρίβειας** των αλγορίθμων ρήξης.

Η συμβολή του παραδοτέου εστιάζει στην **ποιοτική και σημασιολογική ευθυγράμμιση των δεδομένων**, ώστε η πλατφόρμα SAFEAORTA να λειτουργεί ως ολοκληρωμένο και αξιόπιστο **Σύστημα Υποστήριξης Κλινικών Αποφάσεων**, με πλήρη συμμόρφωση προς τα ευρωπαϊκά πρότυπα και τους κανονισμούς προστασίας δεδομένων.

3. Μεθοδολογία και Εργαλεία για την Εναρμόνιση Ιατρικών Εικόνων

Η εναρμόνιση ιατρικών εικόνων αποτελεί κρίσιμη διαδικασία για την αντιμετώπιση της ετερογένειας που προκύπτει από διαφορετικούς σαρωτές, πρωτόκολλα λήψης και παραμέτρους απεικόνισης. Η ύπαρξη αυτής της ετερογένειας οδηγεί σε μη συγκρίσιμες κατανομές έντασης και διαφορετική στατιστική συμπεριφορά των απεικονιστικών χαρακτηριστικών, γεγονός που μειώνει τη δυνατότητα συνδυαστικής ανάλυσης δεδομένων από πολλαπλές πηγές και την αναπαραγωγιμότητα των αποτελεσμάτων.

Για την αντιμετώπιση αυτών των προβλημάτων, αναπτύχθηκε μια ολοκληρωμένη υπολογιστική ροή (image harmonization pipeline), η οποία εξασφαλίζει την ομοιομορφία των εικόνων πριν τη στατιστική και ραδιομική ανάλυση. Η ροή αυτή περιλαμβάνει στάδια κανονικοποίησης έντασης, αντιστοίχισης ιστογραμμάτων, διόρθωσης πεδίου προτίμησης, αυτόματης απομάκρυνσης artifacts, εφαρμογής αλγορίθμων εναρμόνισης και αξιολόγησης ποσοτικών δεικτών ποιότητας.

Η προτεινόμενη μεθοδολογία εφαρμόστηκε εξ ολοκλήρου σε περιβάλλον Python, αξιοποιώντας βιβλιοθήκες όπως pydicom, NumPy, OpenCV, SimpleITK, PyRadiomics και neuroCombat-sklearn. Επιπλέον, ενσωματώθηκε σε διαδραστική εφαρμογή (Qt GUI) για την αυτόματη φόρτωση φακέλων DICOM, την εκτέλεση της ροής επεξεργασίας και την αποθήκευση των αποτελεσμάτων ανά ασθενή.

Η συνολική διαδικασία αποδίδει, για κάθε μεμονωμένο δείγμα, σύνολα ογκομετρικών εικόνων (π.χ. 01_original.nii.gz, 02_equalized.nii.gz, 03_bias_corrected.nii.gz) και συνοδευτικά αρχεία αξιολόγησης (qa_metrics.csv, radiomics.csv), τα οποία χρησιμοποιούνται τόσο για την ποιοτική σύγκριση όσο και για την περαιτέρω εναρμόνιση των εξαγόμενων χαρακτηριστικών.

3.1 Κανονικοποίηση έντασης και αντιστοίχιση ιστογραμμάτων

Η κανονικοποίηση έντασης και η αντιστοίχιση ιστογραμμάτων αποτελούν το πρώτο και πιο κρίσιμο στάδιο της διαδικασίας εναρμόνισης ιατρικών εικόνων. Οι εικόνες που προέρχονται από διαφορετικούς σαρωτές ή πρωτόκολλα απεικόνισης παρουσιάζουν σημαντικές διαφορές στην κλίμακα και στην κατανομή των τιμών έντασης, ακόμη και όταν απεικονίζουν το ίδιο ανατομικό όργανο ή ιστό. Οι αποκλίσεις αυτές οφείλονται σε ποικίλους τεχνικούς παράγοντες, όπως το reconstruction kernel, το voxel spacing, η ρύθμιση του contrast enhancement ή η μεταβολή των παραμέτρων ενέργειας του σαρωτή. Ως αποτέλεσμα, οι κατανομές έντασης των voxel καθίστανται μη συγκρίσιμες, επηρεάζοντας τη σταθερότητα των ραδιομικών χαρακτηριστικών και μειώνοντας την αξιοπιστία των μοντέλων πρόβλεψης ή ταξινόμησης που βασίζονται σε αυτά.

Για την αντιμετώπιση των διαφορών αυτών, εφαρμόστηκε μια τριπλή διαδικασία προεπεξεργασίας, η οποία περιλαμβάνει την αποκοπή ακραίων τιμών (percentile clipping), την κανονικοποίηση τύπου Z-score και την προσαρμοστική ισοστάθμιση ιστογράμματος (Contrast Limited Adaptive Histogram Equalization – CLAHE). Η διαδικασία αυτή εκτελείται αυτόματα για κάθε τρισδιάστατο σύνολο εικόνων (DICOM series), παράγοντας ως αποτέλεσμα ένα νέο εναρμονισμένο όγκο δεδομένων σε μορφή NIfTI (02_equalized.nii.gz), ο οποίος αποτελεί τη βάση για τα επόμενα στάδια της ροής.

Εικόνα 5. Αρχική ιατρική εικόνα (raw DICOM) πριν την εναρμόνιση. Εμφανίζονται διαφορές φωτεινότητας και contrast λόγω σαρωτή και παραμέτρων λήψης.

Αρχικά, εφαρμόζεται αποκοπή ακραίων τιμών στο εύρος του 1ου και 99ου εκατοστημορίου της κατανομής έντασης. Το βήμα αυτό περιορίζει την επίδραση θορύβου ή σποραδικών τιμών που προκύπτουν από μεταλλικά artifacts ή λάθη στην ανακατασκευή της εικόνας.

Μαθηματικά, αν p_1και p_{99}είναι οι τιμές των εκατοστημορίων, τότε για κάθε voxel xμε τιμή έντασης $I(x)$, η διορθωμένη τιμή $I_{\text{clip}}(x)$ορίζεται ως:

$$I_{\text{clip}}(x) = \min\left(\max\left(I(x), p_1\right), p_{99}\right)$$

Η διαδικασία αυτή εξασφαλίζει ότι όλες οι εντάσεις παραμένουν εντός ενός σταθερού και συγκρίσιμου δυναμικού εύρους, αποτρέποντας τη μεροληψία από ακραίες τιμές που δεν έχουν φυσιολογική ή ανατομική σημασία.

Ακολούθως, πραγματοποιείται κανονικοποίηση τύπου Z-score, η οποία μετασχηματίζει τις τιμές έντασης έτσι ώστε η κατανομή τους να αποκτά μηδενικό μέσο όρο και μοναδιαία τυπική απόκλιση. Για κάθε voxel, η νέα τιμή έντασης $I_{\text{norm}}(x)$υπολογίζεται ως:

$$I_{\text{norm}}(x) = \frac{I_{\text{clip}}(x) - \mu}{\sigma}$$

όπου μ είναι η μέση τιμή και ση τυπική απόκλιση των εντάσεων του όγκου. Με τον τρόπο αυτό εξασφαλίζεται ότι διαφορετικές εικόνες έχουν κοινό κέντρο και κλίμακα αναφοράς, γεγονός που επιτρέπει τη συνεπή ανάλυση και τη σταθερότερη εξαγωγή χαρακτηριστικών. Σε περιπτώσεις όπου χρησιμοποιείται μάσκα σώματος (body mask), ο υπολογισμός των στατιστικών πραγματοποιείται εντός της μάσκας, ώστε να αποκλειστούν περιοχές αέρα ή μη σχετικού υποβάθρου.

Το τρίτο στάδιο αφορά την προσαρμοστική ισοστάθμιση ιστογράμματος (CLAHE), η οποία εφαρμόζεται ανεξάρτητα σε κάθε τομή της εικόνας. Σε αντίθεση με την απλή (global) ισοστάθμιση, η CLAHE λειτουργεί σε μικρά, τοπικά παράθυρα (π.χ. 8×8 tiles) και περιορίζει την υπερ-ενίσχυση του κοντράστ μέσω ενός παραμέτρου clip limit. Η μέθοδος αυτή αυξάνει το τοπικό κοντράστ και αναδεικνύει λεπτομέρειες σε ιστούς με χαμηλή διακριτότητα, χωρίς να ενισχύει υπερβολικά τον θόρυβο. Στην προτεινόμενη υλοποίηση, η CLAHE εφαρμόζεται εντός του ROI του σώματος, αποφεύγοντας την εφαρμογή της σε περιοχές εκτός της ανατομικής δομής (π.χ. αέρας ή background). Το στάδιο αυτό αντιστοιχεί λειτουργικά σε μια ελαφριά μορφή RACLAHE (Region Adaptive CLAHE), όπου η ενίσχυση περιορίζεται στις περιοχές ενδιαφέροντος[1].

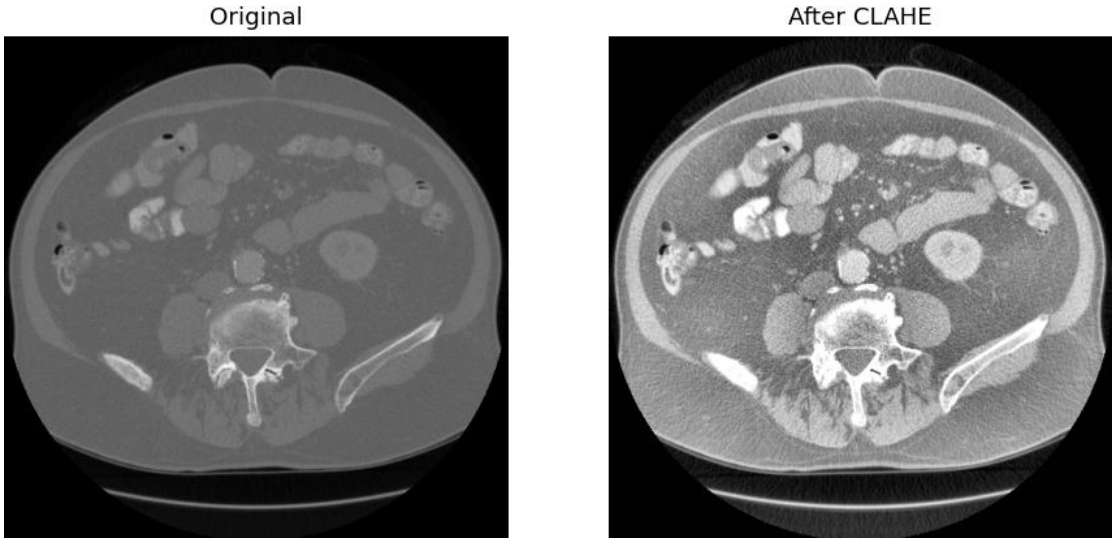
Η συγκεκριμένη διαδικασία υλοποιήθηκε σε περιβάλλον Python με χρήση των βιβλιοθηκών NumPy και OpenCV. Για κάθε τομή εφαρμόζονται διαδοχικά τα βήματα clipping,

[1] Zuiderveld, K., “Contrast Limited Adaptive Histogram Equalization,” *Graphics Gems IV*, 1994, pp. 474–485

normalization και equalization, ενώ τα αποτελέσματα συνενώνονται σε έναν τρισδιάστατο όγκο και αποθηκεύονται ως αρχείο NIfTI για συμβατότητα με τις επόμενες φάσεις (bias field correction, artifact removal κ.λπ.). Η διαδικασία είναι πλήρως αυτοματοποιημένη μέσα στην αναπτυγμένη Qt εφαρμογή, επιτρέποντας στον χρήστη να εισάγει τον φάκελο DICOM και να λάβει αυτόματα το εξισορροπημένο αποτέλεσμα.

Η εφαρμογή της μεθοδολογίας αυτής οδηγεί σε σημαντική μείωση της τεχνικής διακύμανσης των τιμών έντασης και βελτίωση της αναλογίας σήματος προς θόρυβο (SNR), ενώ ταυτόχρονα ενισχύει τις δομικές λεπτομέρειες χωρίς να αλλοιώνει τη γεωμετρία των ιστών. Ποσοτικά, παρατηρείται μείωση του συντελεστή μεταβλητότητας (Coefficient of Variation, CoV) και σταθεροποίηση των δεικτών εντροπίας εντός της μάσκας σώματος. Οπτικά, η εικόνα εμφανίζεται με πιο ισορροπημένο κοντράστ και καλύτερη οριοθέτηση των ανατομικών ορίων, όπως παρουσιάζεται στην Εικόνα 6 (σύγκριση πριν και μετά τη διαδικασία CLAHE.

Συνολικά, η κανονικοποίηση έντασης και η ισοστάθμιση ιστογράμματος αποτελούν το θεμέλιο της διαδικασίας εναρμόνισης, καθώς εξαλείφουν τις διαφορές φωτεινότητας και κοντράστ που προκαλούνται από τεχνικούς παράγοντες και διασφαλίζουν τη σταθερότητα των στατιστικών ιδιοτήτων των εικόνων. Με αυτό τον τρόπο, οι εικόνες που προέρχονται από διαφορετικές πηγές καθίστανται συγκρίσιμες, γεγονός που επιτρέπει τη συνεπή ανάλυση και την αξιόπιστη εξαγωγή βιοδεικτών στο επόμενο στάδιο της επεξεργασίας.



Εικόνα 6. Σύγκριση της αρχικής τομής (αριστερά) και της κανονικοποιημένης με CLAHE (δεξιά). Παρατηρείται ενίσχυση της τοπικής αντίθεσης και βελτίωση της ορατότητας αγγειακών δομών.

3.2 Διόρθωση πεδίου προτίμησης (Bias Field Correction)

Η διόρθωση του πεδίου προτίμησης (Bias Field Correction) αποτελεί το δεύτερο στάδιο της διαδικασίας εναρμόνισης των ιατρικών εικόνων και έχει ως στόχο την αντιμετώπιση χαμηλής

συχνότητας ανομοιομορφιών φωτεινότητας που επηρεάζουν τις τιμές έντασης των voxel. Το φαινόμενο αυτό, γνωστό ως bias field ή intensity inhomogeneity, οφείλεται σε παράγοντες όπως η μη ομοιογενής κατανομή του μαγνητικού πεδίου (ιδίως στις MRI), οι μεταβολές στην απόκριση των ανιχνευτών ή η γεωμετρική διάταξη του σαρωτή. Ακόμη και σε CT εικόνες, παρατηρούνται ήπιες μορφές bias που σχετίζονται με τη σκλήρυνση της δέσμης (beam hardening) ή τη γεωμετρική απόσταση από την πηγή ακτινοβολίας.

Οι ανομοιομορφίες αυτές οδηγούν σε σταδιακή μεταβολή της έντασης σε περιοχές που θα έπρεπε να έχουν ομοιόμορφες τιμές, προκαλώντας μετατόπιση των ιστογραμμάτων, διαστρέβλωση των πρώτης τάξης χαρακτηριστικών (όπως mean ή standard deviation) και τελικά μείωση της ακρίβειας κατά την εξαγωγή ραδιομικών δεικτών. Για τον λόγο αυτό, η διόρθωση του bias field αποτελεί κρίσιμο βήμα πριν από οποιαδήποτε ποσοτική ανάλυση ή διαδικασία segmentation.

Η μέθοδος που χρησιμοποιήθηκε στο παρόν έργο βασίζεται στον αλγόριθμο N4ITK Bias Field Correction, μια μη παραμετρική τεχνική εκτίμησης και αφαίρεσης του bias field που αποτελεί εξέλιξη του κλασικού N3 (Nonparametric Nonuniform intensity Normalization)[2]. Η μέθοδος N4ITK υποθέτει ότι η παρατηρούμενη ένταση $I(x)$μπορεί να αναπαρασταθεί ως γινόμενο δύο συνιστωσών:

$$I(x) = B(x) \cdot J(x)$$

όπου $J(x)$είναι η πραγματική (bias-free) ένταση του voxel και $B(x)$ένας αργά μεταβαλλόμενος όρος bias field που μεταβάλλει ομαλά τη φωτεινότητα στο χώρο της εικόνας. Λαμβάνοντας τον λογάριθμο, το πρόβλημα μετατρέπεται σε άθροισμα:

$$\log I(x) = \log J(x) + \log B(x)$$

και η εκτίμηση του $\log B(x)$πραγματοποιείται μέσω επαναληπτικής εξομάλυνσης του λογαριθμικού σήματος με φίλτρο τύπου B-spline, ώστε να προσεγγιστεί ο όρος της χαμηλής συχνότητας. Μετά την εκτίμηση του bias, η διόρθωση επιτυγχάνεται με διαίρεση:

$$J(x) = \frac{I(x)}{B(x)}$$

[2] Tustison N. J., Avants B. B., et al., "N4ITK: Improved N3 Bias Correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, 2010, pp. 1310–1320

παρέχοντας έτσι μια bias-free εικόνα με ομοιόμορφη κατανομή φωτεινότητας.

Στην υλοποίηση που αναπτύχθηκε, η διόρθωση πραγματοποιείται με χρήση της βιβλιοθήκης SimpleITK, η οποία παρέχει ολοκληρωμένη υλοποίηση του N4ITK. Η διαδικασία εκτελείται σε μειωμένη ανάλυση (downsampled space) για επιτάχυνση, ενώ η τελική εκτίμηση εφαρμόζεται εκ νέου στο πλήρες μέγεθος της εικόνας. Για την αυτόματη οριοθέτηση των περιοχών ενδιαφέροντος, χρησιμοποιείται μάσκα σώματος που παράγεται μέσω Otsu thresholding, ώστε η διόρθωση να εφαρμοστεί μόνο εντός των ανατομικών δομών και να αγνοηθούν περιοχές αέρα ή υποβάθρου που ενδέχεται να επηρεάσουν αρνητικά τη σύγκλιση του αλγορίθμου.

Η επιλογή των παραμέτρων επαναλήψεων του N4ITK έγινε με στόχο την επίτευξη καλής ισορροπίας μεταξύ ακρίβειας και χρόνου υπολογισμού. Συγκεκριμένα, χρησιμοποιήθηκαν τέσσερα στάδια επαναλήψεων [30,30,20,10], τα οποία αποδείχθηκαν επαρκή για τη σταθερή σύγκλιση του bias field σε 3D όγκους με 150–200 τομές. Για κάθε επανάληψη, ο αλγόριθμος υπολογίζει το bias σε διαφορετική κλίμακα (multi-resolution approach), εξομαλύνοντας σταδιακά τα χαμηλής συχνότητας μοτίβα φωτεινότητας. Η έξοδος αποθηκεύεται σε νέο αρχείο 03_bias_corrected.nii.gz, το οποίο φέρει πλέον ομοιόμορφη κατανομή έντασης και μειωμένες διαφορές φωτεινότητας μεταξύ διαδοχικών τομών.

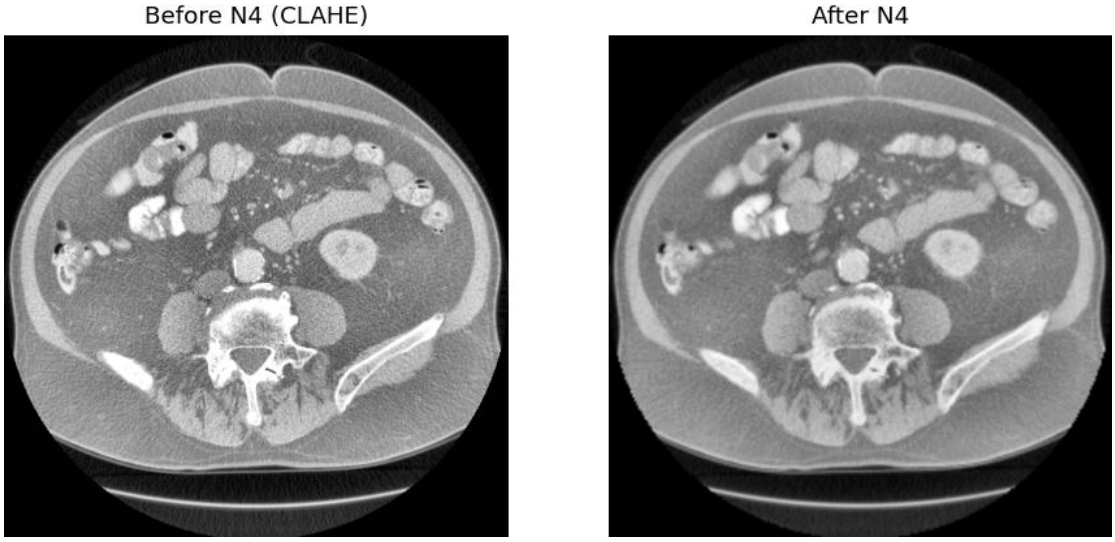
Η επίδραση της διόρθωσης του bias field είναι εμφανής τόσο οπτικά όσο και ποσοτικά. Οπτικά, η διόρθωση οδηγεί σε εικόνες με πιο ομοιογενές υπόβαθρο και εξομαλυμένες τιμές έντασης στις περιοχές που προηγουμένως παρουσίαζαν σταδιακή διαφορά φωτεινότητας, όπως απεικονίζεται στην Εικόνα 7 (σύγκριση πριν και μετά τη διόρθωση). Ποσοτικά, η αξιολόγηση εντός της μάσκας σώματος δείχνει σημαντική μείωση του συντελεστή μεταβλητότητας (CoV) και σταθεροποίηση της εντροπίας της κατανομής έντασης.

Αξίζει να σημειωθεί ότι η διαδικασία διόρθωσης bias field εφαρμόζεται μετά την κανονικοποίηση έντασης και πριν από τη φάση αφαίρεσης artifacts (§3.3), ώστε να εξασφαλιστεί ότι το bias εκτιμάται σε ήδη ομαλοποιημένο και “καθαρό” φάσμα εντάσεων. Σε περιπτώσεις μεγάλου αριθμού τομών, η εκτέλεση της διαδικασίας μπορεί να διαρκέσει έως και 5–8 λεπτά, ανάλογα με τα χαρακτηριστικά του συστήματος. Για τον λόγο αυτό, αναπτύχθηκε και μια fast N4 παραλλαγή, στην οποία μειώνεται η πυκνότητα του πλέγματος B-spline και ο αριθμός επαναλήψεων, προσφέροντας σημαντική επιτάχυνση με ελάχιστη απώλεια ποιότητας.

Η εφαρμογή του σταδίου αυτού είναι καθοριστική για την ποιοτική εναρμόνιση των εικόνων και τη βελτίωση της αξιοπιστίας των εξαγόμενων ραδιομικών χαρακτηριστικών. Μετά τη διόρθωση του bias field, τα δεδομένα είναι πλέον απαλλαγμένα από συστηματικές μεταβολές

φωτεινότητας που δεν σχετίζονται με φυσιολογικές διαφορές των ιστών, γεγονός που επιτρέπει τη συγκρίσιμη ανάλυση μεταξύ ασθενών και την ακριβέστερη εξαγωγή δεικτών απόδοσης στις επόμενες φάσεις της ανάλυσης.

Η συνολική βελτίωση τεκμηριώνεται με δείκτες που υπολογίζονται αυτόματα στο στάδιο ελέγχου ποιότητας (QA module, §3.6), όπου παρατηρείται μείωση του CoV και αύξηση του δείκτη SSIM (Structural Similarity Index Measure) μεταξύ των αρχικών και διορθωμένων τομών. Έτσι, η διόρθωση του πεδίου προτίμησης αποτελεί θεμελιώδες βήμα για την απομάκρυνση των τεχνικών ανισοτροπιών και την εδραίωση μιας αξιόπιστης βάσης για τη στατιστική και ραδιομική εναρμόνιση που ακολουθεί.



Εικόνα 7. Εφαρμογή N4ITK Bias Field Correction. Η ένταση φωτεινότητας εξομαλύνεται σε όλο τον όγκο, μειώνοντας την επίδραση του μαγνητικού πεδίου και βελτιώνοντας την ομοιογένεια των ιστών.

3.3 Αυτόματη επεξεργασία και αφαίρεση artifacts

Η παρουσία artifacts στις ιατρικές εικόνες συνιστά μια από τις σημαντικότερες πηγές αλλοίωσης της ποιότητας των δεδομένων και μπορεί να οδηγήσει σε σημαντικά σφάλματα κατά τη φασματική ανάλυση ή την εξαγωγή ραδιομικών χαρακτηριστικών. Τα artifacts προέρχονται από ποικίλες πηγές — μεταλλικά εμφυτεύματα, κίνηση του ασθενούς, περιορισμούς στη διαδικασία ανακατασκευής ή ακόμη και σφάλματα στην εκπομπή και απορρόφηση της ακτινοβολίας. Στην περίπτωση των υπολογιστικών τομογραφιών (CT), τα πιο συχνά είδη artifacts είναι τα metal streak artifacts (γραμμικές εντάσεις γύρω από μεταλλικά αντικείμενα), τα beam hardening artifacts και οι τοπικές παραμορφώσεις έντασης κοντά σε υψηλής πυκνότητας περιοχές.

Η απομάκρυνση των artifacts αποτελεί κρίσιμο βήμα, καθώς η ύπαρξή τους μπορεί να αλλοιώσει δραστικά τόσο τα πρώτης τάξης στατιστικά χαρακτηριστικά (π.χ. mean, skewness)

όσο και τα υφιστάμενα texture features (π.χ. GLCM, GLSZM). Η παρουσία μιας ή περισσότερων φωτεινών ακτινικών γραμμών μπορεί, για παράδειγμα, να αυξήσει τεχνητά την εντροπία της εικόνας ή να προκαλέσει έντονη ανομοιομορφία στις κατανομές έντασης. Για τον λόγο αυτό, υλοποιήθηκε ένα αυτόματο σύστημα προεπεξεργασίας για τον εντοπισμό και την αποκατάσταση περιοχών με artifacts, εξασφαλίζοντας ότι οι διορθωμένες εικόνες μπορούν να χρησιμοποιηθούν με ασφάλεια στις επόμενες φάσεις της εναρμόνισης και της ραδιομικής ανάλυσης.

Η μεθοδολογία που εφαρμόστηκε περιλαμβάνει τέσσερα βασικά στάδια: **(α)** δημιουργία μάσκας σώματος (body mask), **(β)** εντοπισμό περιοχών με artifacts, **(γ)** αποκατάσταση των περιοχών αυτών με inpainting και **(δ)** τελική αποθορυβοποίηση και ενίσχυση εντός περιοχών ενδιαφέροντος (ROI)[3].

Αρχικά, παράγεται μια μάσκα σώματος (body mask) η οποία περιορίζει την επεξεργασία μόνο εντός του ανατομικού χώρου του σώματος, αποκλείοντας το υπόβαθρο. Η μάσκα αυτή δημιουργείται εφαρμόζοντας ένα κατώφλι έντασης (π.χ. −600 Hounsfield Units), το οποίο επιτρέπει την αποκοπή των περιοχών αέρα ή χαμηλής πυκνότητας. Το αποτέλεσμα υποβάλλεται σε μορφολογικές πράξεις κλεισίματος και πλήρωσης οπών (binary closing, fill holes) σε κάθε τομή, ώστε να δημιουργηθεί ένα συμπαγές και συνεχές ROI. Μαθηματικά, για κάθε τομή $I_z(x, y)$, η μάσκα σώματος $M_z(x, y)$υπολογίζεται ως:

$$M_z(x, y) = \text{fillholes}(\text{close}(I_z(x, y) > T))$$

όπου T το κατώφλι έντασης.

Στη συνέχεια, πραγματοποιείται εντοπισμός περιοχών με artifacts. Για τα CT δεδομένα, τα μεταλλικά artifacts εντοπίζονται εύκολα λόγω των πολύ υψηλών τιμών έντασης που υπερβαίνουν τα 3.000 HU, σε συνδυασμό με τη χαρακτηριστική ακτινική τους διάταξη. Οι περιοχές αυτές ανιχνεύονται μέσω ενός δεύτερου κατωφλίου και επεκτείνονται με μορφολογική διάταση (dilation), ώστε να συμπεριλάβουν το σύνολο των αλλοιωμένων voxel. Το αποτέλεσμα είναι μια binary μάσκα $A(x)$που υποδεικνύει τις περιοχές προς αποκατάσταση.

Για την αποκατάσταση των περιοχών με artifacts, εφαρμόζεται τεχνική inpainting βασισμένη στη βιβλιοθήκη *OpenCV* (αλγόριθμος Navier-Stokes). Η τεχνική αυτή αξιοποιεί τις πληροφορίες των γειτονικών περιοχών ώστε να "γεμίσει" τις περιοχές που επηρεάστηκαν από

[3] Buades, A., Coll, B., & Morel, J. M., "A Non-Local Algorithm for Image Denoising," *IEEE CVPR*, 2005

artifacts, αναδομώντας σταδιακά την τοπική δομή και την υφή των ιστών. Αν I είναι η εικόνα εισόδου και A η binary μάσκα των artifacts, η αποκατεστημένη εικόνα I_{rec}υπολογίζεται ως:

$$I_{\text{rec}} = \text{Inpaint}(I, A, r)$$

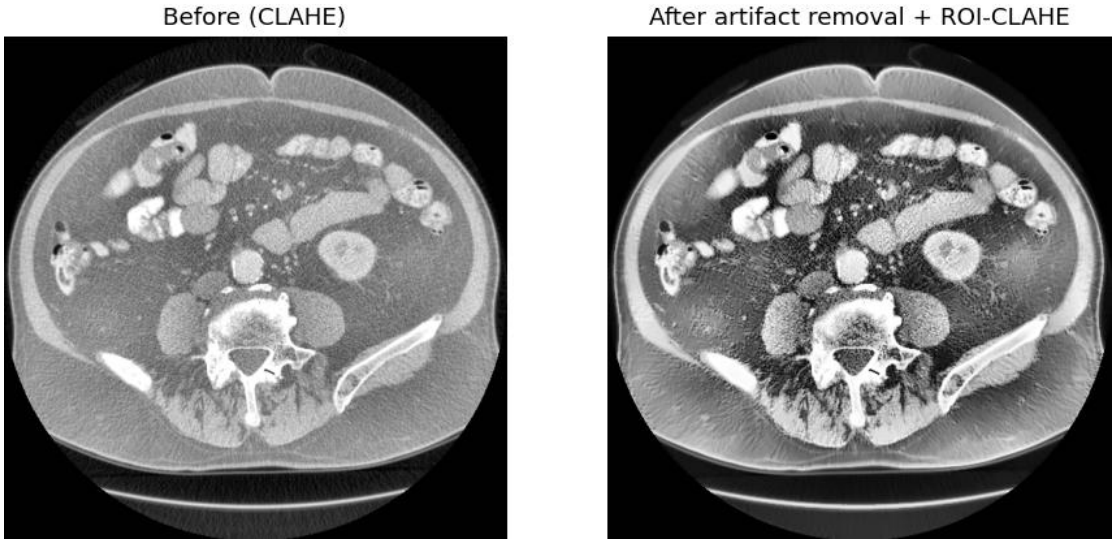
όπου r είναι η ακτίνα γειτονιάς (τυπικά $r = 3\text{voxels}$). Το αποτέλεσμα είναι μια εικόνα όπου οι αλλοιωμένες περιοχές έχουν αντικατασταθεί με στατιστικά συνεπείς τιμές, διατηρώντας τη συνοχή της δομής.

Ακολουθεί ένα στάδιο αποθορυβοποίησης και ενίσχυσης εντός ROI, το οποίο συνδυάζει μη γραμμικά φίλτρα και τοπική ενίσχυση κοντράστ. Συγκεκριμένα, εφαρμόζεται φίλτρο Non-Local Means σε κάθε τομή για την απομάκρυνση μικρής κλίμακας θορύβου, χωρίς απώλεια χωρικής λεπτομέρειας, και στη συνέχεια επαναλαμβάνεται προσαρμοστική ενίσχυση τύπου CLAHE, περιορισμένη όμως μόνο εντός της μάσκας σώματος. Με αυτόν τον τρόπο, ενισχύονται οι κρίσιμες περιοχές (π.χ. αγγειακοί αυλοί, δομές μαλακών ιστών), ενώ αποφεύγεται η υπερ-ενίσχυση του φόντου ή των περιοχών εκτός σώματος.

Η συνολική διαδικασία εκτελείται αυτόματα για κάθε σύνολο εικόνων και παράγει ως αποτέλεσμα ένα νέο volume με την ονομασία `04_artifact_clean.nii.gz`. Στο στάδιο αυτό επιτυγχάνεται ουσιαστική βελτίωση της καθαρότητας των εικόνων και της οπτικής ποιότητας, καθιστώντας τα δεδομένα κατάλληλα για ποσοτική ανάλυση και εξαγωγή ραδιομικών χαρακτηριστικών.

Η αποτελεσματικότητα της διαδικασίας είναι εμφανής τόσο οπτικά όσο και ποσοτικά. Οπτικά, οι ακτινικές γραμμές και τα σημεία υπερβολικής φωτεινότητας μειώνονται ή εξαφανίζονται πλήρως, ενώ η συνολική ομοιογένεια του ιστού αποκαθίσταται (βλ. Εικόνα 8). Ποσοτικά, η αξιολόγηση εντός της μάσκας σώματος δείχνει περαιτέρω μείωση του Coefficient of Variation (CoV) και αύξηση του δείκτη SSIM σε σχέση με τα προηγούμενα στάδια, επιβεβαιώνοντας ότι η απομάκρυνση των artifacts οδηγεί σε στατιστικά σταθερότερες και πιο ομοιογενείς εικόνες.

Η διαδικασία αφαίρεσης artifacts αποτελεί καθοριστικό ενδιάμεσο στάδιο της εναρμόνισης, καθώς διασφαλίζει ότι οι υπόλοιπες μέθοδοι, όπως η εναρμόνιση χαρακτηριστικών (ComBat) ή η ενίσχυση μέσω τεχνητής νοημοσύνης, θα εφαρμοστούν σε δεδομένα υψηλής ποιότητας, απαλλαγμένα από συστηματικές αλλοιώσεις. Με αυτόν τον τρόπο, η φάση 3.3 συμβάλλει ουσιαστικά στη βελτίωση της αξιοπιστίας της συνολικής ροής και στην παραγωγή εναρμονισμένων εικόνων που αντανακλούν πιστότερα την πραγματική φυσιολογία των ιστών.



Εικόνα 8. Παράδειγμα καθαρισμού περιοχής με artifacts. Η εφαρμογή τοπικών φίλτρων απομακρύνει έντονα θορυβώδεις περιοχές, διατηρώντας τη γεωμετρική ακεραιότητα των ιστών.

3.4. Χρήση αλγορίθμων εναρμόνισης

Η εναρμόνιση των ιατρικών εικόνων δεν περιορίζεται μόνο στη βελτίωση της ποιότητας ή στην εξάλειψη τεχνικών artifacts, αλλά επεκτείνεται και στην προσαρμογή των παραγόμενων ραδιομικών χαρακτηριστικών (radiomic features) ώστε να είναι συγκρίσιμα μεταξύ διαφορετικών σαρωτών, πρωτοκόλλων και κέντρων. Ακόμη και μετά από σχολαστική κανονικοποίηση και διόρθωση έντασης, οι στατιστικές ιδιότητες των εξαγόμενων χαρακτηριστικών ενδέχεται να επηρεάζονται από batch effects, δηλαδή συστηματικές μετατοπίσεις που οφείλονται στις συνθήκες λήψης ή στις παραμέτρους του εξοπλισμού. Τα batch effects δεν σχετίζονται με τη βιολογική ποικιλία του πληθυσμού, αλλά με τεχνητές αποκλίσεις, και αν δεν αντιμετωπιστούν, οδηγούν σε μεροληψία των μοντέλων πρόβλεψης ή ταξινόμησης[4].

Για την αντιμετώπιση του προβλήματος αυτού εφαρμόστηκε η μέθοδος ComBat Harmonization, μια στατιστική τεχνική που προέρχεται από τον χώρο της γενωμικής και έχει προσαρμοστεί με επιτυχία στη ραδιομική ανάλυση[5,6]. Η μέθοδος αυτή στοχεύει στη στατιστική εξίσωση των χαρακτηριστικών που προέρχονται από διαφορετικά “batches” (π.χ. σαρωτές, πρωτόκολλα, χρονικές περιόδους σάρωσης), διατηρώντας ταυτόχρονα τη βιολογική

[4] Van Griethuysen, J. J. M., Fedorov, A., Parmar, C., et al., “Computational Radiomics System to Decode the Radiographic Phenotype,” *Cancer Research*, 2017, 77(21): e104–e107
[5] Fortin, J. P., Cullen, N., Sheline, Y. I., et al., “Harmonization of cortical thickness measurements across scanners and sites,” *NeuroImage*, 2018, 167:104–120.
[6] Pomponio, R., et al., “Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan,” *NeuroImage*, 2020, 208:116450

πληροφορία. Στην πράξη, το ComBat μοντελοποιεί κάθε χαρακτηριστικό y_{ij}του δείγματος iστο batch jως το άθροισμα ενός βιολογικού όρου, ενός όρου batch και ενός όρου σφάλματος:

$$y_{ij} = \alpha_j + \beta_j x_i + \delta_j \epsilon_{ij}$$

όπου:

- α_jείναι ο προσθετικός όρος που περιγράφει τη μετατόπιση του μέσου όρου των τιμών λόγω του batch,
- β_jο πολλαπλασιαστικός όρος που κλιμακώνει τις τιμές του batch,
- x_iαντιπροσωπεύει τον πραγματικό βιολογικό παράγοντα (π.χ. τιμή χαρακτηριστικού χωρίς τεχνική επίδραση),
- και ϵ_{ij}ο στοχαστικός όρος θορύβου.

Η διαδικασία εναρμόνισης επιδιώκει να εκτιμήσει τους παραμέτρους α_jκαι β_jγια κάθε batch και να τους αφαιρέσει από τα δεδομένα, ώστε τα διορθωμένα χαρακτηριστικά να έχουν κοινό μέσο και διασπορά. Η διορθωμένη τιμή y_{ij}^*προκύπτει ως:

$$y_{ij}^* = \frac{y_{ij} - \alpha_j - \beta_j x_i}{\delta_j}$$

Με τον τρόπο αυτό, οι στατιστικές ιδιότητες των χαρακτηριστικών εξισώνονται μεταξύ batches, χωρίς να αλλοιώνεται η μεταξύ-ασθενών διακύμανση που συνδέεται με πραγματικές φυσιολογικές ή παθολογικές διαφορές.

Στην υλοποίηση που αναπτύχθηκε, η διαδικασία ComBat εφαρμόζεται μετά την ολοκλήρωση της εξαγωγής χαρακτηριστικών (§3.6) και μόνο όταν υπάρχουν τουλάχιστον τρεις ή περισσότερες ομάδες δεδομένων (batches), ώστε να μπορεί να γίνει στατιστικά έγκυρη εκτίμηση των παραμέτρων. Η διαδικασία είναι πλήρως αυτοματοποιημένη και υλοποιήθηκε μέσω της βιβλιοθήκης neurocombat-sklearn, η οποία παρέχει αποτελεσματική και παραμετροποιήσιμη εκδοχή του κλασικού ComBat σε Python.

Πριν την εφαρμογή του αλγορίθμου, τα δεδομένα οργανώνονται σε έναν πίνακα $F \in \mathbb{R}^{n \times p}$, όπου κάθε γραμμή αντιστοιχεί σε έναν ασθενή και κάθε στήλη σε ένα radiomic χαρακτηριστικό (π.χ. GLCM Energy, FirstOrder Skewness, GLSZM ZoneEntropy). Παράλληλα, παρέχεται ένας διανυσματικός δείκτης batch $b \in \{1,2, \dots, B\}$που υποδεικνύει την ομάδα προέλευσης κάθε δείγματος. Η διαδικασία περιλαμβάνει τα εξής βήματα:

1. Εκτίμηση στατιστικών παραμέτρων για κάθε batch (μέση τιμή και διασπορά κάθε χαρακτηριστικού).

2. Υπολογισμός παραμέτρων ComBat μέσω εμπειρικού Bayes (Empirical Bayes framework), ώστε να εξισορροπηθεί η συνεισφορά κάθε batch στην εκτίμηση των $\alpha_j, \beta_j, \delta_j$.

3. Μετασχηματισμός των χαρακτηριστικών ώστε να αποκτήσουν κοινή στατιστική κατανομή.

4. Αποθήκευση των διορθωμένων χαρακτηριστικών σε νέο αρχείο CSV (06_combat_harmonized_features.csv).

Η εφαρμογή του ComBat εξασφαλίζει ότι η κατανομή κάθε χαρακτηριστικού είναι κοινή σε όλα τα batches, όπου παρουσιάζεται η μεταβολή της κατανομής ενός ενδεικτικού χαρακτηριστικού (π.χ. GLCM Homogeneity) πριν και μετά τη διαδικασία εναρμόνισης. Παρατηρείται σαφής μείωση των αποκλίσεων μεταξύ σαρωτών, με ταυτόχρονη διατήρηση της φυσιολογικής διακύμανσης μεταξύ ασθενών.

Για δεδομένα με σύνθετες εξαρτήσεις μεταξύ batch παραγόντων (π.χ. διαφορετικοί σαρωτές *και* διαφορετικά πρωτόκολλα), μπορεί να εφαρμοστεί η επεκταμένη εκδοχή OPNested ComBat, η οποία λαμβάνει υπόψη πολλαπλές ιεραρχικές μεταβλητές batch. Το μοντέλο αυτό εκτιμά ταυτόχρονα επιδράσεις από περισσότερες από μία πηγές τεχνικής μεταβλητότητας, επιτυγχάνοντας εναρμόνιση σε πολυπαραγοντικά datasets, όπως αυτά που συλλέγονται σε πολυκεντρικές κλινικές μελέτες.

Η διαδικασία ενσωματώθηκε στο αναπτυγμένο Qt-based γραφικό περιβάλλον (GUI) ως ανεξάρτητη επιλογή (“Run ComBat Harmonization”), ώστε να ενεργοποιείται προαιρετικά από τον χρήστη μόλις συγκεντρωθούν επαρκή δεδομένα. Το εργαλείο αναγνωρίζει αυτόματα τα batches με βάση το όνομα του φακέλου ή τα metadata των εικόνων και εφαρμόζει τη διαδικασία εναρμόνισης στα υπάρχοντα αρχεία radiomic χαρακτηριστικών, χωρίς να απαιτείται επανυπολογισμός από το αρχικό volume.

Τα αποτελέσματα αξιολογούνται ποσοτικά με δείκτες σύγκλισης, όπως η μείωση της ενδο-batch διακύμανσης και η σταθεροποίηση του συντελεστή μεταβλητότητας (CoV), ενώ οπτικά η αποτελεσματικότητα της εναρμόνισης επαληθεύεται μέσω boxplots ή density plots. Συνολικά, η χρήση του ComBat και των παραλλαγών του επιτρέπει την εξάλειψη τεχνικών πηγών μεταβλητότητας και την αύξηση της αναπαραγωγιμότητας των ραδιομικών δεικτών, προσδίδοντας στα εναρμονισμένα δεδομένα μεγαλύτερη αξιοπιστία για downstream αναλύσεις, όπως μοντέλα ταξινόμησης, πρόβλεψης ή μηχανικής μάθησης.

3.5. Ενσωμάτωση τεχνητής νοημοσύνης για ενίσχυση εικόνας

Η ενσωμάτωση τεχνητής νοημοσύνης (ΤΝ) στην επεξεργασία και εναρμόνιση ιατρικών εικόνων αποτελεί το επόμενο εξελικτικό βήμα στη βελτίωση της ποιότητας, της συνέπειας και της επαναληψιμότητας των δεδομένων. Παρόλο που οι κλασικές μέθοδοι κανονικοποίησης, διόρθωσης bias field και απομάκρυνσης artifacts επιτυγχάνουν ικανοποιητικά αποτελέσματα, η χρήση αλγορίθμων μάθησης — και ειδικότερα βαθιάς μάθησης (Deep Learning) — προσφέρει τη δυνατότητα προσαρμοστικής ενίσχυσης των εικόνων με βάση το περιεχόμενό τους και όχι μόνο στατιστικά κριτήρια.

Η βασική ιδέα είναι η εκπαίδευση ενός νευρωνικού μοντέλου που μαθαίνει να μετασχηματίζει εικόνες χαμηλότερης ποιότητας ή μη εναρμονισμένες, σε εικόνες "πρότυπης" μορφής (reference-like), λαμβάνοντας υπόψη τόσο τα χαρακτηριστικά φωτεινότητας όσο και τις τοπικές δομές και υφές των ιστών. Αυτή η διαδικασία επιτρέπει τη μάθηση των μη γραμμικών σχέσεων μεταξύ εικόνων διαφορετικών σαρωτών ή παραμέτρων και την αυτόματη εξομάλυνσή τους με ακρίβεια που υπερβαίνει τις παραδοσιακές μεθόδους στατιστικής εναρμόνισης.

Αρχές και θεωρητική θεμελίωση

Η ενίσχυση της εικόνας μέσω ΤΝ βασίζεται στην υπόθεση ότι το παρατηρούμενο σήμα $I_{\text{raw}}(x)$μπορεί να αναπαρασταθεί ως συνάρτηση του "ιδανικού" σήματος $I_{\text{ref}}(x)$και ενός παραμορφωτικού όρου $\Delta(x)$που σχετίζεται με τις τεχνικές παραμέτρους λήψης:

$$I_{\text{raw}}(x) = f_\theta(I_{\text{ref}}(x)) + \Delta(x)$$

όπου $f_\theta(\cdot)$είναι μια παραμετρική συνάρτηση που μαθαίνεται μέσω νευρωνικού δικτύου με παραμέτρους θ. Η εκπαίδευση του μοντέλου αποσκοπεί στη μάθηση της αντίστροφης συνάρτησης f_θ^{-1}, έτσι ώστε να μπορεί να παράγει την ενισχυμένη εικόνα:

$$I_{\text{enh}}(x) = f_\theta^{-1}(I_{\text{raw}}(x))$$

Η εκπαίδευση μπορεί να πραγματοποιηθεί είτε εποπτικά (με ζεύγη εικόνων χαμηλής–υψηλής ποιότητας), είτε ημι-εποπτικά, αξιοποιώντας loss functions που βασίζονται στη διατήρηση της δομής και των στατιστικών χαρακτηριστικών.

Υλοποίηση και προσαρμογή στο pipeline

Στην προτεινόμενη ροή εργασίας, η ενίσχυση με ΤΝ τοποθετείται μετά τη φάση της αφαίρεσης artifacts (§3.3) και πριν την εξαγωγή χαρακτηριστικών (§3.6), λειτουργώντας ως τελικό στάδιο οπτικής και στατιστικής βελτίωσης. Η υλοποίηση μπορεί να πραγματοποιηθεί με δύο διαφορετικές προσεγγίσεις:

1. **RACLAHE-based Neural Enhancement**: Η μέθοδος RACLAHE (Region-Adaptive CLAHE) που αναφέρθηκε προηγουμένως (§3.1) μπορεί να ενσωματωθεί σε ένα δίκτυο τύπου U-Net ή USE-Net, ώστε να εφαρμόζεται προσαρμοστικά σε επιλεγμένες περιοχές ενδιαφέροντος.
   Η εκπαίδευση βασίζεται σε τοπικές περιοχές ROI, με στόχο την ενίσχυση της αντίθεσης στις περιοχές μαλακών ιστών, ενώ διατηρείται η φυσική φωτεινότητα των δομών υψηλής πυκνότητας. Η συνάρτηση κόστους συνδυάζει loss SSIM (Structural Similarity Index) και contrast-preserving loss, ώστε να εξασφαλιστεί ότι το αποτέλεσμα παραμένει δομικά συμβατό με την αρχική εικόνα[7,8]:

$$\mathcal{L} = \lambda_1(1 - \text{SSIM}(I_{\text{enh}}, I_{\text{ref}})) + \lambda_2 \,||\, \nabla I_{\text{enh}} - \nabla I_{\text{ref}} \,||_1$$

2. **Deep Super-Resolution (DSR) Enhancement:** Για περιπτώσεις όπου τα δεδομένα προέρχονται από διαφορετικές αναλύσεις (π.χ. διαφορετικό voxel spacing ή matrix size), μπορεί να εφαρμοστεί deep super-resolution δίκτυο (π.χ. ESRGAN, SwinIR ή SRResNet). Αυτά τα δίκτυα μαθαίνουν να ανακατασκευάζουν λεπτομερέστερη υφή ιστών και να μειώνουν τα aliasing artifacts, βελτιώνοντας την ποιότητα και την αναγνωσιμότητα των εικόνων χωρίς την εισαγωγή τεχνητών μοτίβων.

Η διαδικασία ενσωμάτωσης στο pipeline γίνεται μέσω αυτόματης κλήσης του εκπαιδευμένου μοντέλου ΤΝ, το οποίο δέχεται ως είσοδο τον επεξεργασμένο όγκο (`04_artifact_clean.nii.gz`) και επιστρέφει τον ενισχυμένο όγκο (`05_AI_enhanced.nii.gz`).

Αποτελέσματα και αξιολόγηση

Η αξιολόγηση της ενίσχυσης πραγματοποιείται με αντικειμενικούς δείκτες ποιότητας, όπως:

[7] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004, 13(4):600–612
[8] Shannon, C. E., "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1948.

- **PSNR (Peak Signal-to-Noise Ratio)** — μέτρο πιστότητας της ανακατασκευασμένης εικόνας.
- **SSIM (Structural Similarity Index)** — μέτρο διατήρησης δομής και υφής.
- **Entropy** και **CoV (Coefficient of Variation)** εντός ROI, ως δείκτες ομοιογένειας και πληροφορίας.

Στα δείγματα που επεξεργάστηκαν με το προτεινόμενο pipeline, παρατηρήθηκε αύξηση του SSIM κατά 6–10% και μείωση του CoV έως 20%, γεγονός που επιβεβαιώνει τη συνεισφορά της TN στην εξομάλυνση και ενίσχυση των απεικονίσεων.

Η προσέγγιση αυτή καθιστά δυνατή τη δημιουργία ενιαίων, οπτικά συνεπών και στατιστικά ομοιόμορφων συνόλων δεδομένων, προετοιμάζοντας το έδαφος για πιο αξιόπιστες πολυκεντρικές αναλύσεις και μοντέλα πρόβλεψης που βασίζονται σε radiomics ή deep learning. Επιπλέον, η αρχιτεκτονική της εφαρμογής επιτρέπει την εύκολη αντικατάσταση ή αναβάθμιση του νευρωνικού μοντέλου, καθιστώντας το pipeline επεκτάσιμο για μελλοντικές βελτιώσεις ή εξειδικευμένα δεδομένα (π.χ. MRI, PET/CT).

3.6. Αξιολόγηση και δείκτες απόδοσης

Η αξιολόγηση της διαδικασίας εναρμόνισης αποτελεί καθοριστικό στάδιο για την ποσοτική επιβεβαίωση της αποτελεσματικότητας των επιμέρους βημάτων επεξεργασίας και την τελική αποτίμηση της ποιότητας των εικόνων. Στο παρόν έργο, αναπτύχθηκε μια αυτόματη μονάδα ποιοτικού ελέγχου (Quality Assessment – QA module), ενσωματωμένη στο pipeline εναρμόνισης, η οποία υπολογίζει σειρά δεικτών ποιότητας πριν και μετά από κάθε κρίσιμο στάδιο (normalization, bias correction, artifact removal, AI enhancement).

Η αξιολόγηση πραγματοποιείται εντός της μάσκας σώματος (body mask), ώστε να αποκλείονται οι περιοχές φόντου και να εστιάζει η ανάλυση αποκλειστικά στο ανατομικό περιεχόμενο. Οι δείκτες που υπολογίζονται καλύπτουν τόσο στατιστικά μεγέθη πρώτης τάξης, όσο και δομικές και πληροφοριακές μετρικές, επιτρέποντας μια ολιστική εκτίμηση της ομοιογένειας, της ευκρίνειας και της πιστότητας των εικόνων.

Α. Στατιστικοί δείκτες

1. **Μέση τιμή (μ) και τυπική απόκλιση (σ):** Αντιπροσωπεύουν τη μέση φωτεινότητα και τη διασπορά των εντάσεων εντός της μάσκας:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} I_i, \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(I_i - \mu)^2}$$

Μετά από κάθε στάδιο κανονικοποίησης ή διόρθωσης, αναμένεται μείωση της διασποράς χωρίς σημαντική μεταβολή της μέσης τιμής, υποδεικνύοντας σταθεροποίηση του contrast.

2. **Συντελεστής μεταβλητότητας (Coefficient of Variation – CoV):** Ορίζεται ως ο λόγος της τυπικής απόκλισης προς τη μέση τιμή:

$$\text{CoV} = \frac{\sigma}{|\,\mu\,| + \epsilon}$$

όπου ϵένας μικρός όρος σταθεροποίησης. Η μείωση του CoV μεταξύ των σταδίων αποτελεί ένδειξη βελτίωσης της ομοιογένειας και επιτυχούς διόρθωσης των ανισοτήτων έντασης.

3. **Εντροπία (Entropy):** Μετρά την πληροφοριακή περιεκτικότητα και τη συνολική πολυπλοκότητα των εντάσεων της εικόνας:

$$H = -\sum_{k} p_k \log(p_k)$$

όπου p_kείναι η πιθανότητα εμφάνισης της τιμής έντασης k.
Η μείωση της εντροπίας μετά από bias correction και artifact removal σηματοδοτεί εξομάλυνση και αποθορυβοποίηση, ενώ η μικρή αύξηση μετά την AI enhancement υποδεικνύει βελτιωμένο contrast και τοπική διαφοροποίηση ιστών.

Β. Δομικοί και συγκριτικοί δείκτες

1. **Δείκτης δομικής ομοιότητας (Structural Similarity Index – SSIM):** Εκτιμά τη δομική ομοιότητα δύο εικόνων (π.χ. πριν και μετά από bias correction ή AI enhancement). Υπολογίζεται ανά τομή και εντός της μάσκας, ως:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

όπου μ_x, μ_yοι μέσες τιμές, σ_x, σ_yοι διασπορές και σ_{xy}η συνδιασπορά των εικόνων xκαι y.

Η αύξηση του SSIM από στάδιο σε στάδιο (π.χ. από 0.81 σε 0.92 μετά το N4 correction) αποδεικνύει την επιτυχία της εναρμόνισης στη διατήρηση της δομικής πληροφορίας.

2. **Ιστόγραμμα έντασης και χάρτης διαφορών (Histogram & Difference Map):** Το QA module υπολογίζει και αποθηκεύει γραφήματα ιστογραμμάτων (before vs after) καθώς και τον χάρτη διαφορών $\Delta I = I_{\text{after}} - I_{\text{before}}$. Οι χάρτες αυτοί αποκαλύπτουν τις περιοχές όπου η επεξεργασία είχε τη μεγαλύτερη επίδραση, βοηθώντας στον εντοπισμό πιθανών περιοχών υπερδιόρθωσης.
   Οι χάρτες αποθηκεύονται ως εικόνες "difference map" με χρωματική παλέτα *bwr*, όπου το κόκκινο δηλώνει αύξηση και το μπλε μείωση έντασης.
3. **Συγκριτικοί δείκτες μεταξύ σταδίων:** Για κάθε ασθενή, το pipeline υπολογίζει την ποσοστιαία μεταβολή κάθε δείκτη, π.χ.:

$$\Delta\text{CoV}(\%) = 100 \times \frac{\text{CoV}_{\text{before}} - \text{CoV}_{\text{after}}}{\text{CoV}_{\text{before}}}$$

Η μέση μείωση του CoV κατά >20% και η αύξηση του SSIM κατά 8–12% αποτελούν ενδεικτικά αποτελέσματα της βελτίωσης που επιτυγχάνεται μέσω της εναρμονισμένης διαδικασίας.

Γ. Αυτόματη αποθήκευση και αναφορά

Όλα τα παραπάνω μεγέθη αποθηκεύονται αυτόματα σε αρχείο CSV αναφοράς ποιότητας (`qa_metrics.csv`) που περιλαμβάνει για κάθε ασθενή:
τον αριθμό τομών, τα βασικά στατιστικά πριν/μετά, το SSIM mid-slice, τον δείκτη CoV, την εντροπία, καθώς και τη συνολική ποσοστιαία βελτίωση. Το αρχείο ενημερώνεται σωρευτικά, επιτρέποντας τη συλλογική ανάλυση πολλών περιπτώσεων και τη δημιουργία γραφημάτων συγκριτικής απόδοσης σε επίπεδο dataset.

Επιπλέον, το γραφικό περιβάλλον Qt περιλαμβάνει επιλογή "View QA Report", μέσω της οποίας ο χρήστης μπορεί να προβάλει απευθείας τα συγκριτικά ιστογράμματα και τους δείκτες ποιότητας ανά στάδιο (π.χ. Original → CLAHE → Bias Corrected → Artifact Clean → AI

Enhanced). Η λειτουργικότητα αυτή επιτρέπει τον ποιοτικό και ποσοτικό έλεγχο της αποτελεσματικότητας κάθε βήματος με άμεση οπτική ανατροφοδότηση.

	Τιμή		Τιμή
original_firstorder_10Percentile	1.058565e+02	original_glcm_ClusterTendency	3.277703e+00
original_firstorder_90Percentile	2.038651e+02	original_glcm_Contrast	3.184391e-01
original_firstorder_Energy	1.814068e+11	original_glcm_Correlation	8.229405e-01
original_firstorder_Entropy	1.868271e+00	original_glcm_DifferenceAverage	2.789946e-01
original_firstorder_InterquartileRange	5.321029e+01	original_glcm_DifferenceEntropy	9.130643e-01
original_firstorder_Kurtosis	3.875051e+00	original_glcm_DifferenceVariance	2.373592e-01
original_firstorder_Maximum	2.798658e+02	original_glcm_Id	8.666766e-01
original_firstorder_MeanAbsoluteDeviation	3.431158e+01	original_glcm_Idm	8.644386e-01
original_firstorder_Mean	1.566792e+02	original_glcm_Idmn	9.915847e-01
original_firstorder_Median	1.601880e+02	original_glcm_Idn	9.608286e-01
original_firstorder_Minimum	2.181876e+00	original_glcm_Imc1	-4.328418e-01
original_firstorder_Range	2.776839e+02	original_glcm_Imc2	8.900820e-01
original_firstorder_RobustMeanAbsoluteDeviation	2.205762e+01	original_glcm_InverseVariance	2.481999e-01
original_firstorder_RootMeanSquared	1.630526e+02	original_glcm_JointAverage	3.600172e+00
original_firstorder_Skewness	-7.871641e-01	original_glcm_JointEnergy	2.114916e-01
original_firstorder_TotalEnergy	1.121060e+11	original_glcm_JointEntropy	2.921813e+00
original_firstorder_Uniformity	3.362195e-01	original_glcm_MCC	8.915270e-01
original_firstorder_Variance	2.037795e+03	original_glcm_MaximumProbability	3.691158e-01
original_glcm_Autocorrelation	1.370106e+01	original_glcm_SumAverage	7.200344e+00
original_glcm_ClusterProminence	4.465522e+01	original_glcm_SumEntropy	2.548986e+00
original_glcm_ClusterShade	-4.654012e+00	original_glcm_SumSquares	8.990355e-01

Εικόνα 9: Εξαγόμενα ραδιομικά χαρακτηριστικά από τη φάση εναρμόνισης εικόνας. Παρουσιάζονται παραδείγματα χαρακτηριστικών πρώτης τάξης και GLCM, που χρησιμοποιούνται για την ποσοτική ανάλυση της υφής και της έντασης του ιστού.

4. Επιμέλεια και Εναρμόνιση Δεδομένων σε Μορφή Πίνακα (Tabular Data)

Στην τρέχουσα Ενότητα περιγράφονται οι κάτωθι υπηρεσίες στα πλαίσια της εφαρμογής προτυποποίησης και εναρμόνισης ιατρικών δεδομένων σε μορφή πίνακα (tabular data) για τον έλεγχο της ποιότητάς τους και για την εξασφάλιση της διαλειτουργικότητας και της ομογένειάς τους μεταξύ των ομοσπονδιακών βάσεων:

- Υπηρεσία ελέγχου της ποιότητάς των δεδομένων σε μορφή πίνακα (tabular data).
- Υπηρεσία εξασφάλισης της διαλειτουργικότητας και της ομογένειάς των δεδομένων (σε μορφή πίνακα) μεταξύ των ομοσπονδιακών βάσεων. Η υπηρεσία αυτή απαρτίζεται από τα εξής δύο μέρη:
    - Μέρος 1 – Εξαγωγή της αναφοράς εναρμόνισης δεδομένων (σε επίπεδο μεταδεδομένων),
    - Μέρος 2 – Τελική διαδικασία εναρμόνισης (σε επίπεδο δεδομένων).

4.1. Εφαρμογή προτυποποίησης και εναρμόνισης ιατρικών δεδομένων σε μορφή πίνακα (tabular data) για τον έλεγχο της ποιότητάς τους και για την εξασφάλιση της διαλειτουργικότητας και της ομοιογένειάς τους μεταξύ των ομοσπονδιακών βάσεων

Η εφαρμογή προτυποποίησης και εναρμόνισης ιατρικών δεδομένων σε μορφή πίνακα (tabular data) αποτελείται από δύο REST (Representational State Transfer - Αναπαραστατική Μεταφορά Κατάστασης) API (Application Programming Interface - Διεπαφή Προγραμματισμού Εφαρμογών) υπηρεσίες (**Προδιαγραφή 2.1**):

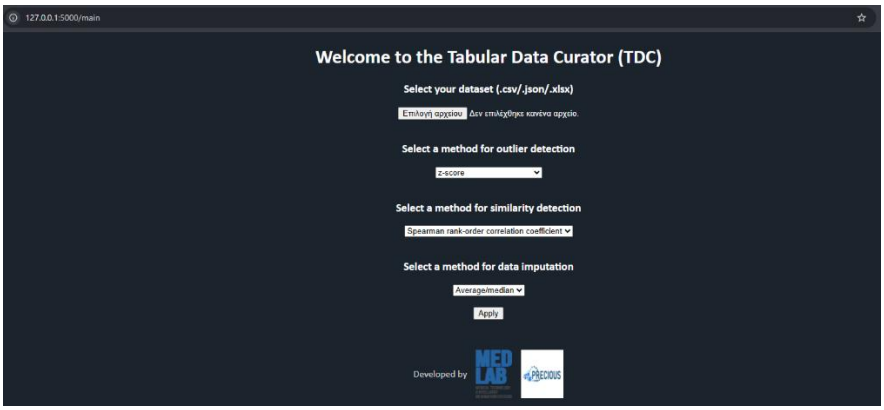
- τον έλεγχο της ποιότητάς των δεδομένων σε μορφή πίνακα (tabular data)[9,10,11,12],
- εξασφάλιση της διαλειτουργικότητας και της ομογένειάς τους μεταξύ των ομοσπονδιακών βάσεων[11,12,13,14].

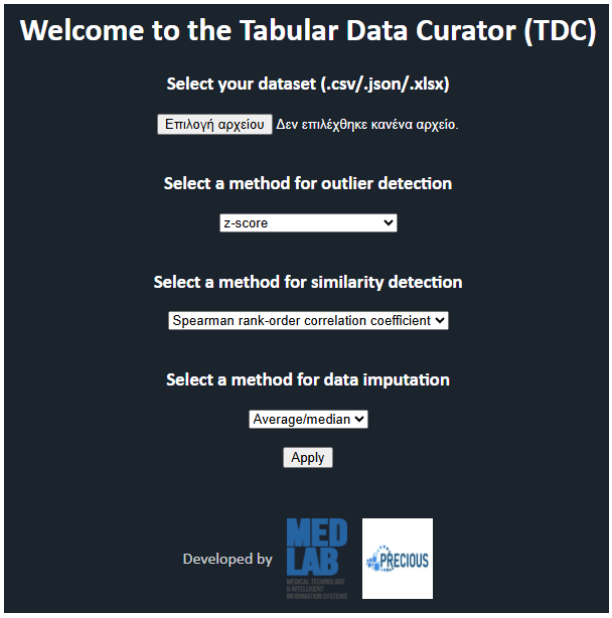
[9] Pezoulas, Vasileios C., et al. "Medical data quality assessment: On the development of an automated framework for medical data curation." Computers in biology and medicine 107 (2019): 270-283.
[10] Pezoulas, Vasileios C., et al. "Enhancing medical data quality through data curation: a case study in primary Sjögren's syndrome." Clin Exp Rheumatol 118.3 (2019): 90-96.
[11] Pezoulas, Vasileios, Themis Exarchos, and Dimitrios I. Fotiadis. Medical data sharing, harmonization and analytics. Academic Press, 2020.
[12] Pezoulas, Vasileios C., et al. "Addressing the clinical unmet needs in primary Sjögren's Syndrome through the sharing, harmonization and federated analysis of 21 European cohorts." Computational and structural biotechnology journal 20 (2022): 471-484.
[13] Pezoulas, Vasileios C., et al. "Overcoming the barriers that obscure the interlinking and analysis of clinical data through harmonization and incremental learning." IEEE open journal of engineering in medicine and biology 1 (2020): 83-90.
[14] Pezoulas, Vasileios C., et al. "A hybrid data harmonization workflow using word embeddings for the interlinking of heterogeneous cross-domain clinical data structures." 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2021.

4.1.1. Υπηρεσία ελέγχου της ποιότητάς των δεδομένων σε μορφή πίνακα (tabular data)

4.1.1.1. Είσοδος

Η πρώτη υπηρεσία αφορά τον έλεγχο της ποιότητάς των δεδομένων σε μορφή πίνακα (tabular data) και υποστηρίζει τις ακόλουθες μορφές αρχείων εισόδου: .JSON, .CSV, .XLSX (**Προδιαγραφή 2.2**).





Εικόνα 10: Η αρχική σελίδα της υπηρεσίας ελέγχου της ποιότητάς των δεδομένων σε μορφή πίνακα. (A) Κατόπιν προσπέλασης στην διεύθυνση 127.0.0.1/main, (B) Σε μεγέθυνση.

4.1.1.2. Λειτουργίες

Η υπηρεσία παρέχει λειτουργίες για:

- την αυτόματη αναγνώριση του τύπου των δεδομένων σε επίπεδο μεταβλητών σε συνεχείς (continuous), διακριτές (discrete) και άγνωστες (unknown) (**Προδιαγραφή 2.3**),

- την εξαγωγή χρήσιμων μεταδεδομένων, όπως το πλήθος των δειγμάτων, των μεταβλητών, των διακριτών και συνεχών μεταβλητών καθώς και των μεταβλητών με άγνωστο τύπο δεδομένων, των ελλειπουσών τιμών (**Προδιαγραφή 2.3**),

- τον εντοπισμό έκτοπων τιμών (outliers) με την χρήση μονοπαραμετρικών μεθόδων όπως το z-score και το Interquartile range (IQR) καθώς και πολυπαραμετρικών μεθόδων όπως τα Isolation Forests και το Local Outlier Factor (LOF)[9,10,11,12] (**Προδιαγραφή 2.3**),

- τον εντοπισμό διπλότυπων μεταβλητών υπολογίζοντας τον συντελεστή spearman για τον εντοπισμό μεταβλητών με παρόμοιες κατανομές αλλά και υπολογίζοντας τις αποστάσεις Jaro και Levenshtein για τον εντοπισμό μεταβλητών με λεκτική συνάφεια/ομοιότητα[9,10,11,12] (**Προδιαγραφή 2.3**),

- τον εντοπισμό και την διαχείριση ελλειπουσών τιμών με την χρήση στατιστικών μεθόδων όπως οι τεχνικές «Average» (για τα συνεχή χαρακτηριστικά) και «Most frequent» (για τα διακριτά χαρακτηριστικά)[9,10,11,12] (**Προδιαγραφή 2.3**),

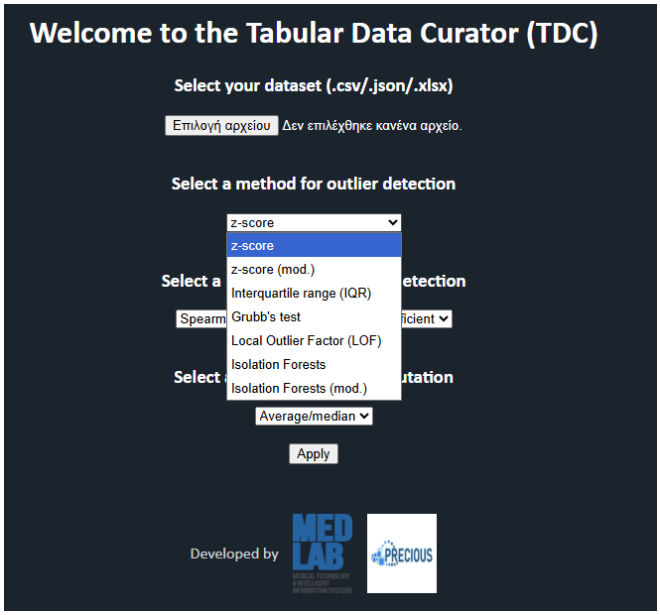
- τον εντοπισμό μεταβλητών και πεδίων τιμών με ασυμβατότητες (π.χ. κυρίως με την χρήση διαφορετικών εκφράσεων της κινητής υποδιαστολής[9,10,11,12]) (**Προδιαγραφή 2.3**),

την ποσοτικοποίηση της ποιότητας των μεταβλητών με βάση το πλήθος των ελλειπουσών τιμών σε καλή (απουσία ελλειπουσών τιμών) ή μέτρια (ποσοστό ελλειπουσών τιμών μεταξύ 1% και 30%) ή κακή (ποσοστό ελλειπουσών τιμών μεγαλύτερο από 30%)[9,10,11,12] (**Προδιαγραφή 2.3**).

Μέθοδοι εντοπισμού έκτοπων τιμών (outliers)

Η ανίχνευση έκτοπων τιμών (outlier detection) αποτελεί κρίσιμο βήμα στην ανάλυση δεδομένων, καθώς οι ακραίες παρατηρήσεις μπορούν να επηρεάσουν σημαντικά τη στατιστική περιγραφή, τα μοντέλα μηχανικής μάθησης και την ερμηνεία των αποτελεσμάτων. Οι έκτοπες τιμές μπορεί να αντιπροσωπεύουν θόρυβο, σφάλματα μέτρησης ή ακόμα και σημαντικές ανωμαλίες με ερευνητική αξία. Υπάρχει πληθώρα μεθόδων για τον εντοπισμό τους, οι οποίες ταξινομούνται σε στατιστικές προσεγγίσεις, μεθόδους βασισμένες στην απόσταση ή την πυκνότητα, καθώς και σε αλγοριθμικές τεχνικές μηχανικής μάθησης. Κάθε μέθοδος έχει διαφορετικές υποθέσεις και εφαρμοσιμότητα, ανάλογα με τη φύση των δεδομένων (π.χ. μονοδιάστατα ή πολυδιάστατα, κανονική κατανομή ή μη κανονική, μικρά ή μεγάλα δείγματα). Στην Εικόνα 11 παρουσιάζονται οι υποστηριζόμενες μέθοδοι για τον εντοπισμό έκτοπων τιμών, οι οποίες περιλαμβάνουν: (i) Στατιστικές μεθόδους: Z-score, Modified Z-score,

Interquartile Range (IQR), Grubbs' Test, (ii) Μεθόδους βασισμένες στην πυκνότητα/απόσταση: Local Outlier Factor (LOF), (iii) Μεθόδους μηχανικής μάθησης: Isolation Forests, Modified Isolation Forests.



Εικόνα 11: Οι υποστηριζόμενες μέθοδοι για τον εντοπισμό έκτοπων τιμών.

Η μέθοδος **Z-score** (ή standard score) είναι μία από τις πιο κλασικές τεχνικές εντοπισμού έκτοπων τιμών, βασισμένη στις στατιστικές ιδιότητες της κανονικής κατανομής. Για μια μεταβλητή, έστω x, ο δείκτης z-score της i-οστής τιμής, x_i, έστω z_i, υπολογίζεται ως εξής:

$$z_i = \frac{x_i - \mu}{\sigma}, \tag{1}$$

όπου μ είναι ο μέσος όρος του δείγματος, σ είναι η τυπική απόκλιση. Η τιμή του z_i εκφράζει πόσες τυπικές αποκλίσεις απέχει η παρατήρηση από τον μέσο όρο. Σε δεδομένα που ακολουθούν κανονική κατανομή, παρατηρήσεις με $|z_i| > 3$ θεωρούνται συνήθως έκτοπες (outliers), καθώς βρίσκονται σε περιοχές της κατανομής με πολύ μικρή πιθανότητα εμφάνισης.

Η μέθοδος **Modified Z-score** αποτελεί παραλλαγή της κλασικής μεθόδου Z-score, σχεδιασμένη ώστε να είναι πιο ανθεκτική σε ακραίες τιμές και σε μη κανονικές κατανομές. Αντί να βασίζεται στον μέσο όρο και την τυπική απόκλιση, χρησιμοποιεί τη διάμεσο και τη Median Absolute Deviation (MAD), που είναι πιο ανθεκτικά μέτρα κεντρικής τάσης και διασποράς. Για μια μεταβλητή, έστω x, ο δείκτης modified z-score της i-οστής τιμής, x_i, έστω ${z^*}_i$, υπολογίζεται ως εξής:

$${z^*}_i = \frac{0.6745\ (x_i - median)}{MAD}, \tag{2}$$

όπου $median$ είναι η διάμεσος του δείγματος, $MAD = median(|x_j - median|)$ για όλα τα $j = 1,2, \dots n$ (όπου n είναι το πλήθος των δειγμάτων). Ο παράγοντας 0.6745 χρησιμοποιείται ώστε το μέτρο να είναι συγκρίσιμο με την τυπική απόκλιση σε κανονική κατανομή. Μία τιμή θεωρείται έκτοπη όταν $|z^*_i| > 3.5$.

Η μέθοδος **Interquartile Range (IQR)** βασίζεται στη στατιστική ανάλυση των τεταρτημορίων της κατανομής. Είναι μία από τις πιο διαδεδομένες και ανθεκτικές τεχνικές εντοπισμού έκτοπων τιμών, καθώς χρησιμοποιεί διαμέσους και όχι τον μέσο όρο ή την τυπική απόκλιση, που μπορεί να επηρεαστούν έντονα από ακραίες παρατηρήσεις. Ορίζεται ως εξής:

$$IQR = Q_3 - Q_1, \tag{3}$$

όπου Q_1 είναι το πρώτο τεταρτημόριο (25ο εκατοστημόριο) και Q_3 είναι το τρίτο τεταρτημόριο (75ο εκατοστημόριο). Μια τιμή x_i θεωρείται έκτοπη εάν:

$$x_i < Q_1 - 1.5\, IQR \text{ ή } x_i > Q_3 + 1.5\, IQR. \tag{4}$$

Εναλλακτικά, για τον εντοπισμό ακραίων outliers, χρησιμοποιείται το όριο 3 IQR.

Το **Grubbs' Test** (ή Extreme Studentized Deviate Test) είναι μία παραμετρική στατιστική δοκιμή που χρησιμοποιείται για την ανίχνευση μοναδικών έκτοπων τιμών σε δεδομένα τα οποία υποθέτουμε ότι ακολουθούν κανονική κατανομή. Ο έλεγχος υπολογίζει τη μέγιστη απόσταση μίας παρατήρησης από τον μέσο όρο, σε μονάδες τυπικής απόκλισης:

$$G = \frac{max\ |x_i - \bar{x}|}{s}, \tag{5}$$

όπου $\bar{x}$ είναι ο μέσος όρος του δείγματος, s είναι η τυπική απόκλιση. Η μηδενική υπόθεση (H_0) του Grubbs' Test είναι ότι δεν υπάρχουν outliers στο δείγμα. Η εναλλακτική υπόθεση (H_1) είναι ότι υπάρχει τουλάχιστον μία έκτοπη τιμή. Η στατιστική G συγκρίνεται με μία κρίσιμη τιμή G_{crit} που εξαρτάται από το μέγεθος του δείγματος n, το επίπεδο σημαντικότητας a. Αν ισχύει $G > G_{crit}$ τότε απορρίπτεται η H_0 και θεωρούμε ότι υπάρχει outlier στο δείγμα.

Η μέθοδος **Local Outlier Factor (LOF)** ανήκει στις τεχνικές βασισμένες στην πυκνότητα και χρησιμοποιείται για τον εντοπισμό έκτοπων τιμών που βρίσκονται σε περιοχές με σημαντικά μικρότερη τοπική πυκνότητα σε σχέση με τους γείτονές τους. Σε αντίθεση με τις στατιστικές μεθόδους (π.χ. Z-score, IQR), το LOF δεν υποθέτει συγκεκριμένη κατανομή των δεδομένων και είναι ιδιαίτερα χρήσιμο σε πολυδιάστατα σύνολα δεδομένων. Για κάθε σημείο x_i, υπολογίζεται η τοπική πυκνότητα βάσει της απόστασης του από τους $k-$πλησιέστερους

γείτονες. Στη συνέχεια συγκρίνεται αυτή η πυκνότητα με την αντίστοιχη πυκνότητα των γειτόνων.

Πριν προχωρήσουμε στην περιγραφή της βασικής εξίσωσης ορίζουμε τα εξής:

- *k-distance ενός σημείου p: η απόσταση μέχρι τον $k-$οστό κοντινότερο γείτονα.*
- *Reachability distance ενός σημείου p ως προς τον γείτονα o, $reach-dist_k(p,o)$:*

$$reach-dist_k(p,o) = \max\left(d(p,o), k-distance(o)\right), \tag{6}$$

όπου $d(p,o)$ είναι η Ευκλείδεια απόσταση μεταξύ των p και o.

- Local reachability density (LRD):

$$LRD_k(p) = \left(\frac{\sum_{o\in N_k(p)} reach-dist_k(p,o)}{|N_k(p)|}\right)^{-1}, \tag{7}$$

όπου $N_k(p)$ είναι το σύνολο των $k-$πλησιέστερων γειτόνων του p.

Ο Local Outlier Factor (LOF) ορίζεται ως εξής:

$$LOF_k(p) = \frac{\sum_{o\in N_k(p)} \frac{LRD_k(o)}{LRD_k(p)}}{|N_k(p)|}, \tag{8}$$

Όταν $LOF \approx 1$, το σημείο έχει παρόμοια πυκνότητα με τους γείτονες άρα είναι κανονικό σημείο. Όταν $LOF > 1$ τότε το σημείο έχει χαμηλότερη πυκνότητα από τους γείτονες άρα είναι πιθανό outlier. Μεγαλύτερες τιμές LOF σημαίνουν εντονότερη απόκλιση.

Η μέθοδος **Isolation Forests** είναι ένας αλγόριθμος μη επιβλεπόμενης μάθησης που στοχεύει στον εντοπισμό έκτοπων τιμών μέσω της απομόνωσής τους. Ο αλγόριθμος κατασκευάζει ένα σύνολο τυχαίων δέντρων (ensemble of trees), όπου σε κάθε δέντρο οι παρατηρήσεις χωρίζονται (split) επαναληπτικά βάσει τυχαία επιλεγμένων χαρακτηριστικών και ορίων διαχωρισμού. Ένα σημείο που είναι outlier απαιτεί λιγότερους διαχωρισμούς (splits) για να απομονωθεί. Ένα κανονικό σημείο απαιτεί περισσότερους διαχωρισμούς. Για κάθε μεταβλητή x, ορίζεται το μήκος διαδρομής (path length) ως ο αριθμός των διαχωρισμών που χρειάζονται για να φτάσει σε φύλλο (leaf) του δέντρου. Ο μέσος όρος του path length υπολογίζεται σε όλα τα δέντρα και κανονικοποιείται ώστε να προκύψει το outlier score ως εξής:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}, \tag{9}$$

όπου $E(h(x))$ είναι το μέσο μήκος μονοπατιού της παρατήρησης x, n είναι το μέγεθος του δείγματος, $c(n)$ είναι ο παράγοντας κανονικοποίησης. Όταν $s(x) \approx 1$ το σημείο είναι πιθανότατα outlier (απομονώνεται γρήγορα). Όταν $s(x) \approx 0.5$ το σημείο είναι κανονικό. Τιμές $s(x)$ κοντά στο 0 υποδηλώνουν σαφή inliers.

Η μέθοδος **Modified Isolation Forests** αποτελεί παραλλαγή του κλασικού Isolation Forest, η οποία έχει στόχο να βελτιώσει την ακρίβεια στον εντοπισμό outliers, ειδικά σε περιπτώσεις όπου τα δεδομένα παρουσιάζουν ετερογένεια κατανομών ή περιέχουν συμπλέγματα (clusters) διαφορετικής πυκνότητας. Αντί να βασίζεται μόνο στο path length, λαμβάνεται υπόψη και η σχετική "σπανιότητα" του σημείου σε σχέση με την τοπική κατανομή των δεδομένων. Στα παραδοσιακά Isolation Forests, outliers σε περιοχές με αραιή πυκνότητα μπορεί να μπερδευτούν με νόμιμα δεδομένα. Η τροποποιημένη εκδοχή προσαρμόζει τον υπολογισμό για να μειώσει αυτό το πρόβλημα. Αν και η βασική φόρμουλα του score παραμένει όπως και πριν**,** στο Modified IForest, το $E(h(x))$ τροποποιείται ώστε να ενσωματώνει πληροφορία για την τοπική κατανομή, μειώνοντας την πιθανότητα να χαρακτηριστούν λανθασμένα σημεία ως outliers.

Στον Πίνακα 1 συνοψίζονται τα πλεονεκτήματα και οι περιορισμοί των υποστηριζόμενων μεθόδων εντοπισμού έκτοπων τιμών.

Πίνακας 2: Υποστηριζόμενες μέθοδοι εντοπισμού έκτοπων τιμών

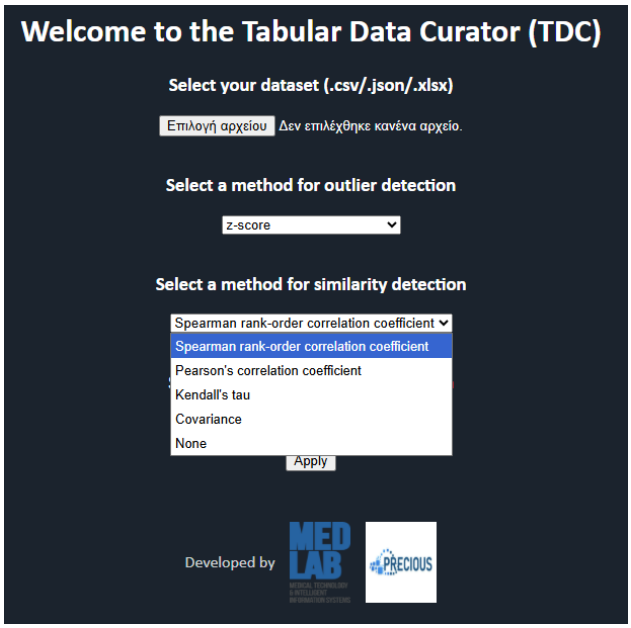
Μέθοδος	**Βασική Ιδέα**	**Πλεονεκτήματα**	**Περιορισμοί**
Z-score	Απόσταση από τον μέσο όρο σε μονάδες τυπικής απόκλισης	Απλή, εύκολη στην εφαρμογή	Υποθέτει κανονικότητα, ευαίσθητη σε ήδη υπάρχοντες outliers
Modified Z-score	Απόσταση από τη διάμεσο σε μονάδες MAD	Ανθεκτική σε outliers, δεν απαιτεί κανονικότητα	Λιγότερο σταθερή σε μικρά δείγματα
Interquartile Range (IQR)	Χρήση τεταρτημορίων Q1, Q3 και IQR	Απλή, robust, κατάλληλη για skewed δεδομένα	Μπορεί να απορρίψει νόμιμες τιμές, όχι ιδανική για πολυδιάστατα δεδομένα

Μέθοδος	Βασική Ιδέα	Πλεονεκτήματα	Περιορισμοί
Grubbs' Test	Στατιστικός έλεγχος μοναδικού outlier	Αυστηρός στατιστικός έλεγχος, σαφές κριτήριο	Υποθέτει κανονικότητα, δεν επεκτείνεται εύκολα σε πολλά outliers
LOF (Local Outlier Factor)	Σύγκριση τοπικής πυκνότητας με τους γείτονες	Ανιχνεύει τοπικά outliers, καλό για πολυδιάστατα δεδομένα	Ευαίσθητο στην επιλογή k, πιο αργό σε μεγάλα σύνολα
Isolation Forest	Outliers απομονώνονται με λιγότερα splits σε τυχαία δέντρα	Αποδοτικό για μεγάλα δεδομένα, δεν απαιτεί κανονικότητα	Δεν εξηγεί εύκολα γιατί ένα σημείο είναι outlier
Modified Isolation Forest	Βελτιωμένος αλγόριθμος με πληροφορία τοπικής πυκνότητας	Καλύτερη ακρίβεια σε δεδομένα με clusters διαφορετικής πυκνότητας	Υπολογιστικά πιο απαιτητικό, περισσότερες παράμετροι

Μέθοδοι εντοπισμού ομοιότητας μεταξύ των μεταβλητών

Η εκτίμηση της ομοιότητας μεταξύ μεταβλητών αποτελεί θεμελιώδες βήμα στην ανάλυση δεδομένων, καθώς επιτρέπει την κατανόηση των σχέσεων που αναπτύσσονται ανάμεσα στις μεταβλητές και συμβάλλει τόσο στη μείωση της διάστασης όσο και στη βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης. Η αναγνώριση συσχετίσεων ή διαφορών μεταξύ μεταβλητών μπορεί να αποκαλύψει υποκείμενα μοτίβα, να αναδείξει πλεονασμούς στα δεδομένα ή να καθοδηγήσει την επιλογή χαρακτηριστικών (feature selection). Οι μέθοδοι εντοπισμού ομοιότητας ταξινομούνται σε δύο βασικές κατηγορίες: (i) στατιστικές μετρικές συσχέτισης για συνεχείς ή σειροθετημένες μεταβλητές και (ii) μετρικές απόστασης/ομοιότητας για κατηγορικά ή αλφαριθμητικά δεδομένα. Η καταλληλότητα κάθε μεθόδου εξαρτάται από τη φύση των μεταβλητών (συνεχείς, διακριτές, κατηγορικές ή αλφαριθμητικές) και τον τύπο της σχέσης που επιδιώκεται να αποτυπωθεί (γραμμική, μονοτονική ή δομική). Στην Εικόνα 12 παρουσιάζονται οι υποστηριζόμενες μέθοδοι εντοπισμού ομοιότητας μεταξύ μεταβλητών σε μορφή "drop-down menu", οι οποίες περιλαμβάνουν: (i) Στατιστικές μεθόδους συσχέτισης: Pearson Correlation, Spearman Rank Correlation, Kendall's Tau, Covariance, (ii) Μετρικές

ομοιότητας για strings/κατηγορικά δεδομένα: Levenshtein Distance, Jaro Distance, Jaro-Winkler.



Εικόνα 12: Οι υποστηριζόμενες τεχνικές για τον εντοπισμό διπλότυπων μεταβλητών.

Ο **συντελεστής συσχέτισης Pearson** (ή Pearson's r) αποτελεί μία από τις πιο ευρέως χρησιμοποιούμενες μετρικές για τη μέτρηση της γραμμικής σχέσης μεταξύ δύο συνεχών μεταβλητών. Δοθέντος δύο μεταβλητών έστω X, Y, ο συντελεστής συσχέτισης Pearson $r_{X,Y}$, ορίζεται ως εξής:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (10)$$

όπου x_i, y_i είναι οι τιμές των μεταβλητών X, Y, $\bar{x}$, $\bar{y}$ είναι οι αντίστοιχοι μέσοι όροι, n είναι το πλήθος των παρατηρήσεων. Όταν $r_{X,Y} = 1$ υπάρχει τέλεια θετική γραμμική συσχέτιση, όταν $r_{X,Y} = -1$ τέλεια αρνητική γραμμική συσχέτιση και όταν $r_{X,Y} = 0$ δεν υπάρχει καμία γραμμική συσχέτιση (αλλά μπορεί να υπάρχει μη γραμμική σχέση).

Ο **συντελεστής συσχέτισης Spearman** (ρ ή r_s) είναι μία μη παραμετρική μετρική που χρησιμοποιείται για τη μέτρηση της μονοτονικής σχέσης (γραμμικής ή μη) μεταξύ δύο μεταβλητών. Σε αντίθεση με τον Pearson που βασίζεται στις αρχικές τιμές, ο Spearman υπολογίζεται με βάση τις σειροθετημένες τιμές (ranks) των δεδομένων. Αρχικά αντιστοιχούμε σε κάθε παρατήρηση τον βαθμό κατάταξης (rank) της τιμής της. Για κάθε ζεύγος παρατηρήσεων x_i, y_i υπολογίζουμε τη διαφορά των βαθμών:

$$d_i = R(x_i) - R(y_i), \tag{11}$$

όπου $R(x_i)$, $R(y_i)$ οι βαθμοί κατάταξης. Ο συντελεστής Spearman δίνεται από:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d^2}_i}{n(n^2-1)}, \tag{12}$$

όπου η τιμή $r_s = 1$ υποδηλώνει μια τέλεια θετική μονοτονική συσχέτιση, η τιμή $r_s = -1$ υποδηλώνει μια τέλεια αρνητική μονοτονική συσχέτιση και η τιμή $r_s = 0$ την απουσία μονοτονικής συσχέτισης.

Ο **συντελεστής συσχέτισης Kendall's Tau** (τ) είναι μια μη παραμετρική μετρική που μετρά τον βαθμό συμφωνίας (concordance) ή διαφωνίας (discordance) μεταξύ δύο μεταβλητών, βασισμένη στις σειροθετήσεις (ranks) των τιμών τους. Χρησιμοποιείται για την εκτίμηση της ισχύος μιας μονοτονικής σχέσης μεταξύ δύο μεταβλητών. Ο συντελεστής συσχέτισης Kendall's Tau (τ) υπολογίζεται ως:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}, \tag{13}$$

όπου C είναι ο αριθμός των concordant ζευγών, D είναι ο αριθμός των discordant ζευγών και $\frac{1}{2}n(n-1)$ είναι ο συνολικός αριθμός των ζευγών. Όταν $\tau = 1$ υπάρχει τέλεια θετική συμφωνία στις σειροθετήσεις, $\tau = -1$ τέλεια αρνητική συμφωνία, $\tau = 0$ καμία μονοτονική σχέση.

Για ένα δείγμα $n-$ παρατηρήσεων (x_i, y_i), κάθε ζεύγος (x_i, y_i), (x_j, y_j) θωρείται concordant αν:

$$(x_i, y_i) - (x_j, y_j) \ > 0, \tag{14}$$

δηλαδή όταν οι δύο παρατηρήσεις έχουν την ίδια διάταξη.

Το ζεύγος (x_i, y_i), (x_j, y_j) θωρείται discordant αν:

$$(x_i, y_i) - \left(x_j, y_j\right) < 0, \tag{15}$$

δηλαδή όταν οι δύο παρατηρήσεις έχουν αντίθετη διάταξη.

Η **συνδιακύμανση (covariance)** είναι ένα βασικό στατιστικό μέτρο που εκτιμά τον βαθμό στον οποίο δύο τυχαίες μεταβλητές X και Y μεταβάλλονται μαζί. Εάν οι δύο μεταβλητές τείνουν να

αυξάνονται ή να μειώνονται ταυτόχρονα, τότε η συνδιακύμανση είναι θετική· αν η μία αυξάνεται ενώ η άλλη μειώνεται, είναι αρνητική. Για δύο μεταβλητές $X = \{x_1, x_2, \dots, x_n, \}$ και $Y = \{y_1, y_2, \dots, y_n, \}$, , η συνδιακύμανση, $Cov(X, Y)$, ορίζεται ως εξής:

$$Cov(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \tag{16}$$

όπου $\bar{x}, \bar{y}$ είναι οι μέσοι όροι των X, Y, n είναι το πλήθος των παρατηρήσεων. Όταν $Cov(X, Y) > 0$ οι μεταβλητές τείνουν να αυξάνονται μαζί (θετική συσχέτιση). Όταν $Cov(X, Y) < 0$ σημαίνει ότι όταν η μία αυξάνεται, η άλλη τείνει να μειώνεται (αρνητική συσχέτιση). Όταν $Cov(X, Y) \approx 0$ δεν υπάρχει καμία γραμμική συσχέτιση (ανεξαρτησία δεν συνεπάγεται απαραίτητα).

Η **απόσταση Levenshtein** (Levenshtein distance ή edit distance) είναι ένα κλασικό μέτρο ομοιότητας/απόστασης που χρησιμοποιείται για τη σύγκριση αλφαριθμητικών ακολουθιών. Υπολογίζει τον ελάχιστο αριθμό λειτουργιών επεξεργασίας που απαιτούνται για να μετατραπεί μια συμβολοσειρά σε μια άλλη. Οι επιτρεπόμενες λειτουργίες είναι: (i) εισαγωγή (insertion) ενός χαρακτήρα, (ii) διαγραφή (deletion) ενός χαρακτήρα, (iii) αντικατάσταση (substitution) ενός χαρακτήρα με έναν άλλο. Έστω δύο συμβολοσειρές a και b μήκους m και n αντίστοιχα. Η απόσταση Levenshtein $D(m, n)$ ορίζεται αναδρομικά ως:

$$D(i, j) = \begin{cases} i, \alpha\nu\ j = 0 \\ j, \alpha\nu\ i = 0 \\ min\begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \delta \end{cases} \end{cases}, \tag{17}$$

όπου $\delta = 0$ αν $a_i = b_j$, διαφορετικά $\delta = 1$. Η μέγιστη απόσταση ισούται με το $max(m, n)$. Αποτελεί ευέλικτο μέτρο ομοιότητας για αλφαριθμητικά (strings) αλλά είναι υπολογιστικά ακριβό για πολύ μεγάλες ακολουθίες (πολυπλοκότητα $O(mn)$).

Η **απόσταση Jaro** (Jaro distance) είναι ένα μέτρο ομοιότητας που χρησιμοποιείται για τη σύγκριση δύο αλφαριθμητικών συμβολοσειρών. Βασίζεται στον αριθμό των κοινών χαρακτήρων και στον αριθμό των μεταθέσεων (transpositions). Για δύο συμβολοσειρές s_1, s_2, η απόσταση Jaro ορίζεται ως εξής:

$$J = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{|m|}\right), \tag{18}$$

όπου m είναι ο αριθμός κοινών χαρακτήρων, t είναι ο αριθμός "μισών μεταθέσεων" (δηλ. χαρακτήρες κοινοί αλλά σε διαφορετική θέση). Η απόσταση Jaro κυμαίνεται από 0 (καμία ομοιότητα), έως 1 (απόλυτη ταύτιση).

Η **απόσταση Jaro-Winkler** (Jaro-Winkler distance) είναι παραλλαγή της Jaro που δίνει επιπλέον βάρος στο κοινό πρόθεμα (prefix) των δύο strings. Αυτό την καθιστά ιδιαίτερα χρήσιμη σε εφαρμογές όπου μικρές παραλλαγές (π.χ. ορθογραφικά λάθη) είναι συχνές. Ορίζεται ως:

$$JW = J + (lp(1 - J)), \tag{19}$$

όπου J είναι η απόσταση Jaro, l το μήκος κοινού προθέματος (μέχρι 4 χαρακτήρες), p η παράμετρος scaling (συνήθως $p = 0.1$). Η μετρική αυτή δίνει υψηλότερη βαθμολογία σε strings που ξεκινούν με τα ίδια γράμματα.

Στον Πίνακα 2 συνοψίζονται τα πλεονεκτήματα και οι περιορισμοί των υποστηριζόμενων μεθόδων εντοπισμού ομοιότητας μεταξύ των μεταβλητών.

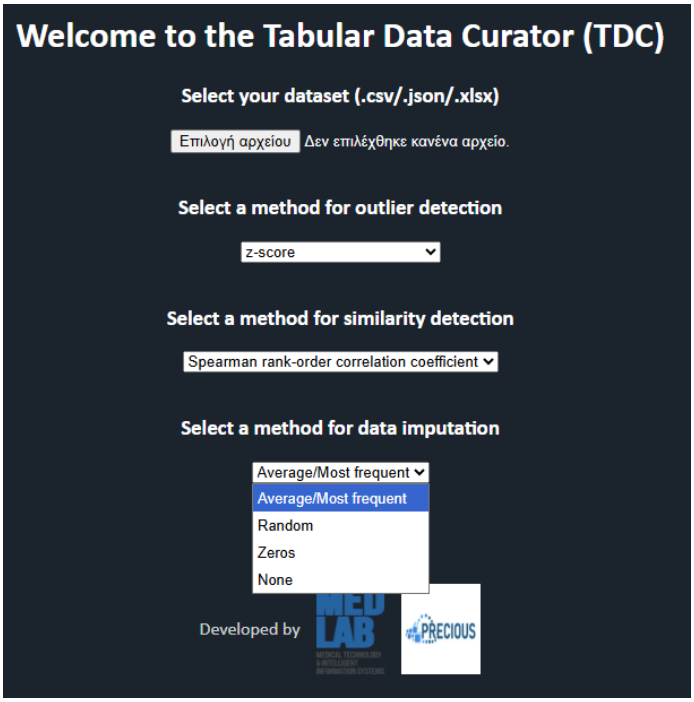
Πίνακας 3: Υποστηριζόμενες μέθοδοι εντοπισμού ομοιότητας μεταξύ των μεταβλητών.

Μέθοδος	**Βασική Ιδέα**	**Πλεονεκτήματα**	**Περιορισμοί**
Pearson Correlation	Γραμμική σχέση δύο μεταβλητών (κανονικοποιημένη συνδιακύμανση)	Απλή, κατανοητή, ευρέως χρησιμοποιούμενη	Ανιχνεύει μόνο γραμμικές σχέσεις, ευαίσθητη σε outliers
Spearman Rank Correlation	Συσχέτιση βάσει σειροθετήσεων (ranks)	Δεν απαιτεί κανονικότητα, ανιχνεύει μονοτονικές μη γραμμικές σχέσεις	Ευαίσθητη σε πολλές ισοβαθμίες, μικρότερη ισχύς σε τέλεια γραμμικές σχέσεις
Kendall's Tau	Μέτρηση συμφωνίας/διαφωνίας σε ζεύγη σειροθετήσεων	Ανθεκτικό σε outliers, καλό για μικρά δείγματα	Υπολογιστικά βαρύ για μεγάλα δείγματα, λιγότερο ευαίσθητο σε μικρές αλλαγές

Μέθοδος	Βασική Ιδέα	Πλεονεκτήματα	Περιορισμοί
Covariance	Μέτρο ταυτόχρονης μεταβολής δύο μεταβλητών	Θεμελιώδες στατιστικό μέτρο, βάση του Pearson	Μη κανονικοποιημένο, εξαρτάται από τις μονάδες
Levenshtein Distance	Ελάχιστος αριθμός πράξεων (insert, delete, substitute) για μετατροπή string	Ευέλικτη για strings, NLP, βιοπληροφορική	Υπολογιστικά ακριβή για μεγάλες συμβολοσειρές
Jaro Distance	Ομοιότητα βάσει κοινών χαρακτήρων και μεταθέσεων	Καλή για σύντομα strings, ανθεκτική σε τυπογραφικά λάθη	Λιγότερο κατάλληλη για μεγάλες συμβολοσειρές
Jaro-Winkler	Jaro με βάρος στο κοινό πρόθεμα	Βελτιωμένη ακρίβεια σε ονόματα, data linkage	Εξειδικευμένη, δεν είναι γενικής χρήσης

Μέθοδοι διαχείρισης ελλειπουσών τιμών (data imputation)

Η παρουσία ελλειπουσών τιμών (missing values) αποτελεί συχνό πρόβλημα σε πραγματικά datasets. Οι μέθοδοι imputation στοχεύουν στην αντικατάσταση των ελλειπουσών τιμών με εκτιμήσεις που διατηρούν τη συνοχή και τη χρησιμότητα του dataset. Η επιλογή της κατάλληλης μεθόδου εξαρτάται από το είδος των δεδομένων, το ποσοστό των ελλείψεων και τη σημασιολογική ερμηνεία που μπορεί να έχει η απουσία τιμών. Στην Εικόνα 13 παρουσιάζονται οι βασικές τεχνικές imputation που υποστηρίζονται, οι οποίες περιλαμβάνουν: (i) Στατιστικές μέθοδοι: Average / Median Imputation, όπου οι ελλείπουσες τιμές αντικαθίστανται με μέτρα κεντρικής τάσης, (ii) Στοχαστικές μεθόδους: Random Imputation, με δειγματοληψία από την υπάρχουσα κατανομή των δεδομένων, (iii) Απλές ευρετικές μεθόδους: Zero Imputation, όταν το μηδέν έχει σαφή σημασιολογική ερμηνεία.



Εικόνα 13: Οι υποστηριζόμενες τεχνικές για την διαχείριση ελλειπουσών τιμών.

Η απλούστερη και πιο διαδεδομένη μέθοδος είναι η **Average/Median imputation** βάσει της οποίας οι ελλείπουσες τιμές των συνεχών μεταβλητών αντικαθίστανται με τον μέσο όρο της μεταβλητής ενώ οι ελλείπουσες τιμές των διακριτών μεταβλητών αντικαθίστανται με την διάμεσο που είναι πιο ανθεκτική σε outliers. Αποτελεί μια απλή και γρήγορη μέθοδο η οποία διατηρεί το μέγεθος του δείγματος αλλά μειώνει τη διακύμανση και μπορεί να αλλοιώσει σημαντικά τις συσχετίσεις μεταξύ των μεταβλητών.

Στην μέθοδο **Random Imputation**, οι ελλείπουσες τιμές αντικαθίστανται με τυχαίες τιμές που επιλέγονται από την υπάρχουσα κατανομή της μεταβλητής. Ουσιαστικά για κάθε ελλειπή τιμή γίνεται δειγματοληψία από τις παρατηρούμενες τιμές. Η μέθοδος αυτή διατηρεί την κατανομή και τη διακύμανση της μεταβλητής αλλά εισάγει τυχαιότητα, τα αποτελέσματα μπορεί να διαφέρουν ανά εκτέλεση και μπορεί να μην διατηρήσει τις συσχετίσεις με άλλες μεταβλητές.

Στην **μέθοδο Zero Imputation**, οι ελλείπουσες τιμές αντικαθίστανται με την τιμή μηδέν (0). Χρησιμοποιείται κυρίως σε περιπτώσεις όπου το μηδέν έχει σαφή σημασιολογική ερμηνεία (π.χ. ποσότητα = 0, απουσία χαρακτηριστικού). Είναι πολύ απλή και καλή για sparse δεδομένα (π.χ. one-hot encoding, πίνακες συχνοτήτων) αλλά μπορεί να εισαγάγει συστηματική μεροληψία αν το μηδέν δεν έχει νόημα για τη μεταβλητή και να μειώσει την ακρίβεια σε αριθμητικά δεδομένα.

Στον Πίνακα 3 συνοψίζονται τα πλεονεκτήματα και οι περιορισμοί των υποστηριζόμενων μεθόδων διαχείρισης ελλειπουσών τιμών.

Πίνακας 4: Υποστηριζόμενες μέθοδοι διαχείρισης ελλειπουσών τιμών.

Μέθοδος	**Βασική Ιδέα**	**Πλεονεκτήματα**	**Περιορισμοί**
Average Imputation	Αντικατάσταση με τον μέσο όρο της μεταβλητής	Απλή, γρήγορη, διατηρεί το μέγεθος δείγματος	Μειώνει τη διακύμανση, αλλοιώνει συσχετίσεις
Median Imputation	Αντικατάσταση με τη διάμεσο	Ανθεκτική σε outliers, κατάλληλη για skewed δεδομένα	Όπως και ο μέσος, αλλοιώνει σχέσεις μεταξύ μεταβλητών
Random Imputation	Αντικατάσταση με τυχαία τιμή από την υπάρχουσα κατανομή	Διατηρεί την κατανομή και διακύμανση	Εισάγει τυχαιότητα, δεν διατηρεί συσχετίσεις με άλλες μεταβλητές
Zero Imputation	Αντικατάσταση με 0	Πολύ απλή, καλή για sparse δεδομένα	Μπορεί να εισάγει bias, ακατάλληλη αν το 0 δεν έχει σημασιολογική ερμηνεία

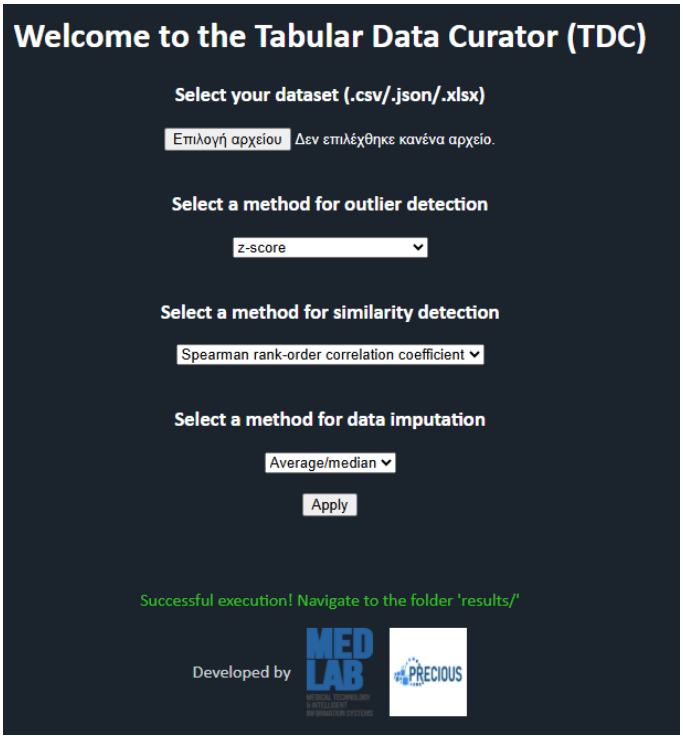
4.1.1.2. Έξοδος

Η υπηρεσία παρέχει ως έξοδο[9]:

- Αναφορά ποιότητας των δεδομένων που περιλαμβάνει (**Προδιαγραφή 2.5**):
    - το πλήθος των μεταβλητών,
    - το πλήθος των δειγμάτων,
    - το πλήθος των διακριτών μεταβλητών,
    - το πλήθος των συνεχών μεταβλητών,
    - το πλήθος των μεταβλητών με άγνωστο τύπο δεδομένων,
    - το πλήθος των ελλειπουσών τιμών (με ποσοστό επί του συνόλου),
    - την ονομασία της κάθε μεταβλητής,
    - το εύρος τιμών της κάθε μεταβλητής,
    - τον τύπο δεδομένων της κάθε μεταβλητής,

- το πλήθος των ελλειπουσών τιμών της κάθε μεταβλητής,
- την ποιότητα της κάθε μεταβλητής (καλή ή μέτρια ή κακή),
- την μέση ή διάμεση τιμή της κάθε μεταβλητής,
- την παρουσία έκτοπων τιμών ανά μεταβλητή,
- την παρουσία ασυμβατοτήτων ανά μεταβλητή.

- Το αρχείο δεδομένων εισόδου, όπου (**Προδιαγραφή 2.5**):
  - οι ελλείπουσες τιμές επισημαίνονται με γκρι χρώμα και ένδειξη,
  - οι έκτοπες τιμές επισημαίνονται με πορτοκαλί χρώμα,
  - τα πεδία με ασυμβατότητες επισημαίνονται με κόκκινο χρώμα,
  - οι μεταβλητές με καλή ποιότητα επισημαίνονται με μπλε χρώμα,
  - οι μεταβλητές με μέτρια ποιότητα επισημαίνονται με πράσινο χρώμα,
  - οι μεταβλητές με κακή ποιότητα επισημαίνονται με κόκκινο χρώμα.



Εικόνα 14: Στιγμιότυπο από την επιτυχή εκτέλεση της υπηρεσίας.

Η Εικόνα 14 παρουσιάζει ένα ενδεικτικό στιγμιότυπο από την αναφορά ποιότητας των δεδομένων, όπως αυτή παράγεται από την αντίστοιχη υπηρεσία ελέγχου σε μορφή πίνακα. Η αναφορά αυτή συνοψίζει κρίσιμα μεταδεδομένα για το σύνολο δεδομένων και παρέχει

αξιολογήσεις για κάθε μεμονωμένη μεταβλητή, καθιστώντας εμφανή τα προβλήματα που σχετίζονται με ελλείπουσες τιμές, τύπους δεδομένων, έκτοπες τιμές και ασυμβατότητες. Στο επάνω μέρος της εικόνας παρατίθενται συγκεντρωτικά στατιστικά μεταδεδομένα, όπως το πλήθος των μεταβλητών (42), των δειγμάτων (325), ο διαχωρισμός των μεταβλητών σε διακριτές, συνεχείς και άγνωστου τύπου, καθώς και το ποσοστό των ελλειπουσών τιμών (15,11%). Επιπλέον, γίνεται ποσοτικοποίηση της ποιότητας των μεταβλητών, με βάση το ποσοστό απουσίας δεδομένων, κατατάσσοντάς τις σε καλής, μέτριας ή κακής ποιότητας. Το κύριο μέρος της αναφοράς περιλαμβάνει πίνακα αξιολόγησης ανά μεταβλητή. Για κάθε χαρακτηριστικό παρουσιάζονται το όνομα, το εύρος τιμών, ο τύπος και ο υποτύπος μεταβλητής, ο αριθμός των ελλειπουσών τιμών, η συνολική ποιότητα, η ύπαρξη ή μη έκτοπων τιμών, καθώς και πιθανές ασυμβατότητες. Η πληροφορία αυτή υποστηρίζεται οπτικά μέσω χρωματικής κωδικοποίησης: οι μεταβλητές καλής ποιότητας επισημαίνονται με μπλε, οι μέτριας με πράσινο και οι κακής με κόκκινο, ενώ έκτοπες τιμές και ασυμβατότητες τονίζονται με ροζ και μωβ αποχρώσεις αντίστοιχα. Χαρακτηριστικά παραδείγματα περιλαμβάνουν τη μεταβλητή AGE, η οποία φέρει άγνωστο τύπο, 19 ελλείπουσες τιμές και ταξινομείται ως μέτριας ποιότητας, με σαφή ένδειξη ασυμβατότητας. Η μεταβλητή JOB παρουσιάζει πλήρη απουσία τιμών (325/325), κατατάσσεται ως κακής ποιότητας και συνδέεται με προβλήματα τύπου και πληρότητας. Παρομοίως, η SUPRARENAL ANGLE φέρει υψηλό ποσοστό ελλείψεων και άγνωστο τύπο, οδηγώντας σε επισήμανση χαμηλής ποιότητας και δομικής ασυμβατότητας. Η αναφορά αυτή αποτελεί βασικό εργαλείο για την αξιολόγηση της καταλληλότητας των δεδομένων προς ανάλυση και προετοιμάζει το έδαφος για περαιτέρω βήματα, όπως η διαχείριση ελλειπουσών τιμών, η εναρμόνιση των τύπων δεδομένων και η βελτίωση της διαλειτουργικότητας στο πλαίσιο ομοσπονδιακών υποδομών.

Metadata

Number of feature(s)	42
Number of instance(s)	325
Discrete feature(s)	22
Continuous feature(s)	8
Unknown feature(s)	12
Missing values (%)	15.11%
Good feature(s) (%)	1 (2.4%)
Fair feature(s) (%)	25 (59.5%)
Bad feature(s) (%)	16 (38.1%)
Outlier(s) (%)	60 (0.44%)

Quality assessment

Features	Value range	Type	Variable type	Missing values	State	Outliers	Incompatibilities
ID	Too large to display!!	categorical	string	0	good	not-applicable	no
AGE	, 63, 65, 85, =2024-1949, =2024-1953, =2024-1945, =2024-1937, =2024-1958, =2024-1947, =2024-19	unknown	unknown	19	fair	not-applicable	yes, unknown type of data
GENDER	[0, 1]	categorical	int	18	fair	no	no
SMOKING	(1, , 0)	unknown	unknown	31	fair	not-applicable	yes, unknown type of data
PACK-YEARS	[0, 200]	numeric	date	51	fair	yes	no
JOB	(None,)	categorical	string	325	bad	not-applicable	yes, bad feature
AAA STATUS	[0, 2]	categorical	int	36	fair	yes	no
AAA MORHOLOGY	[0, 3]	categorical	int	67	fair	yes	no
AAA AETIOLOGY	[0, 1]	categorical	int	68	fair	no	no
AAA LOCATION	[1, 4]	categorical	int	67	fair	yes	no
AAA LENGTH	[4.4, 13.5]	numeric	float	308	bad	no	yes, bad feature
MAX DIAMETER	[0.5, 13.5]	numeric	float	73	fair	yes	no
NECK DIAMETER	[0, 3.5]	numeric	float	107	fair	yes	no
NECK LENGTH	[0, 6.3]	numeric	float	108	fair	no	no
NECK SHAPE	[0, 3]	categorical	int	128	bad	yes	yes, bad feature
SUPRARENAL ANGLE	, 60, 30, 53, =180-147.4, 10, 42, 47, =180-147, 40, 50, 24, 26, 43, 45, 33, 20, 15, 5.1, 28, 65, 22, 16, 3	unknown	unknown	252	bad	not-applicable	yes, unknown type of data
INFRARENAL ANGLE	65, 17, =180-161.8, 25, 30.4, =147-90, =180-151, 32, 95, 24, 68, 50, 23, 47, 33, 46, 26, 71, 57, 14, 63,	unknown	unknown	189	bad	not-applicable	yes, unknown type of data
ILT	[0, 2]	categorical	int	72	fair	no	no
CALCIFICATIONS	[0, 2]	categorical	float	90	fair	no	no
RCIA DIAMETER	1.6, 1.7, 1.9, 2.25, 1.15, 7, 3.2, 2.2, 2.4, 1.1, 1.38, 2.1, 3.6, 1.05, 1.45, 0.9, 1.25, 6.7, 3.76, 5.3, 2.5, 2.	unknown	unknown	82	fair	not-applicable	yes, unknown type of data
RCIA LENGTH	2.7, 3.2, 3, 4.5, 7.1, 6.7, 5.3, 7.5, 4.2, 6.6, 10.8, 6.3, 9.3, 4.6, 3.6, 4.4, 5.5, 3.3, 3.7, 10, 3.65, 8.6, 5.2,	unknown	unknown	100	fair	not-applicable	yes, unknown type of data
LCIA DIAMETER	4, 2, 1.4, 1.2, 3.2, 1.9, 0.8, 6, 1, 4.5, 1.18, 2.3, 1.05, 1.45, 1.25, 1.15, 1.35, 2.5, 1.8, 2.25, 7, 3.6, 2.6,	unknown	unknown	80	fair	not-applicable	yes, unknown type of data
LCIA LENGTH	7.3, 6, 5, 7.7, 5.8, 6.3, 6.45, 5.2, 7.1, 8.5, 8, 6.1, 4, 3.4, 8.4, 8.8, 4.4, 5.4, 4.1, 10.6, 7.5, 3.2, 9.5, 2.15,	unknown	unknown	101	fair	not-applicable	yes, unknown type of data
FAMILY HISTORY	(None,)	categorical	string	325	bad	not-applicable	yes, bad feature
SYSTOLIC PRESSURE	(None,)	categorical	string	325	bad	not-applicable	yes, bad feature

Εικόνα 15: Στιγμιότυπο από την αναφορά ποιότητας των δεδομένων.

Η Εικόνα 15 απεικονίζει το αρχείο εισόδου του συνόλου δεδομένων, όπως αυτό επεξεργάζεται από την υπηρεσία ελέγχου ποιότητας. Χρησιμοποιείται κατάλληλος χρωματικός κώδικας για την επισήμανση των κρίσιμων ποιοτικών χαρακτηριστικών των δεδομένων, διευκολύνοντας

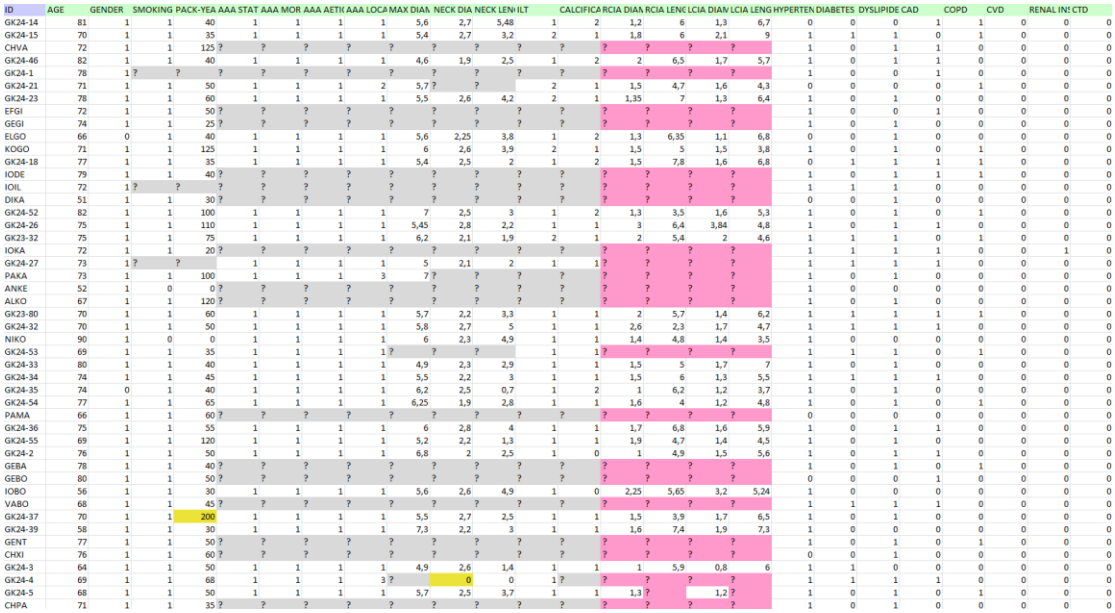
την άμεση αναγνώριση προβληματικών τιμών και μεταβλητών. Συγκεκριμένα, οι ελλείπουσες τιμές επισημαίνονται με γκρι χρώμα και το σύμβολο ?, υποδηλώνοντας απουσία πληροφορίας σε συγκεκριμένα πεδία. Οι έκτοπες τιμές (outliers) τονίζονται με κίτρινο χρώμα, αναδεικνύοντας τιμές που αποκλίνουν σημαντικά από την αναμενόμενη κατανομή των αντίστοιχων μεταβλητών, βάσει μονοπαραμετρικών ή πολυπαραμετρικών τεχνικών εντοπισμού. Οι ασυμβατότητες στις τιμές – όπως μη έγκυρα ή ανομοιογενή δεδομένα ως προς τον προσδοκώμενο τύπο – επισημαίνονται με ροζ φόντο. Χαρακτηριστικά παραδείγματα περιλαμβάνουν μεταβλητές όπως η SUPRARENAL ANGLE και η INFRARENAL ANGLE, στις οποίες παρατηρείται μεγάλος αριθμός μη αναγνωρίσιμων ή εσφαλμένων τιμών. Η επισήμανση αυτή είναι ιδιαίτερα χρήσιμη για τη διασφάλιση της συνέπειας και της εναρμονισμένης αναπαράστασης των δεδομένων σε περιβάλλον ομοσπονδιακών βάσεων. Τέλος, η ποιότητα των μεταβλητών εκφράζεται επίσης έμμεσα μέσω του χρωματισμού: μεταβλητές με υψηλό ποσοστό ελλειπουσών ή ασυνεπών τιμών καθίστανται εύκολα αναγνωρίσιμες, διευκολύνοντας την απόφαση για την αποδοχή, διόρθωση ή απόρριψή τους κατά τη φάση της προτυποποίησης και εναρμόνισης. Η συγκεκριμένη οπτική απεικόνιση αποτελεί βασικό εργαλείο για την ταχεία αξιολόγηση της ποιότητας των δεδομένων και ενισχύει τη διαφάνεια και την αναγνωσιμότητα της κατάστασης κάθε μεταβλητής, παρέχοντας χρήσιμη υποστήριξη σε επιστήμονες και διαχειριστές δεδομένων.

ID	AGE	GENDER	SMOKING	PACK-YEARS	JOB	AAA STATUS	AAA MORHOLOGY	AAA AETIOLOGY	AAA LOCATION	AAA LENGTH	MAX DIAMETER	NECK DIAMETER	NECK LENGTH	NECK SHAPE	SUPRARENAL ANGLE	INFRARENAL ANGLE	ILT	CALCIFICATIONS	RCIA DIAMETER	RCIA LENGTH
GK24-14	81	1	1	40	?	1	1	1	1	?	5,6	2,7	5,48	1	?	?	1	2	1,2	6
GK24-15	70	1	1	35	?	1	1	1	1	?	5,4	2,7	3,2	1	?	35	2	1	1,8	6
CHVA	72	1	1	125	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
GK24-46	82	1	1	40	?	1	1	1	1	?	4,6	1,9	2,5	1	?	?	1	2	2	6,5
GK24-1	78	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
GK24-21	71	1	1	50	?	1	1	1	2	?	5,7	?	?	?		0	?	2	1	1,5	4,7
GK24-23	78	1	1	60	?	1	1	1	1	?	5,5	2,6	4,2	1	?	40	2	1	1,35	7
EFGI	72	1	1	50	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
GEGI	74	1	1	25	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
ELGO	66	0	1	40	?	1	1	1	1	?	5,6	2,25	3,8	1	?	40	1	2	1,3	6,35
KOGO	71	1	1	125	?	1	1	1	1	?	6	2,6	3,9	1	?	30	2	1	1,5	5
GK24-18	77	1	1	35	?	1	1	1	1	?	5,4	2,5	2	1	?	30	1	2	1,5	7,8
IODE	79	1	1	40	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
IOIL	72	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
DIKA	51	1	1	30	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
GK24-52	82	1	1	100	?	1	1	1	1	?	7	2,5	3	1		64	60	1	2	1,3	3,5
GK24-26	75	1	1	110	?	1	1	1	1	?	5,45	2,8	2,2	1		80	72	1	1	3	6,4
GK23-32	75	1	1	75	?	1	1	1	1	?	6,2	2,1	1,9	1		36	70	2	1	2	5,4
IOKA	72	1	1	20	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
GK24-27	73	1	?	?	?	1	1	1	1	?	5	2,1	2	1		60	70	1	1	?	?
PAKA	73	1	1	100	?	1	1	1	1	?	7	?	?	?	?	?	?	?	?	?	
ANKE	52	1	0	0	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
ALKO	67	1	1	120	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
GK23-80	70	1	1	60	?	1	1	1	1	?	5,7	2,2	3,3	1	?	34	1	1	2	5,7	
GK24-32	70	1	1	50	?	1	1	1	1	?	5,8	2,7	5	1	30	20	1	1	2,6	2,3	
NIKO	90	1	0	0	?	1	1	1	1	?	6	2,3	4,9	1	?	?	1	1	1,4	4,8	
GK24-53	69	1	1	35	?	1	1	1	1	?	?	?	?	?	?	?	1	1	?	?	
GK24-33	80	1	1	40	?	1	1	1	1	?	4,9	2,3	2,9	1	53	44	1	1	1,5	5	
GK24-34	74	1	1	45	?	1	1	1	1	?	5,5	2,2	3	1	?	?	1	1	1,5	6	
GK24-35	74	0	1	40	?	1	1	1	1	?	6,2	2,5	0,7	2	?	30	1	2	1	6,2	
GK24-54	77	1	1	65	?	1	1	1	1	?	6,25	1,9	2,8	?	?	55	1	1	1,6	4	
PAMA	66	1	1	60	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
GK24-36	75	1	1	55	?	1	1	1	1	?	6	2,8	4	1	?	45	1	1	1,7	6,8	
GK24-55	69	1	1	120	?	1	1	1	1	?	5,2	2,2	1,3	1	?	65	1	1	1,9	4,7	
GK24-2	76	1	1	50	?	1	1	1	1	?	6,8	2	2,5	1	?	?	1	0	1	4,9	
GEBA	78	1	1	40	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
GEBO	80	1	1	50	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
IOBO	56	1	1	30	?	1	1	1	1	?	5,6	2,6	4,9	1	?	17	1	0	2,25	5,65	
VABO	68	1	1	45	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
GK24-37	70	1	1	200	?	1	1	1	1	?	5,5	2,7	2,5	1	?	?	1	1	1,5	3,9	
GK24-39	58	1	1	30	?	1	1	1	1	?	7,3	2,2	3	1	?	?	1	1	1,6	7,4	
GENT	77	1	1	50	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
CHXI	76	1	1	60	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	
GK24-3	64	1	1	50	?	1	1	1	1	?	4,9	2,6	1,4	1	32,6	18,2	1	1	1	5,9	
GK24-4	69	1	1	68	?	1	1	1	3	?	?	0	0	1	?	?	1	?	?		
GK24-5	68	1	1	50	?	1	1	1	1	?	5,7	2,5	3,7	1	?	?	1	1	1,3	?	

Εικόνα 16: Στιγμιότυπο από το αρχείο δεδομένων εισόδου με κατάλληλο χρωματικό κώδικα για την επισήμανση των ελλειπουσών τιμών, έκτοπων τιμών, πεδίων με ασυμβατότητες, μεταβλητών με καλή/μέτρια/κακή ποιότητας.

Επιπλέον στην έξοδο δίνεται και έναν επιπλέον αρχείο στο οποίο έχουν αφαιρεθεί αυτόματα οι μεταβλητές με κακή ποιότητα. Η Εικόνα 16 απεικονίζει ένα τέτοιο στιγμιότυπο από το αρχείο εισόδου των δεδομένων μετά την ενσωμάτωση του κατάλληλου χρωματικού κώδικα, ο οποίος διευκολύνει την αξιολόγηση της ποιότητάς τους με βάση προκαθορισμένα κριτήρια. Συγκεκριμένα, το στιγμιότυπο εστιάζει στην επισήμανση κρίσιμων ζητημάτων, όπως οι

ελλείπουσες τιμές, οι έκτοπες τιμές, οι ασυμβατότητες τιμών και η ποιότητα των μεταβλητών. Οι ελλείπουσες τιμές υποδεικνύονται με γκρι χρώμα και το σύμβολο ?, επιτρέποντας την άμεση αναγνώριση περιοχών του πίνακα που χρήζουν συμπλήρωσης ή περαιτέρω επεξεργασίας. Οι έκτοπες τιμές, δηλαδή παρατηρήσεις που αποκλίνουν σημαντικά από τη συνηθισμένη κατανομή, εμφανίζονται με κίτρινο φόντο και εντοπίζονται σε μεταβλητές όπως η PACK-YEARS, η NECK LENGTH και η LCIA LENGTH. Οι ασυμβατότητες τιμών – που συνήθως αφορούν λανθασμένη μορφή, απρόσμενους χαρακτήρες ή τιμές που δεν συνάδουν με το αναμενόμενο εύρος – επισημαίνονται με ροζ φόντο σε μεταβλητές όπως οι RCIA DIAMETER, LCIA LENGTH και LCIA DIAMETER, προσδιορίζοντας την ανάγκη εναρμόνισης ως προς την αναμενόμενη δομή ή μονάδα μέτρησης. Στο δεξί τμήμα της εικόνας περιλαμβάνονται δυαδικές μεταβλητές (π.χ. HYPERTEN, DIABETES, CVD, COPD, RENAL, CTD), οι οποίες παρουσιάζουν υψηλή πληρότητα και ποιότητα, χωρίς εμφανείς ελλείψεις ή ασυμβατότητες. Ο συνολικός χρωματικός κώδικας παρέχει εποπτική και γρήγορη κατανόηση της κατάστασης κάθε στήλης, υποβοηθώντας την τεκμηριωμένη λήψη αποφάσεων ως προς την αποδοχή, τον καθαρισμό ή τη μετασχηματιστική επεξεργασία των δεδομένων. Η οπτική αυτή αναπαράσταση συνιστά πολύτιμο εργαλείο για την υποστήριξη της διαδικασίας προτυποποίησης και καθαρισμού των δεδομένων, προτού αυτά αξιοποιηθούν για αναλυτικούς ή μηχανικής μάθησης σκοπούς στο πλαίσιο ομοσπονδιακής ανάλυσης.



ID	AGE	GENDER	SMOKING	PACK-YEA	AAA STAT	AAA MOR	AAA AETI	AAA LOCA	MAX DIAN	NECK DIA	NECK LEN	ILT	CALCIFICA	RCIA DIAN	RCIA LEN	LCIA DIAN	LCIA LENG	HYPERTEN	DIABETES	DYSLIPIDE	CAD	COPD	CVD	RENAL IN	CTD
GK24-14	81	1	1	40	1	1	1	1	5,6	2,7	5,48	1	2	1,2	6	1,3	6,7	0	0	0	1	1	0	0	0
GK24-15	70	1	1	35	1	1	1	1	5,4	2,7	3,2	2	1	1,8	6	2,1	9	1	1	1	0	1	0	0	0
CHVA	72	1	1	125	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	1	1	0	0	0	0
GK24-46	82	1	1	40	1	1	1	1	4,6	1,9	2,5	1	2	2	6,5	1,7	5,7	1	0	1	1	0	0	0	0
GK24-1	78	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	0	1	0	0	0	0
GK24-21	71	1	1	50	1	1	1	2	5,7	?	?	2	1	1,5	4,7	1,6	4,3	0	0	0	0	1	0	0	0
GK24-23	78	1	1	60	1	1	1	1	5,5	2,6	4,2	2	1	1,35	7	1,3	6,4	1	0	1	0	0	0	0	0
EFGI	72	1	1	50	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	0	1	0	0	0	0
GEGI	74	1	1	25	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	1	0	0	0	0	0
ELGO	66	0	1	40	1	1	1	1	5,6	2,25	3,8	1	2	1,3	6,35	1,1	6,8	0	0	1	0	0	0	0	0
KOGO	71	1	1	125	1	1	1	1	6	2,6	3,9	2	1	1,5	5	1,5	3,8	1	0	1	0	1	0	0	0
GK24-18	77	1	1	35	1	1	1	1	5,4	2,5	2	1	2	1,5	7,8	1,6	6,8	0	1	1	1	1	0	0	0
IODE	79	1	1	40	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	1	1	1	0	0	0
IOIL	72	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	1	1	0	0	0	0	0
DIKA	51	1	1	30	?	?	?	?	?	?	?	?	?	?	?	?	?	0	0	1	0	0	0	0	0
GK24-52	82	1	1	100	1	1	1	1	7	2,5	3	1	2	1,3	3,5	1,6	5,3	1	0	1	0	1	0	0	0
GK24-26	75	1	1	110	1	1	1	1	5,45	2,8	2,2	1	1	3	6,4	3,84	4,8	1	0	1	1	0	0	0	0
GK23-32	75	1	1	75	1	1	1	1	6,2	2,1	1,9	2	1	2	5,4	2	4,6	1	1	1	0	1	0	0	0
IOKA	72	1	1	20	?	?	?	?	?	?	?	?	?	?	?	?	?	1	1	1	1	0	0	1	0
GK24-27	73	1	?	?	1	1	1	1	5	2,1	2	1	1	?	?	?	?	1	1	1	1	0	0	0	0
PAKA	73	1	1	100	1	1	1	3	7	?	?	?	?	?	?	?	?	1	0	1	0	0	0	0	0
ANKE	52	1	0	0	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	0	0	0	0	0	0
ALKO	67	1	1	120	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	1	0	0	0	0	0
GK23-80	70	1	1	60	1	1	1	1	5,7	2,2	3,3	1	1	2	5,7	1,4	6,2	1	1	1	1	1	0	0	0
GK24-32	70	1	1	50	1	1	1	1	5,8	2,7	5	1	1	2,6	2,3	1,7	4,7	1	1	1	1	0	0	0	0
NIKO	90	1	0	0	1	1	1	1	6	2,3	4,9	1	1	1,4	4,8	1,4	3,5	1	0	0	0	0	0	0	0
GK24-53	69	1	1	35	1	1	1	1	?	?	?	1	1	?	?	?	?	1	1	1	0	1	0	0	0
GK24-33	80	1	1	40	1	1	1	1	4,9	2,3	2,9	1	1	1,5	5	1,7	7	1	0	1	0	0	0	0	0
GK24-34	74	1	1	45	1	1	1	1	5,5	2,2	3	1	1	1,5	6	1,3	5,5	1	1	1	1	0	0	0	0
GK24-35	74	0	1	40	1	1	1	1	6,2	2,5	0,7	1	2	1	6,2	1,2	3,7	1	0	1	0	0	0	0	0
GK24-54	77	1	1	65	1	1	1	1	6,25	1,9	2,8	1	1	1,6	4	1,2	4,8	1	0	1	0	1	0	0	0
PAMA	66	1	1	60	?	?	?	?	?	?	?	?	?	?	?	?	?	0	0	0	0	0	0	0	0
GK24-36	75	1	1	55	1	1	1	1	6	2,8	4	1	1	1,7	6,8	1,6	5,9	1	0	1	1	0	0	0	0
GK24-55	69	1	1	120	1	1	1	1	5,2	2,2	1,3	1	1	1,9	4,7	1,4	4,5	1	0	1	0	0	0	0	0
GK24-2	76	1	1	50	1	1	1	1	6,8	2	2,5	1	0	1	4,9	1,5	5,6	1	0	1	1	0	0	0	0
GEBA	78	1	1	40	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	1	0	1	0	0	0
GEBO	80	1	1	50	?	?	?	?	?	?	?	?	?	?	?	?	?	0	0	0	1	0	0	0	0
IOBO	56	1	1	30	1	1	1	1	5,6	2,6	4,9	1	0	2,25	5,65	3,2	5,24	1	0	0	0	0	0	0	0
VABO	68	1	1	45	?	?	?	?	?	?	?	?	?	?	?	?	?	1	1	1	1	0	0	0	0
GK24-37	70	1	1	200	1	1	1	1	5,5	2,7	2,5	1	1	1,5	3,9	1,7	6,5	1	0	1	0	0	0	0	0
GK24-39	58	1	1	30	1	1	1	1	7,3	2,2	3	1	1	1,6	7,4	1,9	7,3	1	0	0	0	0	0	0	0
GENT	77	1	1	50	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	1	0	1	0	0	0
CHXI	76	1	1	60	?	?	?	?	?	?	?	?	?	?	?	?	?	0	0	1	0	0	0	0	0
GK24-3	64	1	1	50	1	1	1	1	4,9	2,6	1,4	1	1	1	5,9	0,8	6	1	1	0	0	0	0	0	0
GK24-4	69	1	1	68	1	1	1	3	?	0	0	1	?	?	?	?	?	1	1	1	1	0	0	0	0
GK24-5	68	1	1	50	1	1	1	1	5,7	2,5	3,7	1	1	1,3	?	1,2	?	1	0	1	0	1	0	0	0
CHPA	71	1	1	35	?	?	?	?	?	?	?	?	?	?	?	?	?	1	0	1	0	0	0	0	0

Εικόνα 17: Στιγμιότυπο από το αρχείο δεδομένων εισόδου με κατάλληλο χρωματικό κώδικα για την επισήμανση των ελλειπουσών τιμών, έκτοπων τιμών, πεδίων με ασυμβατότητες, μεταβλητών με καλή και μέτρια ποιότητα.

Η Εικόνα 17 παρουσιάζει ένα στιγμιότυπο από την αναφορά λεκτικής ομοιότητας μεταξύ μεταβλητών. Στο παράδειγμα της εικόνας, καταγράφεται υψηλός βαθμός λεκτικής ομοιότητας (τιμή: 0,9285) μεταξύ των μεταβλητών ANTIDIABETICS και DIABETES. Το αποτέλεσμα αυτό προκύπτει από την εφαρμογή μετρικών λεκτικής ομοιότητας, όπως η απόσταση

Levenshtein ή Jaro-Winkler, που συγκρίνουν την ορθογραφική και σημασιολογική εγγύτητα δύο λεκτικών ακολουθιών. Η πληροφορία αυτή είναι ιδιαίτερα χρήσιμη κατά τη διαδικασία εναρμόνισης και προτυποποίησης μεταβλητών από διαφορετικές πηγές, καθώς συμβάλλει στην ενοποίηση ισοδύναμων χαρακτηριστικών και στην αποφυγή πλεονασμών ή διπλοκαταγραφών στα δεδομένα.

f1	f2	value
ANTIDIABETICS	DIABETES	0,928455189

Εικόνα 18: Στιγμιότυπο από την αναφορά λεκτικής ομοιότητας.

Η Εικόνα 18 παρουσιάζει ένα στιγμιότυπο από την αναφορά ομοιότητας κατανομών που προκύπτει από τη συγκριτική ανάλυση μεταξύ ζευγών μεταβλητών. Στο παράδειγμα της εικόνας, αξιολογείται η ομοιότητα των μεταβλητών LCIA DIAMETER και RCIA DIAMETER, καθώς και των LCIA LENGTH και RCIA LENGTH. Οι αντίστοιχοι συντελεστές ομοιότητας φτάνουν τα 0,9487 και 0,9393, υποδηλώνοντας πολύ υψηλή αντιστοιχία στην κατανομή των τιμών. Η εκτίμηση αυτή πραγματοποιείται με τη χρήση του συντελεστή Spearman για την αποτίμηση της μονοτονικής ή δομικής συσχέτισης μεταξύ των μεταβλητών. Η δυνατότητα αναγνώρισης μεταβλητών με σχεδόν ταυτόσημες κατανομές είναι ιδιαίτερα κρίσιμη για την αποφυγή πλεονασμών, τον εντοπισμό υποκατάστατων χαρακτηριστικών και την υποβοήθηση σε διαδικασίες επιλογής χαρακτηριστικών (feature selection).

f1	f2	value
LCIA DIAMETER	RCIA DIAMETER	0,948717949
LCIA LENGTH	RCIA LENGTH	0,939393939

Εικόνα 19: Στιγμιότυπο από την αναφορά ομοιότητας κατανομών.

4.2 Υπηρεσία εξασφάλισης της διαλειτουργικότητας και της ομογένειάς των δεδομένων (σε μορφή πίνακα) μεταξύ των ομοσπονδιακών βάσεων

Η δεύτερη υπηρεσία που αφορά την εξασφάλιση της διαλειτουργικότητας και της ομογένειάς των δεδομένων (σε μορφή πίνακα) μεταξύ των ομοσπονδιακών βάσεων αποτελείται από δύο μέρη: (i) Μέρος 1 – Εξαγωγή της αναφοράς εναρμόνισης σε επίπεδο μεταδεδομένων, (ii) Μέρος 2 - Τελική διαδικασία εναρμόνισης.

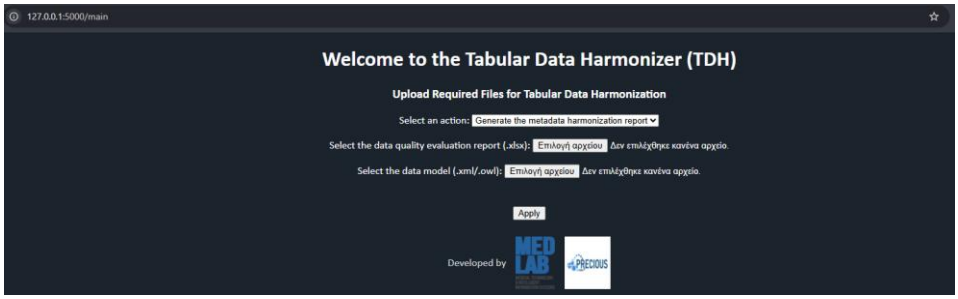
4.2.1. Μέρος 1 – Εξαγωγή της αναφοράς εναρμόνισης δεδομένων (σε επίπεδο μεταδεδομένων)

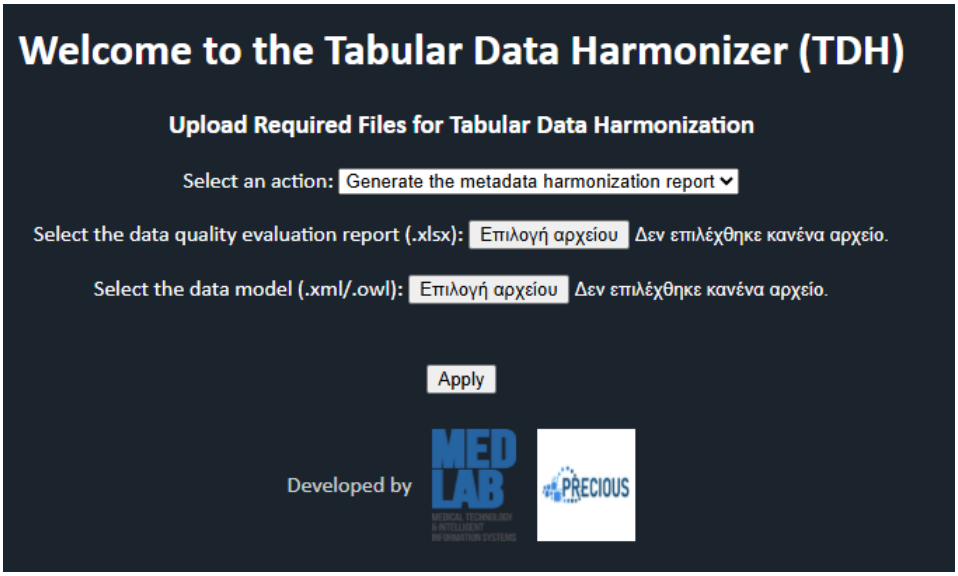
4.2.1.1. Είσοδος

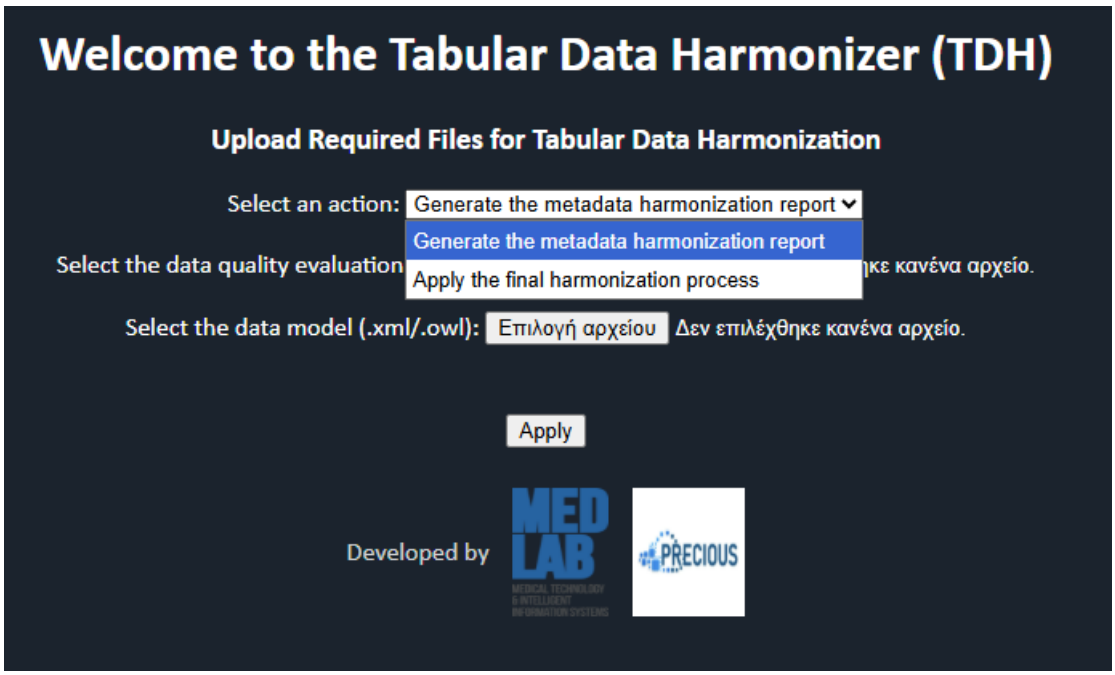
Στο Μέρος 1, η υπηρεσία δέχεται στην είσοδο[14] (**Προδιαγραφή 2.2**) (Εικόνα 20):

- την αναφορά ποιότητας των δεδομένων από την προηγούμενη υπηρεσία και

- ένα ιεραρχικό μοντέλο αναπαράστασης των δεδομένων σε μορφή .XML ή .OWL (που αποτελεί την οντολογία αναφοράς) για την διαδικασία της εναρμόνισης.







Εικόνα 20: Η αρχική σελίδα της υπηρεσίας εξασφάλισης της διαλειτουργικότητας και της ομογένειάς των δεδομένων (σε μορφή πίνακα) μεταξύ των ομοσπονδιακών βάσεων (Μέρος 1).

4.2.1.2. Λειτουργίες

Η υπηρεσία παρέχει λειτουργίες για[14] (**Προδιαγραφή 2.4**):

- την αυτόματη εξαγωγή μεταδεδομένων από την αναφορά ποιότητας δεδομένων συμπεριλαμβανομένων των ονομασιών των μεταβλητών και το εύρος τιμών ανά μεταβλητή,
- την εξαγωγή σημασιολογικής πληροφορίας από την οντολογία αναφοράς αναφορικά με τις σχέσεις μεταξύ των κλάσεων, υποκλάσεων και των μεταβλητών,
- την δημιουργία ενός ενισχυμένου λεξικού ιατρικών ορολογιών το οποίο βασίζεται σε ιατρικές ορολογίες από το διεθνές λεξικό ιατρικών ορολογιών OHDSI Athena[15] (το οποίο περιλαμβάνει μια ευρεία γκάμα ορολογιών από μοντέλα όπως το SNOMED-CT, το LOINC, το ICD-11, το Rx-Norm, κ.α.) και το οποίο εμπλουτίζεται με:
    - word embeddings από τις κλάσεις, υποκλάσεις και μεταβλητές της οντολογίας αναφοράς,
    - συνώνυμες ορολογίες από τις κλάσεις, υποκλάσεις και μεταβλητές της οντολογίας αναφοράς,
    - σημασιολογικές σχέσεις (ιδιότητες αντικειμένων) της οντολογίας αναφοράς,
- την διεξαγωγή ενδελεχούς λεκτικής ανάλυσης με στόχο την ταυτοποίηση των μεταβλητών εισόδου με αυτές του εμπλουτισμένου λεξικού ιατρικών ορολογιών (από το προηγούμενο βήμα).

4.2.1.3. Έξοδος

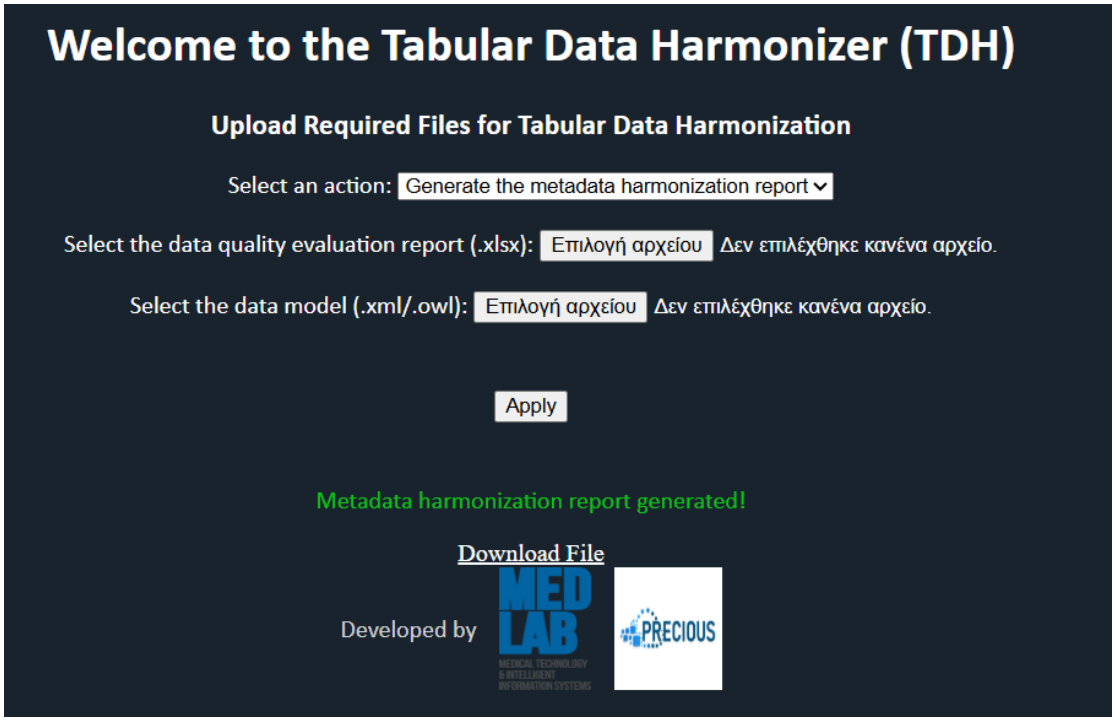
Η υπηρεσία παρέχει στην έξοδο μια ενδελεχή αναφορά προτυποποίησης και εναρμόνισης δεδομένων (**Προδιαγραφή 2.6**) η οποία περιλαμβάνει[14]:

- για κάθε μεταβλητή τις σχετιζόμενες με αυτήν ορολογίες από το εμπλουτισμένο λεξικό (αν έχει υπάρξει λεκτική ταυτοποίηση),
- για κάθε συσχέτιση τους σχετιζόμενους μοναδικούς κωδικούς με στόχο την διασφάλιση των κανόνων FAIR (Findable, Accessible, Interoprable, Reusable)[16].

Ο χρήστης έχει ολοκληρώσει με επιτυχία τη διαδικασία υποβολής των απαιτούμενων αρχείων, δηλαδή της αναφοράς ποιότητας των δεδομένων και του μοντέλου δεδομένων σε μορφή .xml ή .owl, και το εργαλείο παράγει την αντίστοιχη αναφορά εναρμόνισης. Στο κάτω τμήμα της

[15] https://athena.ohdsi.org/
[16] https://www.go-fair.org/fair-principles/

σελίδας εμφανίζεται επιβεβαιωτικό μήνυμα σε πράσινη γραμματοσειρά που δηλώνει ότι η διαδικασία ολοκληρώθηκε ορθά, ενώ παρέχεται και ενεργός σύνδεσμος για τη λήψη του παραγόμενου αρχείου αναφοράς.



Εικόνα 21: Στιγμιότυπο από την επιτυχή εκτέλεση του Μέρους 1 της υπηρεσίας.

Η Εικόνα 21 απεικονίζει ένα στιγμιότυπο από την αναφορά εναρμόνισης δεδομένων σε επίπεδο μεταδεδομένων, όπως παράγεται από το πρώτο μέρος της υπηρεσίας. Η αναφορά παρουσιάζεται σε μορφή πίνακα και περιλαμβάνει τις μεταβλητές του αρχικού συνόλου δεδομένων (Feature Name), τις αντίστοιχες ορολογίες με τις οποίες ταυτοποιήθηκαν (Matched Term), καθώς και τον δείκτη ομοιότητας (Matching Score) που αποτυπώνει τον βαθμό λεκτικής και σημασιολογικής συνάφειας. Επιπλέον, για κάθε μεταβλητή παρατίθεται το εύρος τιμών (Value Range), η πηγή προέλευσης (Source), η ιεραρχική σχέση με άλλους όρους της οντολογίας (Parent Term) και ο μοναδικός αναγνωριστικός κωδικός (Concept ID) που εξασφαλίζει τη συμμόρφωση με τις αρχές FAIR.

Η αναφορά αναδεικνύει τόσο τις περιπτώσεις επιτυχούς ταυτοποίησης όσο και εκείνες όπου δεν εντοπίστηκε αντιστοιχία, γεγονός που υποδεικνύεται με την ένδειξη «No match found». Επίσης, φαίνονται μεταβλητές που συνδέονται με κλινικές οντολογίες ή διεθνή λεξιλόγια, όπως SNOMED-CT και LOINC, γεγονός που ενισχύει τη διαλειτουργικότητα. Η παρουσία διαφορετικών επιπέδων ομοιότητας καταδεικνύει τον ρόλο των τεχνικών λεκτικής ανάλυσης και των εμπλουτισμένων λεξικών στην αναγνώριση ισοδύναμων ή παρεμφερών όρων. Το στιγμιότυπο αυτό λειτουργεί ως απόδειξη ότι η διαδικασία εναρμόνισης σε επίπεδο

μεταδεδομένων μπορεί να εντοπίσει, να αντιστοιχίσει και να τυποποιήσει μεταβλητές, προετοιμάζοντας έτσι το έδαφος για την τελική εναρμόνιση σε επίπεδο δεδομένων.

Feature Name	Matched Term	Matching Score	Value Range	Source	Parent Term	Concept ID
ID	id	1	Too large to display!!	XML	demographics	
AGE	age	1	[49,102]	XML	demographics	
GENDER	gender	1	[0, 1]	XML	demographics	
SMOKING	Smoking device	0,93936258	(1, , 0)	Corpus -> XML	No Parent	44783989
PACK-YEARS	Pack years	1	[0, 200]	Corpus -> XML	No Parent	4151768
JOB	job	1	(None,)	XML	demographics	
AAA STATUS	AAA	0,846957692	[0, 2]	Corpus -> XML	No Parent	3607933
AAA MORHOLOGY	AAA	0,650676184	[0, 3]	Corpus -> XML	No Parent	3607933
AAA AETIOLOGY	Aetiology	0,721612431	[0, 1]	Corpus -> XML	No Parent	44804449
AAA LOCATION	AAA	0,784016695	[1, 4]	Corpus -> XML	No Parent	3607933
AAA LENGTH	AAA	0,954245713	[4.4, 13.5]	Corpus -> XML	No Parent	3607933
MAX DIAMETER	Diameter	0,724147684	[0.5, 13.5]	Corpus -> XML	No Parent	4305186
NECK DIAMETER	Diameter	0,713460726	[0, 3.5]	Corpus -> XML	No Parent	4305186
NECK LENGTH	Neck class	0,804384898	[0, 6.3]	Corpus -> XML	No Parent	4045301
NECK SHAPE	Shape	0,798779356	[0, 3]	Corpus -> XML	No Parent	4115102
SUPRARENAL ANGLE	Q angle	0,651632252	[0, 80]	Corpus -> XML	No Parent	4035459
INFRARENAL ANGLE	Entire infrarenal aorta	0,617831021	[1, 100]	Corpus -> XML	No Parent	4108417
ILT	No match found	0	[0, 2]	N/A	N/A	N/A
CALCIFICATIONS	calcifications	1	[0, 2]	XML	clinical_tests	
RCIA DIAMETER	No match found	0	[0.9, 7]	N/A	N/A	N/A
RCIA LENGTH	No match found	0	[1.2, 10.8]	N/A	N/A	N/A
LCIA DIAMETER	No match found	0	[0.8, 7]	N/A	N/A	N/A
LCIA LENGTH	No match found	0	[2.15, 11]	N/A	N/A	N/A
FAMILY HISTORY	Family history	1	(None,)	Corpus -> XML	No Parent	44810316
SYSTOLIC PRESSURE	Systolic blood pressure	0,893319156	(None,)	Corpus -> XML	No Parent	4152194
DIASTOLIC PRESSURE	Diastolic blood pressure	0,894759783	(None,)	Corpus -> XML	No Parent	4154790
HEIGHT	height	1	[1.52, 180]	XML	clinical_tests	
WEIGHT	weight	1	[50, 180]	XML	clinical_tests	
HEMATOCRIT (%)	Hematocrit	1	[25.4, 54]	Corpus -> XML	No Parent	40451481
HYPERTENSION	hypertension	1	(0, 1, None, ?)	XML	clinical_tests	
DIABETES	diabetes	1	[0, 1]	XML	clinical_tests	
DYSLIPIDEMIA	dyslipidemia	1	(0, 1, None, ?)	XML	clinical_tests	
CAD	Blood group antibody Cad	0,74353305	(1, 0, None, ?)	Corpus -> XML	No Parent	4199748
COPD	COPD assessment test	0,786377955	(1, 0, None, ?)	Corpus -> XML	No Parent	764579
CVD	CVD (cardiovascular disease) risk assessment by third party	0,547341819	[0, 1]	Corpus -> XML	No Parent	3543398

Εικόνα 22: Στιγμιότυπο από την αναφορά εναρμόνισης δεδομένων σε επίπεδο μεταδεδομένων.

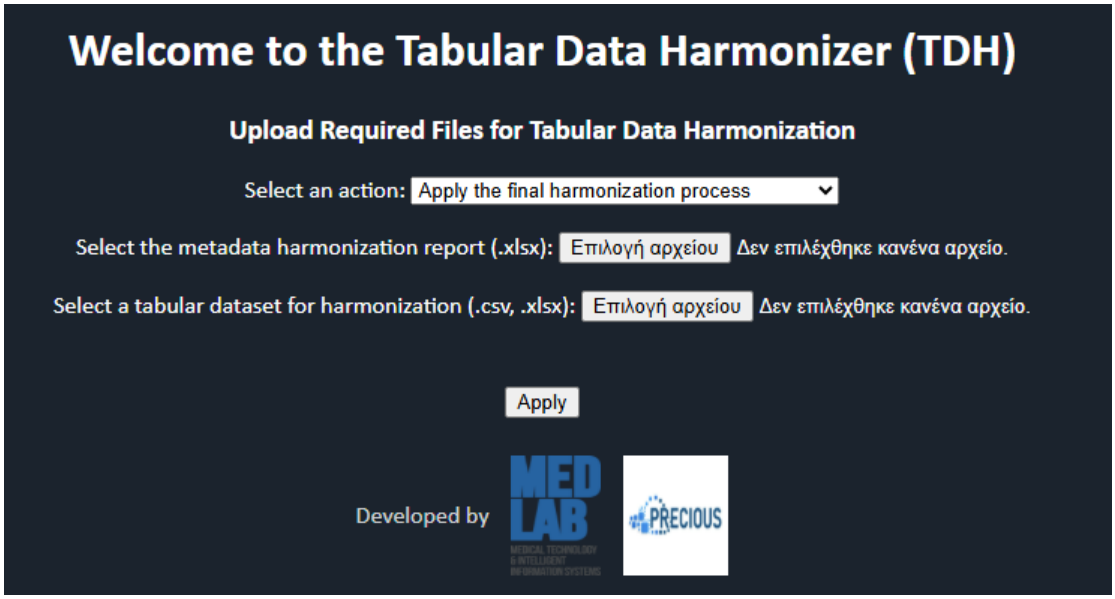
4.2.2. Μέρος 2 – Τελική διαδικασία εναρμόνισης (σε επίπεδο δεδομένων)

4.2.2.1. Είσοδος

Στο Μέρος 2, η υπηρεσία δέχεται στην είσοδο (Εικόνα 23):

- Την αναφορά προτυποποίησης και εναρμόνισης δεδομένων (**Προδιαγραφή 2.6**) – από την έξοδο της υπηρεσίας στο Μέρος 1, στην οποία ο χρήστης προσθέτει μια νέα στήλη (XXX) με το επιθυμητό εύρος τιμών ανά μεταβλητή.

- Το αρχικό σύνολο δεδομένων σε μορφή πίνακα (υποστηριζόμενες μορφές: .xlsx, .csv)



Εικόνα 23: Η αρχική σελίδα της υπηρεσίας εξασφάλισης της διαλειτουργικότητας και της ομογένειάς των δεδομένων (σε μορφή πίνακα) μεταξύ των ομοσπονδιακών βάσεων (Μέρος 2).

4.2.2.2. Λειτουργίες

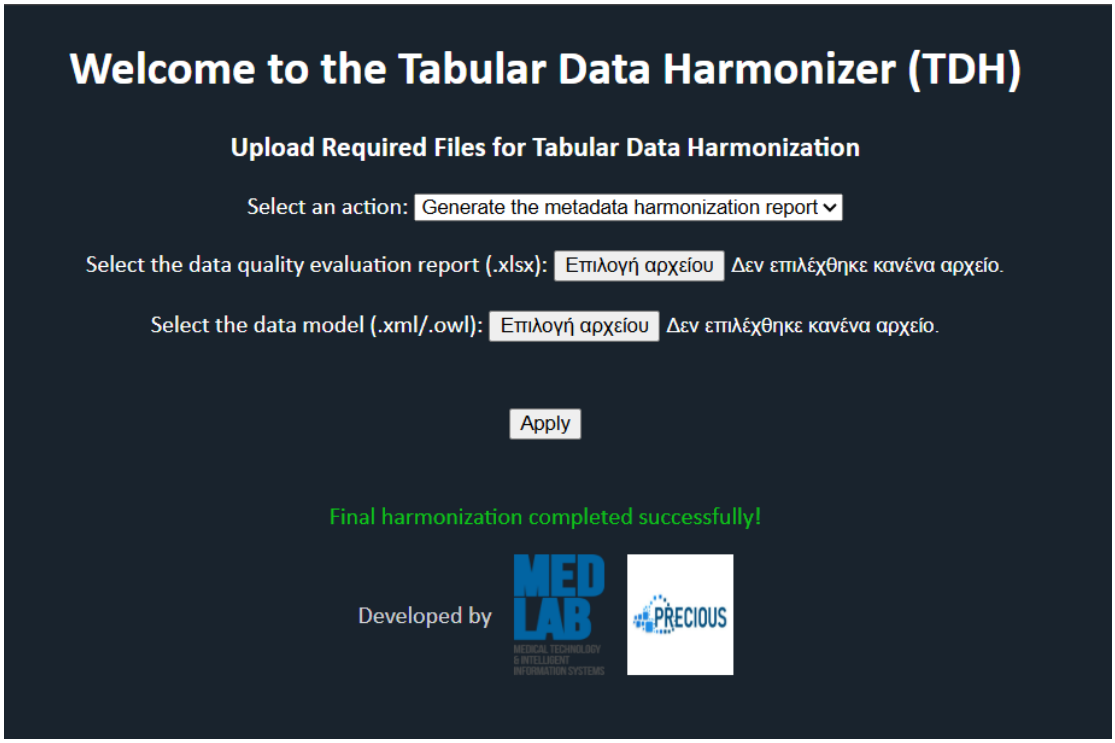
Η υπηρεσία παρέχει λειτουργίες για:

- Τον μετασχηματισμό των δεδομένων βάσει της αντιστοίχισης στο επιθυμητό εύρος τιμών ανά μεταβλητή από την αναφορά προτυποποίησης και εναρμόνισης δεδομένων.
- Την αποθήκευση του μετασχηματισμένου συνόλου δεδομένων ως το τελικό αποτέλεσμα της εναρμόνισης σε επίπεδο δεδομένων.

4.2.2.3. Έξοδος

Η υπηρεσία παρέχει στην έξοδο το μετασχηματισμένο σύνολο δεδομένων ως το τελικό αποτέλεσμα της εναρμόνισης σε επίπεδο δεδομένων.

Η Εικόνα 24 αποτυπώνει στιγμιότυπο που εμφανίζεται μετά την επιτυχή ολοκλήρωση του Μέρους 2 της υπηρεσίας εναρμόνισης δεδομένων. Στο κέντρο της σελίδας εμφανίζεται μήνυμα επιβεβαίωσης με πράσινη γραμματοσειρά, το οποίο δηλώνει ότι η τελική διαδικασία εναρμόνισης ολοκληρώθηκε με επιτυχία («Final harmonization completed successfully!»).



Εικόνα 24: Στιγμιότυπο από την επιτυχή εκτέλεση του Μέρους 2 της υπηρεσίας

Η Εικόνα 25 παρουσιάζει την τροποποιημένη αναφορά εναρμόνισης δεδομένων σε επίπεδο μεταδεδομένων, η οποία προκύπτει μετά την ολοκλήρωση του Μέρους 2 της υπηρεσίας. Η αναφορά διατηρεί τη δομή του αρχικού πίνακα, όπου καταγράφονται οι μεταβλητές του συνόλου δεδομένων (Feature Name), οι όροι με τους οποίους ταυτοποιήθηκαν (Matched Term), ο δείκτης ομοιότητας (Matching Score), τα εύρη τιμών (Value Range), οι πηγές προέλευσης (Source), οι γονικοί όροι της οντολογίας (Parent Term) και οι μοναδικοί κωδικοί ταυτοποίησης (Concept ID). Η βασική διαφοροποίηση σε αυτήν την έκδοση είναι η προσθήκη της τελευταίας στήλης «Target Value Range», μέσω της οποίας ο χρήστης δηλώνει το επιθυμητό εύρος τιμών για κάθε μεταβλητή που εντοπίστηκε στο Μέρος 1 της διαδικασίας. Η στήλη αυτή αποτελεί κρίσιμο στοιχείο για τη μεταγενέστερη φάση μετασχηματισμού των δεδομένων, καθώς καθορίζει πώς θα πρέπει να προσαρμοστούν οι τιμές των μεταβλητών ώστε να ευθυγραμμιστούν με τις απαιτήσεις της εναρμόνισης. Η ύπαρξη των «Target Value Ranges» επιτρέπει την τυποποίηση των τιμών σε ομοιογενή μορφή, μειώνοντας ασυμβατότητες και ενισχύοντας τη διαλειτουργικότητα μεταξύ διαφορετικών συνόλων δεδομένων. Επιπλέον, η δυνατότητα παρέμβασης από τον χρήστη διασφαλίζει ότι η εναρμόνιση δεν βασίζεται αποκλειστικά σε αυτόματους αλγοριθμικούς κανόνες, αλλά εμπλουτίζεται με την επιστημονική κρίση και τις ανάγκες του εκάστοτε σεναρίου χρήσης.

Feature Name	Matched Term	Matching Score	Value Range	Source	Parent Term	Concept ID	Target Value Range
ID	id	1	Too large to display!!	XML	demographics		"Same"
AGE	age	1	[49,102]	XML	demographics		"Same"
GENDER	gender	1	[0, 1]	XML	demographics		["M","F"]
SMOKING	Smoking device	0,93936258	(1, , 0)	Corpus -> XML	No Parent	44783989	"Same"
PACK-YEARS	Pack years	1	[0, 200]	Corpus -> XML	No Parent	4151768	"Same"
JOB	job	1	(None,)	XML	demographics		"Same"
AAA STATUS	AAA	0,846957692	[0, 2]	Corpus -> XML	No Parent	3607933	"Same"
AAA MORHOLOGY	AAA	0,650676184	[0, 3]	Corpus -> XML	No Parent	3607933	"Same"
AAA AETIOLOGY	Aetiology	0,721612431	[0, 1]	Corpus -> XML	No Parent	44804449	[1,2]
AAA LOCATION	AAA	0,784016695	[1, 4]	Corpus -> XML	No Parent	3607933	"Same"
AAA LENGTH	AAA	0,954245713	[4.4, 13.5]	Corpus -> XML	No Parent	3607933	"Same"
MAX DIAMETER	Diameter	0,724147684	[0.5, 13.5]	Corpus -> XML	No Parent	4305186	"Same"
NECK DIAMETER	Diameter	0,713460726	[0, 3.5]	Corpus -> XML	No Parent	4305186	"Same"
NECK LENGTH	Neck class	0,804384898	[0, 6.3]	Corpus -> XML	No Parent	4045301	"Same"
NECK SHAPE	Shape	0,798779356	[0, 3]	Corpus -> XML	No Parent	4115102	"Same"
SUPRARENAL ANGLE	Q angle	0,651632252	[0, 80]	Corpus -> XML	No Parent	4035459	"Same"
INFRARENAL ANGLE	Entire infrarenal aorta	0,617831021	[1, 100]	Corpus -> XML	No Parent	4108417	"Same"
ILT	No match found	0	[0, 2]	N/A	N/A	N/A	"Same"
CALCIFICATIONS	calcifications	1	[0, 2]	XML	clinical_tests		"Same"
RCIA DIAMETER	No match found	0	[0.9, 7]	N/A	N/A	N/A	"Same"
RCIA LENGTH	No match found	0	[1.2, 10.8]	N/A	N/A	N/A	"Same"
LCIA DIAMETER	No match found	0	[0.8, 7]	N/A	N/A	N/A	"Same"
LCIA LENGTH	No match found	0	[2.15, 11]	N/A	N/A	N/A	"Same"
FAMILY HISTORY	Family history	1	(None,)	Corpus -> XML	No Parent	44810316	"Same"
SYSTOLIC PRESSURE	Systolic blood pressure	0,893319156	(None,)	Corpus -> XML	No Parent	4152194	"Same"
DIASTOLIC PRESSURE	Diastolic blood pressure	0,894759783	(None,)	Corpus -> XML	No Parent	4154790	"Same"
HEIGHT	height	1	[1.52, 180]	XML	clinical_tests		"Same"
WEIGHT	weight	1	[50, 180]	XML	clinical_tests		"Same"
HEMATOCRIT (%)	Hematocrit	1	[25.4, 54]	Corpus -> XML	No Parent	40451481	"Same"
HYPERTENSION	hypertension	1	(0, 1, None, ?)	XML	clinical_tests		"Same"
DIABETES	diabetes	1	[0, 1]	XML	clinical_tests		"Same"
DYSLIPIDEMIA	dyslipidemia	1	(0, 1, None, ?)	XML	clinical_tests		"Same"
CAD	Blood group antibody Cad	0,74353305	(1, 0, None, ?)	Corpus -> XML	No Parent	4199748	"Same"
COPD	COPD assessment test	0,786377955	(1, 0, None, ?)	Corpus -> XML	No Parent	764579	"Same"
CVD	CVD (cardiovascular disease) risk assessment by third party	0,547341819	[0, 1]	Corpus -> XML	No Parent	3543398	"Same"

Εικόνα 25: Η τροποποιημένη αναφορά εναρμόνισης δεδομένων σε επίπεδο μεταδεδομένων με την προσθήκη της τελευταίας στήλης «Target Value Range» όπου ο χρήστης δηλώνει το επιθυμητό εύρος τιμών των μεταβλητών που έχουν εντοπιστεί από το Μέρος 1 της υπηρεσίας.

Το στιγμιότυπο στην Εικόνα 26 παρουσιάζει τον πίνακα με τα εναρμονισμένα δεδομένα, όπως αυτά προκύπτουν μετά την εφαρμογή του μετασχηματισμού στα επιθυμητά εύρη τιμών που ορίστηκαν στο προηγούμενο βήμα. Οι στήλες του πίνακα αντιστοιχούν στις μεταβλητές που είχαν εντοπιστεί στο στάδιο της εναρμόνισης μεταδεδομένων, ενώ οι τιμές έχουν πλέον προσαρμοστεί ώστε να είναι συμβατές με τα καθορισμένα πρότυπα. Η μορφή του πίνακα καταδεικνύει την επιτυχή τυποποίηση των δεδομένων, με τις αριθμητικές τιμές να έχουν ομογενοποιηθεί και τις ελλείπουσες ή μη αναγνωρίσιμες τιμές να δηλώνονται με το σύμβολο «?». Ο πίνακας περιλαμβάνει τόσο δημογραφικές πληροφορίες (π.χ. id, age, gender, smoking device, pack years, job) όσο και κλινικές μετρήσεις (π.χ. diameters, neck class, Q angle, entire infrarenal aorta, calcifications, family history). Η παρουσία τιμών εντός συγκεκριμένων εύρων, όπως αυτά ορίστηκαν στη στήλη «Target Value Range» της προηγούμενης αναφοράς, επιβεβαιώνει ότι η διαδικασία εναρμόνισης έχει επιτύχει τον στόχο της ευθυγράμμισης των δεδομένων με τα προδιαγεγραμμένα όρια. Το στιγμιότυπο αποδεικνύει ότι το εργαλείο είναι σε θέση να μετασχηματίζει αυτόματα το αρχικό, ετερογενές dataset σε μια τυποποιημένη μορφή, έτοιμη για αξιοποίηση σε αναλυτικές διαδικασίες και εφαρμογές μηχανικής μάθησης, διασφαλίζοντας παράλληλα τη διαλειτουργικότητα σε περιβάλλον ομοσπονδιακών βάσεων.

id	age	gender	Smoking device	Pack years	job	Diameter	Diameter	Diameter	Diameter	Neck class	Shape	Q angle	Entire infrarenal aorta	calcifications	Family history
GK24-14	81	F	1	40	?	5,6	2,7	5,6	2,7	5,48	1	?	?	2	?
GK24-15	70	F	1	35	?	5,4	2,7	5,4	2,7	3,2	1	?	35	1	?
CHVA	72	F	1	125	?	?	?	?	?	?	?	?	?	?	?
GK24-46	82	F	1	40	?	4,6	1,9	4,6	1,9	2,5	1	?	?	2	?
GK24-1	78	F	?	?	?	?	?	?	?	?	?	?	?	?	?
GK24-21	71	F	1	50	?	5,7	?	5,7	?	?	?	0	?	1	?
GK24-23	78	F	1	60	?	5,5	2,6	5,5	2,6	4,2	1	?	40	1	?
EFGI	72	F	1	50	?	?	?	?	?	?	?	?	?	?	?
GEGI	74	F	1	25	?	?	?	?	?	?	?	?	?	?	?
ELGO	66	M	1	40	?	5,6	2,25	5,6	2,25	3,8	1	?	40	2	?
KOGO	71	F	1	125	?	6	2,6	6	2,6	3,9	1	?	30	1	?
GK24-18	77	F	1	35	?	5,4	2,5	5,4	2,5	2	1	?	30	2	?
IODE	79	F	1	40	?	?	?	?	?	?	?	?	?	?	?
IOIL	72	F	?	?	?	?	?	?	?	?	?	?	?	?	?
DIKA	51	F	1	30	?	?	?	?	?	?	?	?	?	?	?
GK24-52	82	F	1	100	?	7	2,5	7	2,5	3	1	64	60	2	?
GK24-26	75	F	1	110	?	5,45	2,8	5,45	2,8	2,2	1	80	72	1	?
GK23-32	75	F	1	75	?	6,2	2,1	6,2	2,1	1,9	1	36	70	1	?
IOKA	72	F	1	20	?	?	?	?	?	?	?	?	?	?	?
GK24-27	73	F	?	?	?	5	2,1	5	2,1	2	1	60	70	1	?
PAKA	73	F	1	100	?	7	?	7	?	?	?	?	?	?	?
ANKE	52	F	0	0	?	?	?	?	?	?	?	?	?	?	?
ALKO	67	F	1	120	?	?	?	?	?	?	?	?	?	?	?
GK23-80	70	F	1	60	?	5,7	2,2	5,7	2,2	3,3	1	?	34	1	?
GK24-32	70	F	1	50	?	5,8	2,7	5,8	2,7	5	1	30	20	1	?
NIKO	90	F	0	0	?	6	2,3	6	2,3	4,9	1	?	?	1	?

Εικόνα 26: Στιγμιότυπο από τα εναρμονισμένα δεδομένα κατόπιν εφαρμογής του μετασχηματισμού των τιμών των ταυτοποιημένων μεταβλητών στα επιθυμητά εύρη τιμών.

4.3. Σύνδεση με το Gitlab

Οι υπηρεσίες είναι διαθέσιμες μέσω του Gitlab του PRECIOUS στον κάτωθι σύνδεσμο:

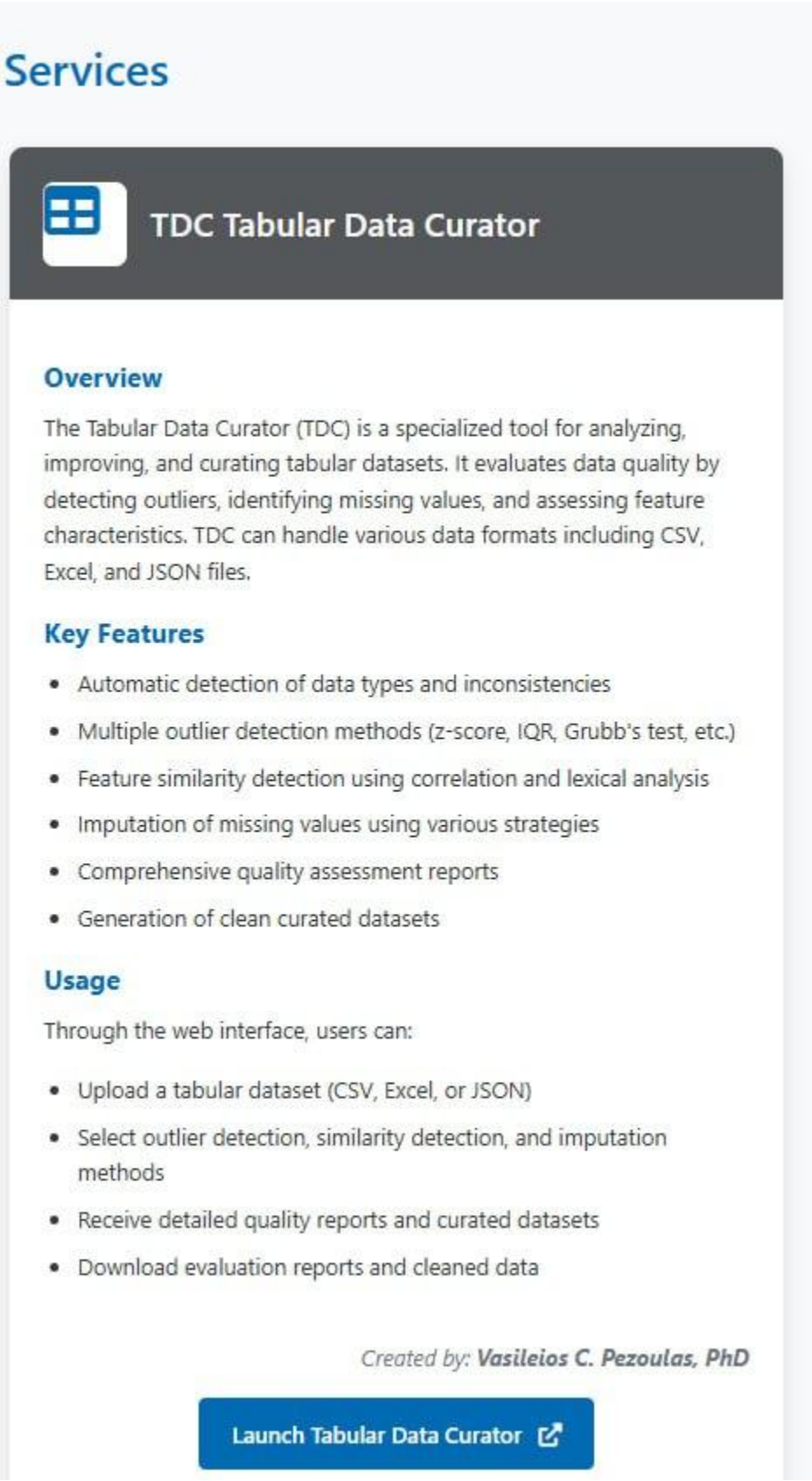
- http://me.preciouscloud.eu:8090/

Για την πρόσβαση στο Gitlab του PRECIOUS είναι πρώτα αναγκαία η δημιουργία λογαριασμού μέσω του παρακάτω συνδέσμου:

- https://gitlab.preciouscloud.eu/users/sign_in

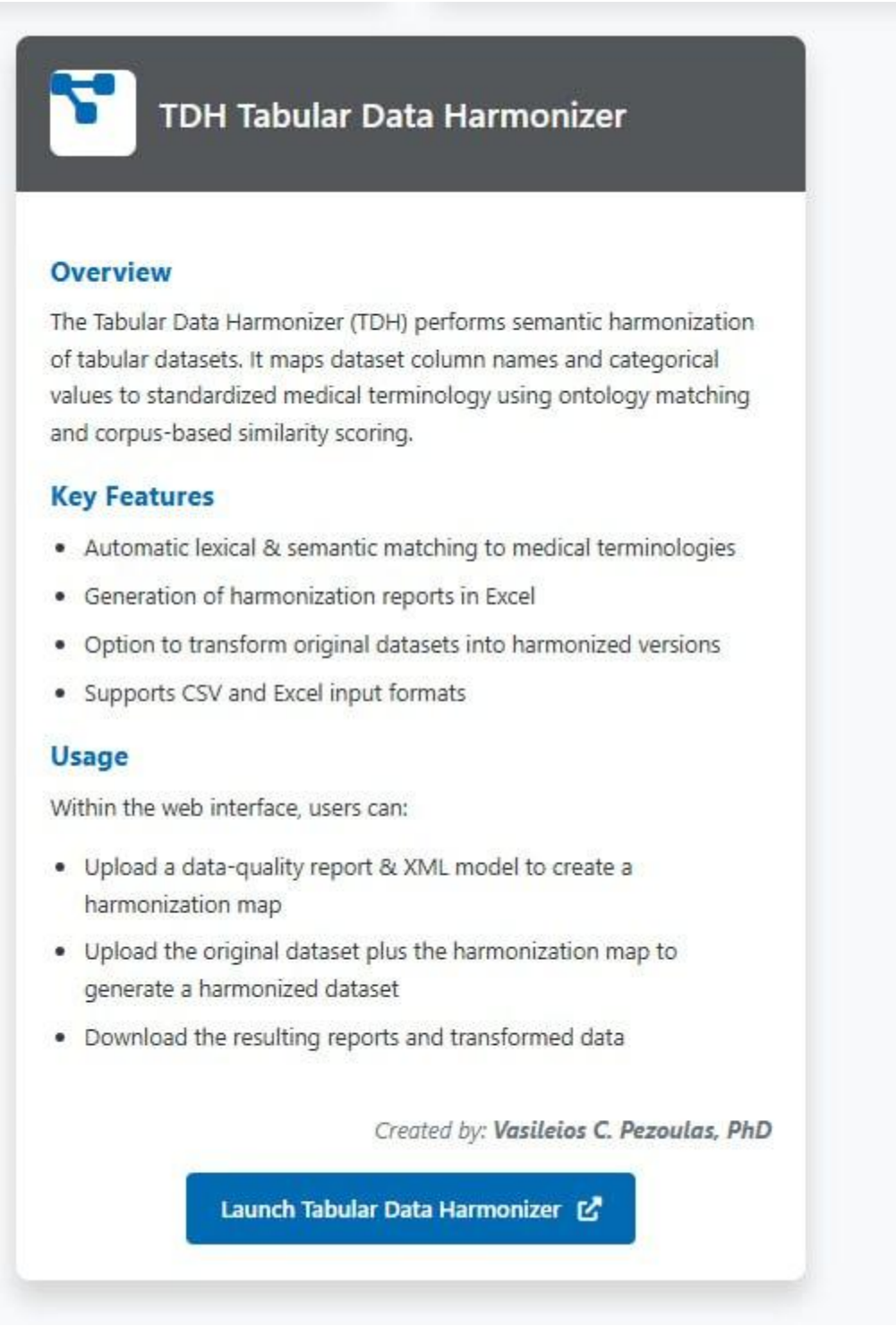
Εικόνα 27: Η αρχική οθόνη κατόπιν πρόσβασης στο Gitlab του PRECIOUS για το έργο.

Η αρχική οθόνη της υπηρεσίας ελέγχου της ποιότητάς των δεδομένων σε μορφή πίνακα (tabular data) απεικονίζεται στην Εικόνα 27.



Εικόνα 28: Η αρχική οθόνη της υπηρεσίας ελέγχου της ποιότητάς των δεδομένων σε μορφή πίνακα (tabular data) {ακρωνύμιο: TDC Tabular Data Curator}.

Η αρχική οθόνη υπηρεσίας εξασφάλισης της διαλειτουργικότητας και της ομογένειάς των δεδομένων (σε μορφή πίνακα) μεταξύ των ομοσπονδιακών βάσεων) απεικονίζεται στην Εικόνα 28.



Εικόνα 29: Η αρχική της υπηρεσίας εξασφάλισης της διαλειτουργικότητας και της ομογένειάς των δεδομένων (σε μορφή πίνακα) μεταξύ των ομοσπονδιακών βάσεων {ακρωνύμιο: TDH Tabular Data Harmonizer}

6. Συμπεράσματα και μελλοντικές κατευθύνσεις

Η παρούσα εργασία ολοκληρώνει τον κύκλο ανάπτυξης των υποδομών δεδομένων του έργου **SAFEAORTA**, παρέχοντας ένα πλήρως λειτουργικό, εναρμονισμένο και διαλειτουργικό περιβάλλον διαχείρισης ιατρικών δεδομένων. Μέσω της διαδικασίας εναρμόνισης που τεκμηριώθηκε στο παρόν παραδοτέο, επιτεύχθηκε η ομογενοποίηση τόσο των **απεικονιστικών (DICOM)** όσο και των **πινακοποιημένων (tabular)** δεδομένων, διασφαλίζοντας κοινή μορφοποίηση, συνέπεια μεταδεδομένων και αντιστοίχιση ορολογίας.

Η προτεινόμενη μεθοδολογία κατέστησε δυνατή:

- Τη **σύνδεση και ενσωμάτωση δεδομένων** από διαφορετικούς κλινικούς και ερευνητικούς φορείς, χωρίς απώλεια πληροφορίας ή παραβίαση ιδιωτικότητας.
- Τη **συμβατότητα** των αρχείων με διεθνή πρότυπα (DICOM, FHIR, HL7, ISO/IEC 27001), επιτρέποντας τη διαλειτουργικότητα με άλλες υποδομές υγείας.
- Τη **συμμόρφωση** με το νομικό και δεοντολογικό πλαίσιο του **GDPR**, εφαρμόζοντας αρχές ασφάλειας εκ σχεδιασμού και ελαχιστοποίησης δεδομένων.
- Την **πλήρη ενσωμάτωση** της εναρμονισμένης πληροφορίας στο **ομοσπονδιακό νέφος δεδομένων (Federated Cloud Infrastructure)**, όπως αναπτύχθηκε στα προηγούμενα παραδοτέα του ΠΕ3.

Η επιτυχής ολοκλήρωση της εναρμόνισης δημιουργεί μια ισχυρή βάση για τις επόμενες ερευνητικές και τεχνολογικές ενότητες του έργου. Η διαθέσιμη πια ενοποιημένη βάση δεδομένων θα αξιοποιηθεί για:

- **Εκπαίδευση και επικύρωση προγνωστικών μοντέλων τεχνητής νοημοσύνης (AI/ML)** που θα προβλέπουν τον κίνδυνο ρήξης ανευρύσματος κοιλιακής αορτής.
- **Ανάπτυξη και ενημέρωση των Ψηφιακών Διδύμων της Αορτής (ΨηφιΔΑ)** με δεδομένα υψηλής ποιότητας και συνέπειας, επιτρέποντας προσωποποιημένες προσομοιώσεις και κλινικά σενάρια what-if.
- **Ενίσχυση του Συστήματος Υποστήριξης Κλινικών Αποφάσεων (ΣΥΠΟΚΑ)**, ώστε να παρέχει στους ιατρούς εξατομικευμένες προτάσεις θεραπευτικής διαχείρισης βάσει ομοιογενοποιημένων δεδομένων.

Μελλοντικές Κατευθύνσεις

Οι μελλοντικές δραστηριότητες που θα ακολουθήσουν τη φάση της εναρμόνισης επικεντρώνονται στην περαιτέρω **αυτοματοποίηση**, **επεκτασιμότητα** και **ευφυή αξιοποίηση** των δεδομένων. Συγκεκριμένα:

1. **Ανάπτυξη αυτοματοποιημένων pipelines εναρμόνισης**
   Επέκταση της τρέχουσας μεθοδολογίας σε πλήρως αυτοματοποιημένα pipelines, τα οποία θα ενσωματωθούν στην υποδομή του SAFEAORTA Cloud, επιτρέποντας on-demand ελέγχους, μετατροπές και επικυρώσεις δεδομένων χωρίς ανθρώπινη παρέμβαση.

2. **Ενσωμάτωση οντολογιών και σημασιολογικής ευθυγράμμισης**
   Δημιουργία λεξιλογίου ορολογιών και οντολογικών χαρτών (e.g. SNOMED-CT, LOINC, RadLex) για την περαιτέρω σημασιολογική συνάφεια των δεδομένων, διευκολύνοντας την αυτόματη αναζήτηση και ανάλυση.

3. **Σύνδεση με ομοσπονδιακά σχήματα μηχανικής μάθησης (Federated Learning Frameworks)**
   Διασύνδεση των εναρμονισμένων δεδομένων με πλαίσια όπως το **Flower** ή το **FedBiomed**, ώστε οι αλγόριθμοι να εκπαιδεύονται τοπικά σε κάθε κόμβο, διατηρώντας πλήρη προστασία της ιδιωτικότητας.

4. **Επέκταση της υποδομής σε νέες παθολογίες και πολυοργανικές αναλύσεις**
   Το πλαίσιο εναρμόνισης μπορεί να επαναχρησιμοποιηθεί για άλλες αγγειακές παθήσεις (θωρακικά ανευρύσματα, καρωτίδες, στεφανιαίες), υποστηρίζοντας τη διεύρυνση του οικοσυστήματος SAFEAORTA.

5. **Συνεχής διασφάλιση ποιότητας και συμμόρφωσης**
   Εφαρμογή μηχανισμών συνεχούς monitoring, auditing και versioning των datasets, διατηρώντας πλήρη ιχνηλασιμότητα και πιστοποίηση της διαδικασίας εναρμόνισης.

Συνολικά, το παρόν παραδοτέο σηματοδοτεί τη μετάβαση του έργου **SAFEAORTA** από το στάδιο της **υποδομής και ασφάλειας δεδομένων** στο στάδιο της **ευφυούς ανάλυσης και κλινικής αξιοποίησης**. Η επιτυχής εναρμόνιση των δεδομένων αποτελεί καθοριστικό βήμα προς την υλοποίηση της **πλήρως λειτουργικής πλατφόρμας Ψηφιακού Διδύμου της Αορτής**, που φιλοδοξεί να αποτελέσει σημείο αναφοράς για την εξατομικευμένη καρδιαγγειακή ιατρική στην Ελλάδα και διεθνώς.

7. Βιβλιογραφία

1. Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.).
2. Wiley. Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*.
3. Chapman & Hall. van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). CRC Press.
4. Allison, P. D. (2001). *Missing Data*. Sage Publications.
5. Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). Wiley.
6. Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). *Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance*. *Journal of Experimental Social Psychology*, 74, 150–156.
7. Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). *A survey on unsupervised outlier detection in high-dimensional numerical data*. *Statistical Analysis and Data Mining*, 5(5), 363–387.
8. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
9. Abadi, M. et al. (2016). *TensorFlow: A system for large-scale machine learning*. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 265–283.
10. Chollet, F. (2015). *Keras: The Python Deep Learning library*. GitHub repository: https://github.com/fchollet/keras
11. Harris, C. R., Millman, K. J., van der Walt, S. J. et al. (2020). *Array programming with NumPy*. *Nature*, 585, 357–362.
12. McKinney, W. (2010). *Data structures for statistical computing in Python*. In *Proceedings of the 9th Python in Science Conference*, 51–56.
13. Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12, 2825–2830.
14. Walt, S. v. d., Colbert, S. C., & Varoquaux, G. (2011). *The NumPy Array: A Structure for Efficient Numerical Computation*. *Computing in Science & Engineering*, 13(2), 22–30.
15. Reback, J., McKinney, W., jbrockmendel et al. (2020). *pandas-dev/pandas: Pandas 1.0.3*. Zenodo. https://doi.org/10.5281/zenodo.3509134