Μεγάλα Δεδομένα στον Τομέα της Υγείας

Κυριάκος Χρήστος

Προχωρημένη Διαχείριση Δεδομένων 2020-21
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας, Βόλος
{ckyriakos}@e-ce.uth.gr

Περίληψη Τα δεδομένα που αποκτώνται από τον τομέα της υγείας αυξάνονται καθημερινά με ραγδαίους ρυθμούς. Τα δεδομένα αυτά όπως και σε άλλους τομείς μεγαλώνουν σε ταχύτητα, όγκο , αξία, ποικιλία και η επιλογή των χρησιμότερων γίνεται όλο και πιο δύσκολη. Το Health Data Management είναι η διαδικασία ανεύρεσης από έναν μεγάλο όγκο δεδομένων των πιο σημαντικών και η χρήση τους προς όφελος των οργανισμών υγείας και κατ'επέκταση των ιατρών και των ασθενών. Στο παρόν κείμενο θα γίνει αρχικά μια προσπάθεια διασαφήνισης των εννοιών γύρω από τα μεγάλα δεδομένα καθώς και τις μορφές τους στον τομέα της υγείας. Επιπλέον, θα συζητηθούν οι διάφορες τεχνικές ανάλυσης και διαχείρισης δεδομένων στον τομέα αυτό. Ακόμα, θα επικεντρωθούμε στη συμβολή του Health Data Management στην αντιμετώπιση του Covid-19 που αποτελεί ακόμα ένα παγκόσμιο πρόβλημα. Τέλος, θα αναφερθούν κάποιες προκλήσεις στην προσπάθεια αξιοποίησης των Big Data καθώς και τρόποι προστασίας των δεδομένων υγείας.

1 Εισαγωγή

Η σημερινή ”τεχνολογική εποχή” χαρακτηρίζεται από συνεχή παραγωγή μεγάλου όγκου δεδομένων. Αυτό έχει ως αποτέλεσμα όλο και περισσότεροι οργανισμοί να αντιμετωπίζουν πρόβλημα με τη διαχείριση των δεδομένων αυτών. Συναλλαγές, διεργασίες σε επιχειρήσεις, διαδικτυακές σελίδες, αποτελούν μονάχα κάποιες από τις πηγές τεράστιου όγκου δεδομένων, διαφορετικών μορφών, δομημένα και μη, τα οποία επιχειρείται να αξιοποιηθούν. Όσο περισσότερες πληροφορίες έχουμε, τόσο καλύτερα είναι τα αποτελέσματα. Συνεπώς, η συλλογή δεδομένων συνιστά μια πολύ σημαντική διαδικασία για κάθε επιχείρηση και οργανισμό, αφού μπορούν να χρησιμοποιήσουν αυτές τις πληροφορίες για προβλέψεις τάσεων στην αγορά, μελλοντικά προβλήματα και άλλες εφαρμογές. Στο ίδιο πλαίσιο μπορούμε να εντάξουμε και την συλλογή δεδομένων στον τομέα της υγείας. Παρ΄όλα αυτά, η διαχείριση και επεξεργασία, καθώς και η αξιοποίηση των δεδομένων αυτών αποτελούν πρόκληση. Στο παρόν κείμενο θα δώσουμε τα βασικά χαρακτηριστικά των Big Data καθώς και τα οφέλη τους στους κλάδους της υγείας. Θα αναφερθούμε στις διάφορες προκλήσεις που συνταντάμε στην επεξεργασία και διαχείριση τους. Τέλος, θα μιλήσουμε και για τη σημασία τους στην επίλυση προβλημάτων, όπως αυτό του Covid-19.

1.1 Τι είναι τα Big Data;

Πριν περιγράψουμε τα ποικιλόμορφα οφέλη τους στις επιστήμες υγείας, οφείλουμε να εξηγήσουμε τί ακριβώς εννοούμε με τον όρο ”Big Data”. Τα Big Data είναι στην ουσία ένας μεγάλος όγκος πληροφοριών, που πηγάζουν από όλες τις καθημερινές μας δραστηριότητες (π.χ. τραπεζικές συναλλαγές, αγορές, αναζήτηση στο διαδίκτυο κ.α.). Λόγω του όγκου και της πολυπλοκότητάς τους δεν καθίσταται δυνατή η επεξεργασία τους με τα παραδοσιακά μέσα(π.χ υπάρχοντες αλγόριθμοι, λογισμικό κ.α). [1] Τα τελευταία χρόνια, αποκτούν ολοένα και περισσότερη σημασία, καθώς απασχολούν πολλούς τομείς ερευνών και επιχειρήσεις λόγω των πολλαπλών πλεονεκτημάτων που μπορεί να προσφέρουν. Στην προσπάθεια μας να τα διασαφηνίσουμε, έχουμε καταλήξει σε 5 βασικά χαρακτηριστικά(”5 ’Vs of Big Data”) [2], τα οποία παρουσιάζονται συνοπτικά παρακάτω:

- **Volume(Όγκος):** αφορά στο μεγάλο αριθμό παραγόμενων δεδομένων, γεγονός που καθιστά δύσκολη την αποθήκευση και την ανάλυση τους με τους παραδοσιακούς τρόπους(π.χ. SQL βάσεις).
- **Velocity(Ταχύτητα):** αφορά στην ταχύτητα παραγωγής και μετακίνησης των δεδομένων στα δίκτυα.
- **Variety(Ποικιλία):** αφορά στους διάφορους τύπους δεδομένων που χρησιμοποιούνται, δηλαδή δομημένα (Structured), μη δομημένα(Unstructured) και την ανάμειξη τους.
- **Veracity(Εγκυρότητα):** αφορά στο πόσο μπορούμε να εμπιστευτούμε τα δεδομένα που παράγονται ως προς την ποιότητα και την ευστοχία τους, που επηρεάζει σημαντικά ο μεγάλος όγκος.
- **Value(Αξία):** αφορά στην δυνατότητα να αντλήσουμε πολύτιμη πληροφορία και να οδηγηθούμε σε ουσιαστικά συμπεράσματα.

1.2 Ο τομέας υγείας ως αποθετήριο δεδομένων

Ο **τομέας υγείας** αφορά στη συντήρηση, στη βελτίωση της υγείας μέσα από την πρόληψη, στη διάγνωση, στην ανάρρωση ή στην θεραπεία μιας ασθένειας ή τραυματισμού και γενικότερα προβλήματα υγείας(ψυχικής, σωματικής). Στον τομέα αυτό συμμετέχουν διάφοροι φορείς, οργανισμοί και ιδιώτες, με κύριους ”κόμβους” τα νοσοκομεία, τις κλινικές. Καθημερινά, μέσα από την εξέταση των ασθενών και την αντιμετώπιση των προβλημάτων τους, καθώς και από έρευνες, παράγονται άπειρα δεδομένα. Η πληροφορία είναι το κλειδί για την ανάπτυξη και την οργάνωση. Όσο περισσότερες πληροφορίες έχουμε, τόσο καλύτερα τα αποτελέσματα. Συνεπώς, η συλλογή δεδομένων είναι μια πολύ σημαντική διαδικασία για κάθε επιχείρηση και οργανισμό, αφού μπορούν να χρησιμοποιήσουν τις πληροφορίες για προβλέψεις τάσεων στην αγορά, μελλοντικά προβλήματα κι άλλες εφαρμογές. Είναι ,λοιπόν, φυσικό επακόλουθο να εντάξουμε την συλλογή δεδομένων στο σύστημα υγείας. Η εξέλιξη της τεχνολογίας και της επιστήμης των υπολογιστών επιτρέπουν την ψηφιοποίηση των αρχείων των ασθενών. Τα αρχεία αυτά περιλαμβάνουν το ιατρικό ιστορικό τους και αναφέρουν τις διάφορες παθήσεις τους. Με την ένταξή τους σε ένα ψηφιακό σύστημα μπορούν με ευκολία να καταγραφούν, επεξεργαστούν και μεταδοθούν με κύριο στόχο την βελτίωση του υπάρχοντος συστήματος υγείας και των υπηρεσιών του .

1.3 Ηλεκτρονικά αρχεία υγείας (EHR)

Τα Ηλεκτρονικά Αρχεία Υγείας είναι η ηλεκτρονική εκδοχή του ιατρικού ιστορικού ενός ασθενούς. Το αρχείο αυτό συντηρείται από έναν ειδικό ανά τακτά χρονικά διαστήματα και μπορεί να περιλαμβάνει όλα τα δεδομένα του ασθενούς που σχετίζονται με την υποστήριξη του από κάποιον φορέα. Με άλλα λόγια περιέχει: δημογραφικά στοιχεία, σημειώσεις σχετικές με την πορεία της υγείας του, διάφορα προβλήματα που τυχόν παρουσιάστηκαν αλλά και τη φαρμακευτική αγωγή που του χορηγήθηκε. Τέλος, περιέχονται στοιχεία εμβολιασμού, εργαστηριακά δεδομένα και ραδιολογικές αναφορές.[3] Οι υγειονομικοί φορείς έχουν επίσης πρόσβαση σε web εφαρμογές και πλατφόρμες που βελτιώνουν την δουλειά τους με αυτοματοποιημένες ειδοποιήσεις και μηνύματα σχετικά με εμβολιασμούς, παθολογικά αποτελέσματα εξετάσεων, εμφανίσεις καρκίνου και άλλους περιοδικούς ελέγχους. Τα EHR επιτρέπουν ταχύτερη ανάκτηση δεδομένων και διευκολύνουν την αναφορά καίριων ενδείξεων υγείας. Σε συνδυασμό με το διαδίκτυο κάνουν δυνατή την πρόσβαση σε σημαντικές πληροφορίες για έναν ασθενή. Εκτός από τα EHR υπάρχουν και άλλες παραλλαγές αρχείων που αποθηκεύουν πληροφορίες για την υγεία ενός ασθενούς, όπως τα EMR που περιέχουν ιατρικά και κλινικά δεδομένα ασθενών, τα PHR που περιλαμβάνουν το προσωπικό ιατρικό ιστορικό και λογισμικό για την διαχείριση ιατρικών πράξεων. Τα παραπάνω επιτρέπουν μεν τη βελτίωση του τομέα αλλά παράγουν τεράστιο αριθμό ”raw data”.

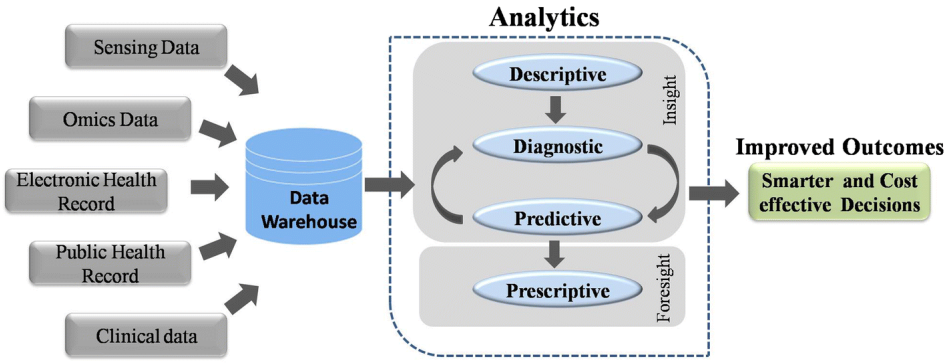
Τα Big Data σε αυτόν τον τομέα είναι κυρίως μη-δομημένα. Τα δεδομένα που περιγράφουν έναν ασθενή, όπως τον τρόπο ζωής του, τις προτιμήσεις του, τη διάγνωση του, δεν είναι ίδια για κάθε άτομο. Αυτό καθιστά ιδιαίτερα δύσκολη την αποθήκευση και οργάνωση τους σε παραδοσιακές δομές ή βάσεις SQL. Επιπλέον, η απουσία κάποιου μοντέλου ικανού να τα επεξεργαστεί εμποδίζει την αξιοποίηση τους, με αποτέλεσμα να μη μπορούμε να εξάγουμε από αυτά πολύτιμες πληροφορίες. Στο πλαίσιο αυτό γίνονται

διάφορες προσπάθειες για την υλοποίηση αλγορίθμων και δομών που θα αποτελέσουν την βάση για τον εκσυγχρονισμό του συστήματος υγείας.
[4]

2 Διαχείριση και ανάλυση δεδομένων

Τα Big Data Analytics στην υγεία μπορούν να ταξινομηθούν σε [5] :

- **Περιγραφική Αναλυτική:** αναφέρεται στην περιγραφή ιατρικών καταστάσεων
- **Διαγνωστική Αναλυτική:** εξηγεί τους λόγους και παράγοντες πίσω από την εμφάνιση συγκεκριμένων καταστάσεων. Για παράδειγμα,επιχειρεί να κατανοήσει τους λόγους πίσω από τις συχνές επανεισαγωγές μερικών ασθενών, χρησιμοποιώντας διάφορες μεθόδους, όπως η συσταδοποίηση και τα δέντρα απόφασης.
- **Προγνωστική Αναλυτική:** επικεντρώνεται στην δυνατότητα να προβλέψουμε κάποιο γεγονός-πρόβλημα υγείας διαπιστώνοντας τάσεις και πιθανότητες. Αυτές οι μέθοδοι βασίζονται κυρίως σε ML τεχνικές.
- **Καθοδηγητική Αναλυτική:** η ανάλυση για την καλύτερη δυνατή απόφαση. Για παράδειγμα, η αποφυγή κάποιας θεραπείας με κριτήριο πιθανές παρενέργειες και επιπλοκές.



Εικ. 1. Κατηγορίες Ανάλυσης

Τα Big Data είναι πολύ μεγάλος όγκος δεδομένων που παρουσιάζουν μεγάλη ποικιλία και παράγονται με ραγδαίους ρυθμούς. Τα δεδομένα αυτά είναι απαραίτητα για την βελτιστοποίηση της βιοϊατρικής έρευνας και του συστήματος υγείας. Η μορφή τους, που δεν επιτρέπει εύκολη επεξεργασία και διαχείριση, αποτελεί τη μεγαλύτερη πρόκληση στην προσπάθεια να γίνουν διαθέσιμα στην ευρύτερη επιστημονική-ιατρική κοινότητα. Τα δεδομένα πρέπει να είναι αποθηκευμένα σε κάποια μορφή, που προσφέρει εύκολη πρόσβαση και επεξεργασία, ώστε η ανάλυσή τους να είναι αποτελεσματική. Ειδικότερα για τον τομέα της υγείας, η δημιουργία εργαλείων και πρωτοκόλλων τελευταίας τεχνολογίας που θα αντλούν στοιχεία από διαφορετικούς κλάδους(βιολογία, στατιστική κ.α)

αποτελεί μία ακόμη μεγάλη πρόκληση. Τα εργαλεία αυτά πρέπει να προσφέρουν δυνατότητες συλλογής και εξόρυξης δεδομένων, αποθήκευσης σε δίσκους τοπικά ή/και στο νέφος αλλά και μηχανικής μάθησης ώστε να μπορεί να εξαχθεί γνώση από τα παραγόμενα δεδομένα. Κύριο έργο είναι να σχολιάσουμε, να εντάξουμε και να παρουσιάσουμε τα πολύπλοκα δεδομένα στο χώρο αυτό για να γίνουν κατανοητά, διότι διαφορετικά είναι δύσκολο να χρησιμοποιηθούν στην έρευνα και στην πράξη. Σε αυτό το πλούσιο σε δεδομένα περιβάλλον είναι απαραίτητη η χρήση τεχνικών εξόρυξης δεδομένων για την αποτελεσματική τους αξιοποίηση. Μερικές από αυτές τις τεχνικές είναι[6]:

- **Κατηγοριοποίηση (Classification):** η διαδικασία οργάνωσης των δεδομένων σε κατηγορίες για την πιο αποτελεσματική και επικερδή χρήση τους. Η τεχνική αυτή εφαρμόζεται ευρέως στην εξόρυξη δεδομένων στον τομέα της υγείας. Η πρωτοβάθμια φροντίδα επηρεάζει την υγεία ενός παιδιού με τη διαχείριση των ασθενειών, την παροχή προληπτικού ελέγχου και υπηρεσιών αναβάθμισης. Στη Νέα Ζηλανδία, στο πλαίσιο εγγραφής σε κάποιο πάροχο υγείας κατά τη γέννα, αναλύονται ευκολότερα μοτίβα παιδικής θνησιμότητας. Σε μία έρευνα των MacRae et al. (2015)[7],προτάθηκε η δημιουργία ενός μοντέλου κατηγοριοποίησης των μη δομημένων δεδομένων-των σημειώσεων από τους γιατρούς με στόχο την επέκταση της χρήσης Pattern Recognition Over Standard Aesculupian Information Collections (PROSAIC) για την αναγνώριση της κατάστασης του αναπνευστικού συστήματος ενός παιδιού.Ο αλγόριθμος αυτός με τη χρήση Big Data επιτρέπει την ακριβή εκτίμηση της αναπνευστικής λειτουργίας ενός παιδιού μετά τη γέννα, καθώς και τους πόρους που θα χρειαστεί για την αντιμετώπιση τυχόν επιπλοκών. Σε άλλες έρευνες Frantzidis et al. (2010)[8], εφαρμόστηκε κατηγοριοποίηση για την αναγνώριση συναισθημάτων σε εφαρμογές βελτίωσης της υγείας με βάση το μοντέλο που λαμβάνει τα συναισθήματα ως μείξη δυο διαστάσεων. Συγκεκριμένα, ο διαχωρισμός έγινε με χρήση κανόνων κατηγοριοποίησης προερχόμενων από C4.5 αλγόριθμο που βασίζεται στην απόσταση Mahalanobis. Στη συνέχεια προωθεί το ρόλο πολυφυσιολογικών καταγραφών για την βελτίωση της διάκρισης μεταξύ των συναισθημάτων και την χρήση των μεταδεδομένων. Αυτό είναι εφικτό με τη βοήθεια δομών, όπως η XML, που διασυνδέουν τα διάφορα στοιχεία του συστήματος. Άλλες έρευνες παρουσιάζουν διάφορες παραλλαγές του αλγόριθμου για την αναγνώριση διάφορων μορφών καρκίνου Fan et al. (2011)[9], για την βοήθεια στη λήψη αποφάσεων με βάση εγκεφαλικά MRI Estella et al. (2012) και γενικότερα για μείωση των διαστάσεων των δεδομένων, επιλογή feature και κατηγοριοποίηση Azar & Hassanien (2015)[10].
- **Συσταδοποίηση(Clustering):** η διαδικασία κατά την οποία ομαδοποιούμε τα δεδομένα ώστε αυτά που ανήκουν στην ίδια συστάδα είναι ομοιότερα μεταξύ τους και λιγότερο με αυτά που ανήκουν σε άλλες συστάδες. Οι τεχνικές αυτές εφαρμόζονται ευρέως στον κλάδο της υγείας για διερευνητική ανάλυση δεδομένων σε εφαρμογές, όπως διαχωρισμό ασθενών, ανίχνευση ακραίων σημείων (outliers) στα δεδομένα, πρόβλεψη ασθενειών. Με χρήση k-means συσταδοποίησης οι Elbattah & Molloy (2017) προσπάθησαν να χωρίσουν τους ασθενείς με βάση τα δεδομένα λαμβάνοντας υπόψιν τυχόν ανωμαλίες, καθώς και την εξαγωγή χαρακτηριστικών που θεωρούνται ενδείξεις για την ποιότητα της περίθαλψης. Στη μελέτη τους οι Huang and Yao (2016)[11] πρότειναν την συσταδοποίηση πολυδιάστατων δεδομένων βασισμένα σε μία τεχνητή αποικία μυρμηγκιών δείχνοντας ότι είναι δυνατή

η χρήση της για συλλογή δεδομένων υγείας για περαιτέρω ανάλυση. Οι Paul and Hoque (2010)[12] πρότειναν την χρήση συσταδοποίησης για την πρόβλεψη της πιθανότητας εμφάνισης μιας ασθένειας. Ο αλγόριθμος μπορεί να διαχειριστεί συνεχή και διακριτά δεδομένα.

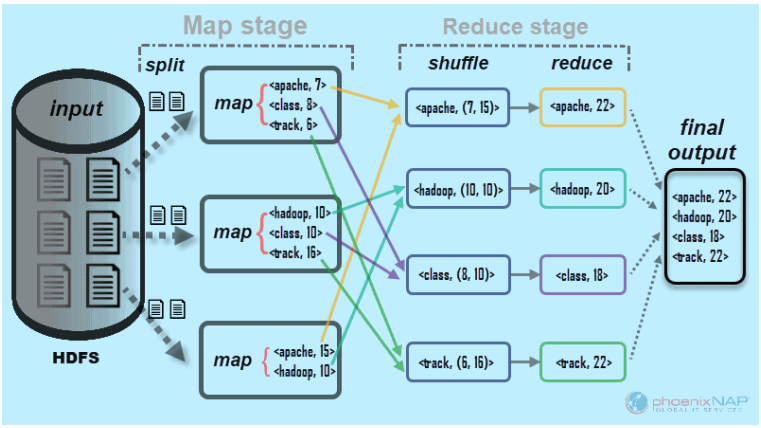
- **Παλινδρόμηση (Regression analysis):** είναι μια διαδεδομένη τεχνική ανάλυσης Big Data για τον υπολογισμό των σχέσεων μεταξύ μεταβλητών ή ιδιοτήτων. Σε διάφορες έρευνες χρησιμοποιήθηκαν τεχνικές παλινδρόμησης για την πρόβλεψη επιπλοκών ή και θνησιμότητας κατά τη διάρκεια περίθαλψης του ασθενούς ή μεταξύ 30 ημερών από την εγχείρηση του[Anderson and Chang (2015)[13]], αλλά και γενικότερα για την διαχείριση του ρίσκου στις ιατρικές εργασίες. Σύμφωνα με τους Oztekin et al. (2009)[14] χρησιμοποιώντας ένα μεγάλο data set μπορεί να δημιουργηθεί ένα Cox regression model που θα προβλέπει την επιβίωση μετά την μεταμόσχευση καρδιάς ή πνεύμονα.
- **Κανόνες Συσχέτισης (Association rules):** η εξόρυξη κανόνων συσχέτισης έχει στόχο να ανακαλύψει τη σχέση μεταξύ αντικειμένων σε μία μεγάλη βάση δεδομένων. Οι τυπικές τεχνικές εξόρυξης κανόνων είναι η Apriori και Frequent Pattern -tree growth. Είναι μια διαδικασία συνήθως 2 βημάτων.Στο πρώτο βήμα ανακαλύπτονται τα συχνότερα αντικείμενα και στο δεύτερο εξάγονται οι κανόνες γύρω από τη συσχέτιση των αντικειμένων αυτών με βάση το ενδιαφέρον που έχουν για το πρόβλημά μας. Οι Antonie et al. (2001) χρησιμοποίησαν έναν Apriori αλγόριθμο με στόχο να ανακαλύψουν κανόνες μεταξύ χαρακτηριστικών που προέρχονται από βάσεις μαστογραφίας και την κατηγορία που ανήκει η κάθε μία. Χρησιμοποιώντας αυτούς τους κανόνες κατασκεύασαν ένα σύστημα κατηγοριοποίησης που διαχωρίζει τις μαστογραφίες σε κανονικές, κακοήθεις και καλοήθεις,

Με άλλα λόγια είναι απαραίτητο να αναπτύξουμε/ αξιοποιήσουμε πλατφόρμες που θα προσφέρουν δυνατότητες εξόρυξης γνώσης από τα διάφορα ετερογενή δεδομένα ενώ παράλληλα προσφέρουν ασφαλή αποθήκευση, διαχείριση και επεκτασιμότητα. Κάποιες από τις σημαντικότερες υποδομές αναφέρονται παρακάτω.

2.1 Hadoop

Η φόρτωση μεγάλου όγκου δεδομένων σε μια μνήμη δεν αποτελεί καλή πρακτική στην διαχείριση των Big Data, ακόμα και όταν αναφερόμαστε σε ισχυρά υπολογιστικά συστήματα. Ο διαμοιρασμός των δεδομένων σε πολλαπλά συστήματα-κόμβους δεν είναι επίσης αποτελεσματικός καθώς μπορεί να χρειαστούν χιλιάδες για την επεξεργασία μεγάλου όγκου δεδομένων σε λογικά χρονικά πλαίσια. Ένα λογισμικό ανοιχτού κώδικα που είναι αρκετά διαδεδομένο στον τομέα αυτό είναι το Hadoop. Το Hadoop με την χρήση του Mapreduce αλγόριθμου κάνει δυνατή την παραγωγή και επεξεργασία μεγάλων datasets. Το Mapreduce, όπως προδίδει το όνομα, χρησιμοποιεί ένα ”map” για να αντιστοιχίζει την είσοδο σε ένα ζευγάρι κλειδί-τιμής και ”reduce” για να συνδυάσει όλες τις τιμές με το ίδιο κλειδί.Το πρόγραμμα αυτό παραλληλίζει με επιτυχία τον υπολογισμό, την αντιμετώπιση λαθών και την επικοινωνία μεταξύ πολλών μηχανών μεγάλης κλίμακας. Το Hadoop Distributed File System (HDFS) είναι το σύστημα αρχείων που προσφέρει έναν τρόπο αποθήκευσης των δεδομένων σε διαφορετικούς κόμβους-μέλη συστάδων καθιστώντας το μια επεκτάσιμη και αποτελεσματική λύση. Τα παραπάνω

σε συνδυασμό με άλλα εργαλεία, διαθέσιμα στην πλατφόρμα, έχουν επιτρέψει στους μελετητές να χρησιμοποιούν και να διαχειρίζονται τα data sets που προκύπτουν στο Healthcare. Κάποιες από τις εφαρμογές του Hadoop είναι η δυνατότητα συσχετισμού δεδομένων που αφορούν στην ποιότητα του αέρα με την εμφάνιση άσθματος,η ανάπτυξη φαρμάκων με την ανάλυση γονιδιακών και πρωτεομικών δεδομένων.



Εικ. 2. Hadoop Mapreduce

2.2 Image Analytics

Οι διάφοροι αλγόριθμοι Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης έχουν συμβάλει σημαντικά στην αύξηση των δυνατοτήτων γύρω από την επεξεργασία πληροφοριών εξάγοντας δεδομένα από τα Electronic Health Records. Για παράδειγμα, η επεξεργασία φυσικής γλώσσας, ένας κλάδος που αναπτύσσεται ραγδαία, δίνει τη δυνατότητα να επεξεργαστούμε συντακτικές δομές σε ελεύθερο κείμενο, βοηθούν στην αναγνώριση λόγου και στην εξαγωγή νοήματος απ'αυτά. Τα εργαλεία NLP παράγουν αρχεία, όπως μια περιγραφή μιας επίσκεψης στο γιατρό. Επιπλέον, δίνουν τη δυνατότητα υπαγορεύσης κάποιων κλινικών σημειώσεων. Το AI χρησιμοποιείται για την περιγραφή των Big Data στο Healthcare. Οι αλγόριθμοι αυτοί μετατρέπουν το διαγνωστικό σύστημα ιατρικών εικόνων σε αυτοματοποιημένο σύστημα λήψης αποφάσεων.

Κάποιες από τις πιο διαδεδομένες τεχνικές απεικόνισης στον τομέα της υγείας είναι:

- **Αξονική/Υπολογιστική Τομογραφία (CT)**. Οι Kavasidis et. al[15] πρότειναν ένα AI-powered pipeline, βασισμένο σε βαθιά μάθηση για αυτόματη διάγνωση COVID-19 και κατηγοριοποίηση των προβλημάτων που δημιουργεί στον πνεύμονα, με χρήση CT scans.
- **Απεικόνιση Μαγνητικού Συντονισμού (MRI) και (fMRI)**
- **X-ray**
- **Μοριακή Απεικόνιση**

– **Υπέρηχος**

Οι παραπάνω τεχνικές αιχμαλωτίζουν μεγάλης ποιότητας εικόνες ιατρικού ενδιαφέροντος. Με άλλα λόγια, καταγράφουν μεγάλο όγκο δεδομένων ενός ασθενούς. Τα δεδομένα αυτά παρότι χρησιμοποιούνται επιτυχώς, δεν αξιοποιούνται επαρκώς λόγω έλλειψης εξειδικευμένων επαγγελματιών για τη διάγνωση πολλών ασθενειών. Γι' αυτό το λόγο αναπτύχθηκαν αποδοτικά συστήματα όπως τα **Picture Archiving and Communication (PACS)**. Τα συστήματα αυτά προσφέρουν εύκολη αποθήκευση και πρόσβαση σε ιατρικές εικόνες και δεδομένα αναφορών, ενώ παράλληλα διανέμουν τις εικόνες σε τοπικούς κόμβους επεξεργασίας. Τα δεδομένα που διανέμονται μέσω των PACS είναι δομημένα με αποτέλεσμα να χάνεται ένας σημαντικός αριθμός πληροφοριών που περιέχεται στα ιατρικά αρχεία. Με τη βοήθεια του ML και της αναγνώρισης προτύπων γίνεται δυνατή η εξαγωγή βιοδεικτών από τον μεγάλο όγκο δεδομένων που παρέχονται από τις βιοϊατρικές απεικονίσεις έχοντας ως σκοπό την βελτίωση της διάγνωσης, θεραπείας και επίβλεψης των ασθενών. Επιπλέον, διάφορα εργαλεία γονιδιακής εγγραφής, κατάτμησης, οπτικοποίησης, ανακατασκευής, προσομοίωσης και διάχυσης έχουν δημιουργηθεί με σκοπό την ιατρική ανάλυση εικόνων και την εξόρυξη κρυμμένων πληροφοριών. Για παράδειγμα, το **Visualization Toolkit** `https://vtk.org/` επιτρέπει ισχυρή επεξεργασία και ανάλυση δισδιάστατων και τρισδιάστατων εικόνων από ιατρικές εξετάσεις. Άλλα λογισμικά, όπως το **SPM** αλλά και τα **GIMIAS, Elastix, MITK** επιτρέπουν την επεξεργασία και ανάλυση πέντε διαφορετικών τύπων εγκεφαλικών εικόνων(e.g. MRI, fMRI, PET, CT-Scan and EEG),

2.3 Εμπορικές πλατφόρμες

Πέρα από τις διάφορες δομές (frameworks) και αλγόριθμους που αποτελούν τη βάση διαχείρισης δεδομένων στο Healthcare, μεγάλες εταιρείες έχουν προσπαθήσει να δώσουν λύσεις σε εμπορικό επίπεδο. Αυτό το κάνουν με τη δημιουργία πλατφορμών. Οι πλατφόρμες αυτές πραγματοποιούν καλύτερες αναλύσεις με τη χρήση AI σε δημοσιευμένα αποτελέσματα εξετάσεων, κειμενικά δεδομένα, δεδομένα εικόνων ώστε να εξάγουν ουσιαστικά συμπεράσματα για την κατάσταση ενός ασθενούς. Κάποιες τέτοιες πλατφόρμες αναφέρονται παρακάτω:

– **IBM Watson Health:** είναι μια πλατφόρμα τεχνητής νοημοσύνης για την ανταλλαγή και την ανάλυση δεδομένων υγείας μεταξύ νοσοκομείων, παρόχων και ερευνητών.
– **Roam Analytics:** παρέχει την δυνατότητα εξόρυξης πληροφορίας από μεγάλα μηδομημένα δεδομένα και τις υποδομές για αξιοποίηση NLP αλγορίθμων.
– **Flatiron Health:** προσφέρει εφαρμογές για την οργάνωση και βελτίωση δεδομένων ογκολογίας, για λύσεις στην έρευνα και θεραπεία του καρκίνου.
– **Health Fidelity:**η πλατφόρμα αυτή παρέχει υπηρεσίες - λύσεις για τη διαχείριση ρίσκου στη ροή εργασιών του Healthcare και μεθόδους βελτιστοποίησης της.
– **Apixio:** δίνει την δυνατότητα επεξεργασίας health records σε μορφή PDF αλλά και κλινικών δεδομένων γενικότερα για εξόρυξη γνώσης.
– **Enlitic:** προσφέρει δυνατότητες βαθιάς μάθησης σε μεγάλα σύνολα δεδομένων από κλινικά τεστ για καλύτερη διάγνωση.

Ειδικότερα η IBM Watson Health, που είναι μια από τις μοναδικές τεχνολογικές εφαρμογές της εταιρείας IBM, στοχεύει σε αναλύσεις Big Data σε σχεδόν κάθε επαγγελματικό τομέα. Αυτή η πλατφόρμα χρησιμοποιεί εκτεταμένα αλγορίθμους βασισμένους σε εκμάθηση μηχανών και τεχνητή νοημοσύνη, ώστε να εξαγάγει τις μέγιστες πληροφορίες από ελάχιστες εισροές. Ο IBM Watson είναι ένας υπερυπολογιστής, που επιβάλλει το σχήμα ενσωμάτωσης ενός ευρέος φάσματος τομέων της υγείας για την παροχή ουσιαστικών και δομημένων δεδομένων. Σε μια προσπάθεια να αποκαλυφθούν νέα φάρμακα για συγκεκριμένα μοντέλα καρκινικών νοσημάτων, ο IBM Watson και η Pfizer έχουν διαμορφώσει μια παραγωγική συνεργασία για την επιτάχυνση της ανακάλυψης νέων συνδυασμών ανοσο-ογκολογίας

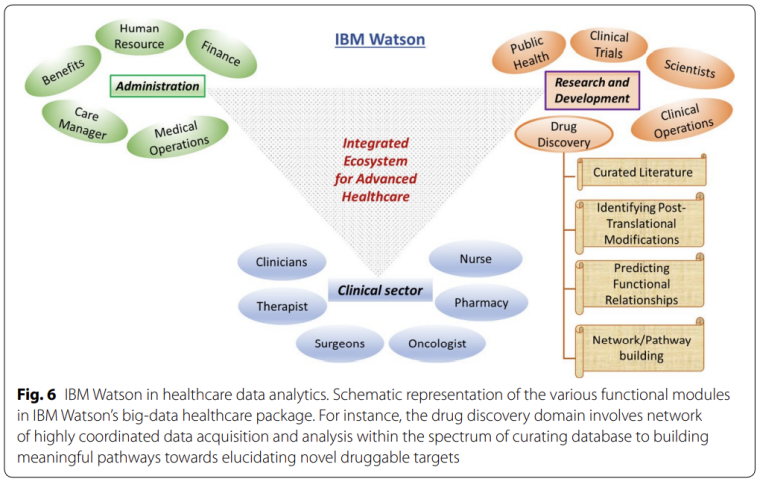


Fig. 6 IBM Watson in healthcare data analytics. Schematic representation of the various functional modules in IBM Watson's big-data healthcare package. For instance, the drug discovery domain involves network of highly coordinated data acquisition and analysis within the spectrum of curating database to building meaningful pathways towards elucidating novel druggable targets

Εικ. 3. IBM Watson Health

3 Big Data και COVID-19

Ο Covid-19, ένας ιός που εξαπλώθηκε ταχύτατα σε όλο τον πλανήτη, είχε σοβαρές επιπτώσεις στην παγκόσμια υγεία και οικονομία. Τη στιγμή που γράφεται το παρόν κείμενο, πέρα από την παραγωγή εμβολίων, δεν έχει βρεθεί ουσιαστική λύση για την σωστή λειτουργία των φορέων υγείας. Στο αποκορύφωμα της πανδημίας οι υπάρχουσες ιατρικές υποδομές αποδείχθηκαν ανεπαρκείς να αντεπεξέλθουν στον αυξημένο όγκο ασθενών. Στην κατάσταση αυτή οι πάροχοι υπηρεσιών υγείας και οι ίδιοι οι ασθενείς αναγκάστηκαν να λάβουν δύσκολες αποφάσεις λόγω της απουσίας στοιχειώδους πληροφορίας για τον ιό. Η κατάσταση αυτή κατέστησε απαραίτητη την δημιουργία εργαλείων ικανών να αυξήσουν τους πόρους στον κλάδο της υγείας. Μέθοδοι Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης μπορούν να εφαρμοστούν για την κατανόηση υποομάδων ασθενών, για την λήψη ιατρικών αποφάσεων και για την βελτίωση του συστήματος, λαμβάνοντας ως βάση τον τεράστιο όγκο δεδομένων που παράγονται καθη-

μερινά. Αρχικά, είναι δυνατή η χρήση τέτοιων εργαλείων στην αρχική επίσκεψη και εκτίμηση συμπτωματικών ατόμων εκτός ιατρείων, διαχωρίζοντας άτομα υψηλού κινδύνου,που είναι πιθανότερο να χειροτερεύσουν από αυτά χαμηλού κινδύνου. Με αυτό τον τρόπο οι φορείς μπορούν να επικεντρωθούν στα άτομα που το χρειάζονται περισσότερο,ενώ παράλληλα μπορούν να αξιοποιηθούν εφαρμογές εικονικής φροντίδας και να αποφευχθεί η άσκοπη επίσκεψη στο νοσοκομείο. Τα μοντέλα μηχανικής μάθησης θα μπορούσαν να προβλέψουν την πιθανότητα μιας επιβεβαιωμένης διάγνωσης Covid-19 και τη σοβαρότητα της παίρνοντας απαντήσεις από συμπτωματικά άτομα και ενισχύοντας τες με κλινικές πληροφορίες από τα EHR. Το μοντέλο θα μπορούσε επίσης να προβλέπει τα επίπεδα δύσπνοιας μέσω τηλεφωνικής κλήσης με τον ασθενή, με κριτήριο το συναίσθημα, τη βραχνάδα και το βήχα μέσα στο λόγο. Παράλληλα, είναι δυνατό με βάση τη σοβαρότητα να υποδειχθεί και το απαραίτητο επίπεδο φροντίδας. Με άλλα λόγια τα Big Data σε συνδυασμό με τα ML μοντέλα μπορούν να ανακουφίσουν τα ήδη γεμάτα νοσοκομεία.

Το τμήμα επειγόντων περιστατικών (ED) αποτελεί τον δεύτερο κλάδο που μπορεί να αξιοποιηθούν τα Big Data μέσω ML/AI. Οι γιατροί καλούνται να συλλέξουν, να αφομοιώσουν και να αναλύσουν μεγάλο όγκο δεδομένων σε μικρό χρονικό διάστημα, υπό την πίεση της σοβαρότητας των συμπτωμάτων και της ανάγκης για άμεση περίθαλψη. Σε ασθενείς με Covid-19 μπορούν να παρέχουν υποστήριξη σε κάθε στάδιο υπολογίζοντας την πιθανότητα εισόδου στο νοσοκομείο(triage),σκιαγραφώντας τους κινδύνους με πραγματικά δεδομένα προερχόμενα από την κλινική εκτίμηση και προβλέποντας την πορεία της υγείας του ασθενούς, καθώς και τα αποτελέσματα τυχόν άμεσης διασωλήνωσης. Μια ακόμα είσοδος για τα μοντέλα αυτά θα μπορούσε να είναι οι ακτινογραφίες και τομογραφίες, που αναφέρθηκαν παραπάνω, και με τις οποίες έχει αποδειχθεί ότι μπορούμε να αναγνωρίσουμε σοβαρές περιπτώσεις COVID-19. Όσον αφορά,λοιπόν, τα επείγοντα,ο υπολογισμός της πιθανότητας ανεπάρκειας του αναπνευστικού δίνει τη δυνατότητα στους γιατρούς να δώσουν προτεραιότητα και πόρους, που είναι ελλιπείς, σε όσα άτομα χρειάζεται μειώνοντας σημαντικά τις απώλειες. Ένα άλλο στάδιο που μπορούμε να παρέμβουμε είναι μεταξύ της μεταφοράς του ασθενούς από τα επείγοντα στο θάλαμο αλλά και καθ' όλη τη διάρκεια της παραμονής του στο νοσοκομείο. Ο συγκλονιστικός αριθμός των ατόμων που εισάγονται στο νοσοκομείο κατά τη διάρκεια της πανδημίας κατακλύζει και τα πιο οργανωμένα νοσοκομεία. Σύμφωνα με έρευνες, η μεγάλη εισροή ασθενών, εμποδίζει την έγκαιρη θεραπεία τους, και όσο αυξάνεται ο αριθμός τους τόσο αυξάνεται η πιθανότητα για ανεπιθύμητα συμβάντα. Στη δύσκολη κατάσταση που έχει δημιουργήσει ο COVID-19 , κάθε καθυστέρηση μπορεί να αποβεί μοιραία, οδηγώντας σε ανθρώπινες απώλειες. Υπό αυτές τις συνθήκες, τεχνικές ML/AI επιτρέπουν την καλύτερη και γρήγορη κατανόηση κάθε περιστατικού, συμβάλλοντας σημαντικά στην διαχείριση του μεγάλου όγκου περιστατικών. Με είσοδο τα Big Data που παράγονται στον χώρο της υγείας (εργαστηριακές εξετάσεις, τεστ κλπ) υπολογίζουν το ρίσκο.

Οι επιστήμονες και κυρίως οι επιδημιολόγοι κατασκευάζουν μοντέλα για να κάνουν προβλέψεις ώστε να προετοιμαστούν και δυνητικά να μετριάσουν την εξάπλωση ασθενειών (π.χ. Covid-19) και τον αντίκτυπό τους.Οι προβλέψεις αυτές μπορεί να είναι:

- **Βραχυπρόθεσμες** προβλέψεις (4-8 ώρες): μπορούν να χρησιμοποιηθούν από νοσοκόμες και γιατρούς για να δώσουν προτεραιότητα στη φροντίδα του ασθενούς.

- **Μεσοπρόθεσμες** προβλέψεις (12-24 ώρες): αφορούν στην αναγνώριση ασθενών που είναι λιγότερο πιθανό να εμφανίσουν λειτουργική επιδείνωση (decompensation) κάποιου οργάνου (π.χ. πνεύμονα).Επιπλέον,είναι δυνατή η προσαρμογή της φροντίδας που θα δεχτούν μέχρι την έξοδο τους από το νοσοκομείο. [16]
- **Μακροπρόθεσμες** προβλέψεις (περισσότερο από 24 ώρες): δίνουν τη δυνατότητα σε άτομα ευθύνης να διαμοιράσουν πολύτιμους πόρους , όπως αναπνευστήρες, κλίνες και απαραίτητο προσωπικό μεταξύ των ασθενών.

Παρακάτω παρουσιάζεται ένα μοντέλο ML/AI στα στάδια εξέτασης και εισαγωγής του ασθενούς στο νοσοκομείο. [17]

A. Οι μαύροι αστερίσκοι αντιπροσωπεύουν τα πολλά σημεία αποφάσεων κατά τη διάρκεια της φροντίδας του ασθενούς τα οποία θα μπορούσαν να επαυξηθούν από ML/AI. Οι πράσινες γραμμές αντιπροσωπεύουν τις αρνητικές διαγνώσεις ή την ανάρρωση. Οι πορτοκαλί και κόκκινες γραμμές το μικρό ή μεγάλο ρίσκο αντίστοιχα, όπως ορίζονται από το μοντέλο.

B. Μία εκτεταμένη βάση δεδομένων για τον COVID-19, παρουσιάζει τον αυξημένο αριθμό δεδομένων σχετικά με τον ιό στο Northwell Health center της Νέας Υόρκης. Αυτή η αύξηση αποτελεί βάση για την εφαρμογή μοντέλων ML/AI για την υποστήριξη του συστήματος υγείας.

C. Εξελισσόμενα προφίλ ασθενών και η συχνότητα με την οποία παίρνουν εξιτήριο από το νοσοκομείο μπορούν να επηρεάσουν την αποτελεσματικότητα του μοντέλου. Για παράδειγμα, η μέση διάρκεια παραμονής στο νοσοκομείο των ασθενών που πεθαίνουν (σκούρο μπλέ με τετράγωνα σύμβολα) και αυτών που επιζούν (γαλάζιο με τρίγωνα σύμβολα) είχε απόκλιση τον μήνα Απρίλιο. Αυτό δηλώνει ότι ένα μοντέλο που είχε καλή αρχική απόδοση μπορεί να παρουσιάσει πτώση λόγω της διαφοράς μεταξύ των ασθενών.

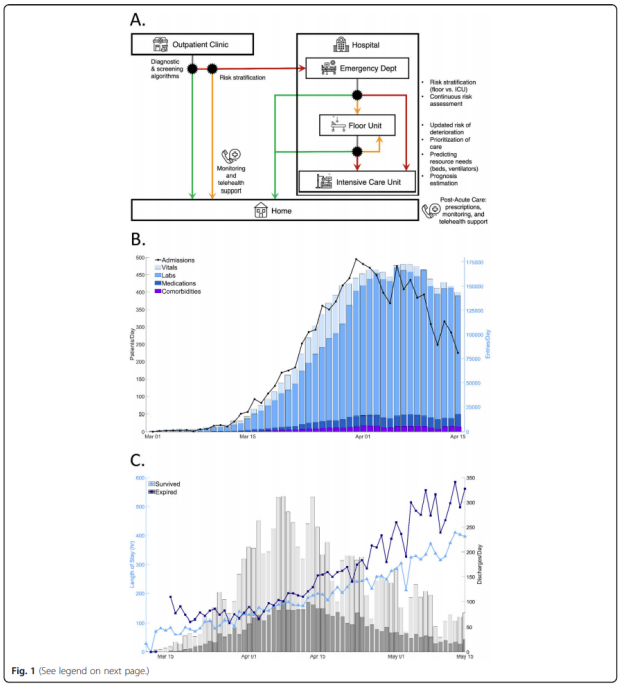


Fig. 1 (See legend on next page.)

Εικ. 4. Northwell COVID-19 Hospital Pathway

4 Προκλήσεις

4.1 Αποθήκευση Δεδομένων (Data Storage)

Οι τρέχουσες δυσκολίες στην αποθήκευση δεδομένων οφείλονται κυρίως στα μεγάλα κόστη. Ο μεγάλος αριθμός ιατρικών δεδομένων είναι μία από τις πηγές του κόστους αποθήκευσης. Με την ανάπτυξη των τεχνικών εξόρυξης αντλούμε συνεχώς περισσότερες πληροφορίες. Τα δεδομένα που παράγονται σε αυτή την ιατρική "βιομηχανια" κυμαίνονται από εικόνες σχετικές με τη διάγνωση ενός ασθενούς μέχρι την παθολογική ανάλυση χαρτών. [18] Για παράδειγμα, τοπικά ιατρικά δεδομένα εξάγονται συνήθως από μια περιοχή με εκατομμύρια άτομα και εκατοντάδες φορείς υγείας και ο αριθμός τους αυξάνεται συνεχώς. Τα δεδομένα ενός ασθενούς τυπικά πρέπει να διατηρούνται για περισσότερο από 50 χρόνια. Το αρχείο του περιέχει αφενός έναν μεγάλο αριθμό από online ή/και δεδομένα πραγματικού χρόνου αφετέρου μια ποικιλία από δεδομένα, όπως διάγνωση και προτάσεις φαρμάκων, σε CD. Επίσης περιέχει διάφορες δομές από πίνακες δομημένων δεδομένων, μη δομημένα αρχεία κειμένου, ιατρικές εικόνες και άλλες πληροφορίες. Το τεράστιο μέγεθος των δεδομένων αναγκαστικά αυξάνει σημαντικά το κόστος και τη δυσκολία αποθήκευσης τους. Υπάρχουν ακόμα κόστη που σχετίζονται με την κίνηση των δεδομένων από ένα μέρος σε ένα άλλο καθώς και με την ανάλυση τους. Τέλος, οι τύποι των ιατρικών δεδομένων είναι διαφορετικοί και ποικίλοι, αφού περιλαμβάνουν αριθμητικά δεδομένα για την καταγραφή ελέγχου ασθενειών, εικόνες, αρχεία ήχου και βίντεο. Αυτά τα μη-δομημένα δεδομένα είναι πιο δύσκολα να αποθηκευτούν, να αναλυθούν και επεξεργαστούν σε σύγκριση με τα παραδοσιακά δομημένα, αυξάνοντας το κόστος αποθήκευσης. Η διασφάλιση της ασφάλειας και ιδιωτικότητας κατά την γενικότερη τους διαχείριση, αποθήκευση, εξαγωγή και κατέβασμα δεδομένων σχετικά με τους ασθενείς, αποτελεί μια σοβαρή πρόκληση.(Youssef, 2014)

4.2 Περιορισμένη τυποποίηση και διαλειτουργικότητα δεδομένων (Limited data standardization and interoperability)

Οι σύγχρονες προδιαγραφές και τεχνολογίες δεν πληρούν τις προϋποθέσεις για την ένταξη των εφαρμογών των Big Data στον τομέα της υγείας. Αρχικά, τα δεδομένα δεν παρουσιάζουν κάποια ενιαία standards, συνεπή μορφή περιγραφής και μεθόδους παρουσίασης. Επιπλέον, η ενσωμάτωση τριών επιπέδων για τους διάφορους τύπους δεδομένων είναι ιδιαίτερα δύσκολη. Ταυτόχρονα, κάθε βάση δεδομένων χρησιμοποιεί διαφορετικό λογισμικό και τύπους δεδομένων, καθιστώντας την μεταξύ τους σύγκριση, ανάλυση, μεταφορά και διαμοιρασμό δύσκολες διαδικασίες. (Chawla & Davis, 2013 [19]; Mohr, Burns, Schueller, Clarke, & Klinkman, 2013 [20]; W. Raghupathi & Raghupathi, 2014 [21]). Συγκεκριμένα στο Healthcare ο διαμοιρασμός δεδομένων μεταξύ των φορέων είναι πολλές φορές περιορισμένος έως και αδύνατος (π.χ Κίνα) λόγω εμποδίων στη μετάδοση της πληροφορίας (Kruse, Goswamy, Raval, & Marawi, 2016 [22]). Με την παγκοσμιοποίηση των Big Data το Healthcare θα πρέπει να αντιμετωπίσει τις διαφορετικές γλώσσες, την ορολογία και τις διαφορετικές τεχνικές τυποποίησης που χαρακτηρίζουν τα αρχεία από τα οποία αντλούνται τα δεδομένα.

4.3 Ιδιωτικότητα δεδομένων (Data privacy)

Τα δεδομένα στον τομέα της υγείας είναι ευαίσθητου περιεχομένου και άμεσα συνδεδεμένα με το άτομο που τα παράγει (centralized) σε αντίθεση με άλλους τομείς. Υπάρχουν πολλοί ενδοιασμοί σχετικά με την εχεμύθεια (Mancini, 2014b; D. C. Mohr et al., 2013 [20]). Όμως, για το πρόβλημα της ιδιωτικότητας των δεδομένων του ασθενούς δεν έχουν βρεθεί τέλειες λύσεις. Η διαρροή δεδομένων μπορεί να έχει απρόβλεπτα αποτελέσματα, όπως στιγματισμό και διάκριση. Τα Big Data εγκυμονούν πολλούς κινδύνους για τα προσωπικά ιατρικά δεδομένα. Μία πηγή κινδύνου είναι τα ίδια τα δεδομένα που μπορούν να αντιγραφούν και διατηρηθούν χωρίς χωρικούς και χρονικούς περιορισμούς. Μια άλλη είναι ότι ακόμα και αν η βάση χρησιμοποιεί κρυπτογραφημένα δεδομένα, αυτά παραμένουν συνδεδεμένα με έναν χρήστη που με διάφορες τεχνικές είναι δυνατόν να αναγνωριστεί (Ward,2014) [23].

5 Συμπεράσματα

Τα Big Data έχουν μια πληθώρα πλεονεκτημάτων στον χώρο της υγείας. Ο μεγάλος όγκος δεδομένων συμβάλλει σημαντικά στην δημιουργία και εκπαίδευση αλγορίθμων και μοντέλων της μηχανικής μάθησης και τεχνητής νοημοσύνης οι οποίες με τη σειρά τους χρησιμοποιούνται για τη βελτίωση των υπηρεσιών στον τομέα αυτό. Προβλέψεις που βοηθούν στην καλύτερη διαχείριση των πόρων μιας μονάδας, η ποιοτικότερη περίθαλψη που εστιάζει στις ανάγκες κάθε ασθενή, η σημαντική μείωση της θνησιμότητας λόγω της άμεσης και ταχείας αντιμετώπισης των διάφορων προβλημάτων και τέλος η πρόβλεψη των επιπλοκών είναι μονάχα κάποιες από τις πολυάριθμες εφαρμογές τους. Επιπλέον, η αξιοποίηση των Big Data μπορεί να συμβάλλει στην αντιμετώπιση πανδημιών, όπως παρουσιάστηκε για την περίπτωση του Covid-19. Παρ' όλα αυτά, για την σωστή και αποτελεσματική χρήση τους στο χώρο της υγείας πρέπει να αντιμετωπίσουμε πολλές προκλήσεις. Με την συνεχή εξέλιξη των blockchain αυξάνονται και οι δυνατότητες για τη δημιουργία αποκεντρωμένων εφαρμογών που θα προσφέρουν ιδιωτικότητα και διαφάνεια (transparency). Σε αυτό το πλαίσιο είναι δυνατή η δημιουργία ενός EHR το οποίο θα αξιοποιεί ένα υπάρχον blockchain και θα δίνει τη δυνατότητα ασφαλούς μεταφοράς των ιατρικών δεδομένων, ευκολότερη διαχείριση αλλά και επεκτασιμότητα [24]. Τέλος, πρέπει να σημειωθεί ότι τα ηλεκτρονικά δεδομένα για την υγεία παραμένουν σε μεγάλο βαθμό αναξιοποίητα,και κατά συνέπεια είναι επιτακτική η ανάγκη μετατροπής των μη επεξεργασμένων δεδομένων σε χρήσιμες και αποτελεσματικές πληροφορίες.

Βιβλιογραφία

[1] Wikipedia contributors. Big data — Wikipedia, the free encyclopedia, 2021. URL https://en.wikipedia.org/w/index.php?title=Big_data&oldid=1017308489. [Online; accessed 14-April-2021].

[2] Ishwarappa and J. Anuradha. A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia Computer Science*, 48:319–324, 2015. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2015.04.188. URL https://www.sciencedirect.com/science/article/pii/S1877050915006973. International Conference on Computer, Communication and Convergence (ICCC 2015).

[3] Rebecca Hubbard. Analysis of big healthcare databases.

[4] Naoual El aboudi and Laila Benhlima. Big data management for healthcare systems: Architecture, requirements, and implementation. *Advances in Bioinformatics*, 2018:4059018, Jun 2018. ISSN 1687-8027. doi: 10.1155/2018/4059018. URL https://doi.org/10.1155/2018/4059018.

[5] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):54, Jun 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0217-0. URL https://doi.org/10.1186/s40537-019-0217-0.

[6] Liang Hong, Mengqi Luo, Ruixue Wang, Peixin Lu, Wei Lu, and Long Lu. Big data in health care: Applications and challenges. *Data and Information Management*, 2(3):175–197, 2018. doi: doi:10.2478/dim-2018-0014. URL https://doi.org/10.2478/dim-2018-0014.

[7] Jayden Macrae, Ben Darlow, Lynn Mcbain, O Jones, Maria Stubbe, Nikki Turner, and A Dowell. Accessing primary care big data: The development of a software algorithm to explore the rich content of consultation records. *BMJ open*, 5, 08 2015. doi: 10.1136/bmjopen-2015-008160.

[8] Christos Frantzidis, Charalampos Bratsas, Manousos Klados, Evdokimos Konstantinidis, Chrysa Lithari, Ana Vivas, Christos Papadelis, Eleni Kaldoudi, Costas Pappas, and Panagiotis Bamidis. On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 14:309–18, 03 2010. doi: 10.1109/TITB.2009.2038481.

[9] Xiomara Blanco, Sara Rodríguez, Juan M. Corchado, and Carolina Zato. Case-based reasoning applied to medical diagnosis and treatment. In Sigeru Omatu, José Neves, Juan M. Corchado Rodriguez, Juan F Paz Santana, and Sara Rodríguez Gonzalez, editors, *Distributed Computing and Artificial Intelligence*, pages 137–146, Cham, 2013. Springer International Publishing. ISBN 978-3-319-00551-5.

[10] Ahmad Taher Azar and Aboul Ella Hassanien. Dimensionality reduction of medical big data using neural-fuzzy classifier. *Soft Computing*, 19(4):1115–1127, Apr 2015. ISSN 1433-7479. doi: 10.1007/s00500-014-1327-4. URL https://doi.org/10.1007/s00500-014-1327-4.

[11] Ravichandran C. Gopalakrishnan and Veerakumar Kuppusamy. Ant colony optimization approaches to clustering of lung nodules from ct images. *Computational and mathematical methods in medicine*, 2014:572494–572494, 2014. ISSN 1748-6718. doi: 10.1155/2014/572494. URL https://pubmed.ncbi.nlm.nih.gov/25525455. 25525455[pmid].

[12] R. Paul and A. S. L. Hoque. Clustering medical data to predict the likelihood of diseases. *2010 Fifth International Conference on Digital Information Management (ICDIM)*, pages 44–49, 2010.

[13] Jamie E. Anderson and David C. Chang. Using Electronic Health Records for Surgical Quality Improvement in the Era of Big Data. *JAMA Surgery*, 150(1):24–29, 01 2015. ISSN 2168-6254. doi: 10.1001/jamasurg.2014.947. URL https://doi.org/10.1001/jamasurg.2014.947.

[14] Asil Oztekin, Dursun Delen, and Zhenyu (James) Kong. Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology. *International Journal of Medical Informatics*, 78(12):e84–e96, 2009. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2009.04.007. URL https://www.sciencedirect.com/science/article/pii/S1386505609000707. Mining of Clinical and Biomedical Text and Data Special Issue.

[15] Matteo Pennisi, Isaak Kavasidis, Concetto Spampinato, Vincenzo Schininà, Simone Palazzo, Francesco Rundo, Massimo Cristofaro, Paolo Campioni, Elisa Pianura, Federica Di Stefano, Ada Petrone, Fabrizio Albarello, Giuseppe Ippolito, Salvatore Cuzzocrea, and Sabrina Conoci. An explainable ai system for automated covid-19 assessment and lesion categorization from ct-scans, 2021.

[16] Hoyt Burdick, Carson Lam, Samson Mataraso, Anna Siefkas, Gregory Braden, R. Phillip Dellinger, Andrea McCoy, Jean-Louis Vincent, Abigail Green-Saxena, Gina Barnes, Jana Hoffman, Jacob Calvert, Emily Pellegrini, and Ritankar Das. Prediction of respiratory decompensation in covid-19 patients using machine learning: The ready trial. *Computers in Biology and Medicine*, 124:103949, 2020. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2020.103949. URL https://www.sciencedirect.com/science/article/pii/S0010482520302845.

[17] Shubham Debnath, Douglas P. Barnaby, Kevin Coppa, Alexander Makhnevich, Eun Ji Kim, Saurav Chatterjee, Viktor Tóth, Todd J. Levy, Marc d. Paradis, Stuart L. Cohen, Jamie S. Hirsch, Theodoros P. Zanos, Lance B. Becker, Jennifer Cookingham, Karina W. Davidson, Andrew J. Dominello, Louise Falzon, Thomas McGinn, Jazmin N. Mogavero, Gabrielle A. Osorio, and the Northwell COVID-19 Research Consortium. Machine learning to assist clinical decision-making during the covid-19 pandemic. *Bioelectronic Medicine*, 6(1):14, Jul 2020. ISSN 2332-8886. doi: 10.1186/s42234-020-00050-8. URL https://doi.org/10.1186/s42234-020-00050-8.

[18] Katerina Rohlenova, Jermaine Goveia, Melissa García-Caballero, Abhishek Subramanian, Joanna Kalucka, Lucas Treps, Kim D. Falkenberg, Laura P.M.H. de Rooij, Yingfeng Zheng, Lin Lin, Liliana Sokol, Laure-Anne Teuwen, Vincent Geldhof, Federico Taverna, Andreas Pircher, Lena-Christin Conradi, Shawez Khan, Steve Stegen, Dena Panovska, Frederik De Smet, Frank J.T. Staal, Rene J.

Mclaughlin, Stefan Vinckier, Tine Van Bergen, Nadine Ectors, Patrik De Haes, Jian Wang, Lars Bolund, Luc Schoonjans, Tobias K. Karakach, Huanming Yang, Geert Carmeliet, Yizhi Liu, Bernard Thienpont, Mieke Dewerchin, Guy Eelen, Xuri Li, Yonglun Luo, and Peter Carmeliet. Single-cell rna sequencing maps endothelial metabolic plasticity in pathological angiogenesis. *Cell Metabolism*, 31(4):862–877.e14, 2020. ISSN 1550-4131. doi: https://doi.org/10.1016/j.cmet.2020.03.009. URL https://www.sciencedirect.com/science/article/pii/S1550413120301248.

[19] Nitesh V. Chawla and Darcy A. Davis. Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, 28(S3):660–665, June 2013. doi: 10.1007/s11606-013-2455-8. URL https://doi.org/10.1007/s11606-013-2455-8.

[20] David C. Mohr, Michelle Nicole Burns, Stephen M. Schueller, Gregory Clarke, and Michael Klinkman. Behavioral intervention technologies: Evidence review and recommendations for future research in mental health. *General Hospital Psychiatry*, 35(4):332–338, July 2013. doi: 10.1016/j.genhosppsych.2013.03.008. URL https://doi.org/10.1016/j.genhosppsych.2013.03.008.

[21] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1), February 2014. doi: 10.1186/2047-2501-2-3. URL https://doi.org/10.1186/2047-2501-2-3.

[22] Clemens Scott Kruse, Rishi Goswamy, Yesha Raval, and Sarah Marawi. Challenges and opportunities of big data in health care: A systematic review. *JMIR Medical Informatics*, 4(4):e38, November 2016. doi: 10.2196/medinform.5359. URL https://doi.org/10.2196/medinform.5359.

[23] Elizabeth Ward, Carol DeSantis, Anthony Robbins, Betsy Kohler, and Ahmedin Jemal. Childhood and adolescent cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, 64(2):83–103, January 2014. doi: 10.3322/caac.21219. URL https://doi.org/10.3322/caac.21219.

[24] André Mayer, Cristiano André da Costa, and Rodrigo Righi. Electronic health records in a blockchain: A systematic review. *Health Informatics Journal*, 26: 146045821986635, 09 2019. doi: 10.1177/1460458219866350.