ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εφαρμογές Τεχνητής Νοημοσύνης για την Βελτίωση της Ποιότητας Ζωής σε Ασθενείς με Άνοια

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΥΓΕΡΙΝΟΣ ΠΕΤΡΟΣ

Επιβλέπων : Αθανάσιος Βουλόδημος
Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εφαρμογές Τεχνητής Νοημοσύνης για την Βελτίωση της Ποιότητας Ζωής σε Ασθενείς με Άνοια

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΥΓΕΡΙΝΟΣ ΠΕΤΡΟΣ

Επιβλέπων : Αθανάσιος Βουλόδημος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Σεπτεμβρίου 2024.

Αθανάσιος Βουλόδημος	*Νικόλαος Δουλάμης*	*Ανδρέας-Γεώργιος Σταφυλοπάτης*
Καθηγητής Ε.Μ.Π.	*Καθηγητής Ε.Μ.Π.*	*Καθηγητής Ε.Μ.Π.*

Αθήνα, Σεπτέμβριος 2024

.....
Αυγερινός Πέτρος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η έρευνα αυτή εστιάζει στο τομέα του Explainable AI και στις εφαρμογές της για την ανάπτυξη μοντέλων μηχανικής μάθησης για την ανίχνευση άνοιας μέσω ανάλυσης ομιλίας. Η έρευνα πραγματοποιήθηκε στα πλαίσια του έργου COMFORTAGE, το οποίο έχει στόχο την ανάπτυξη εργαλείων για την παρακολούθηση και διάγνωση ασθενών με άνοια. Η μελέτη μας περιλαμβάνει μια επισκόπηση των μεθόδων XAI και των μετρικών τους, επισημαίνοντας τα πλεονεκτήματα και τα μειονεκτήματα των διάφορων προσεγγίσεων. Εξετάσαμε επίσης τις ηθικές και νομικές εκτάσεις της χρήσης εφαρμογών τεχνητής νοημοσύνης στην υγεία, με γνώμονα τη διαφάνεια, την ευθύνη και τη δικαιοσύνη στην ανάπτυξη τέτοιων μοντέλων, και τον ρόλο που έχει να παίξει το XAI στον συγκεκριμένο τομέα. Η έρευνα μας κορυφώθηκε με την ανάπτυξη ενός XAI εργαλείου ονόματι DEMET, το οποίο έκανε χρήση ensemble learning συνδυάζοντας ταξινομητές και transformers για την ανίχνευση άνοιας από δείγματα ομιλίας. Το DEMET έδειξε ότι το ensemble learning δίνει σημαντικά αποτελέσματα στην απόδοση, επιτυγχάνοντας πάνω από 97% ακρίβεια στο σύνολο δεδομένων DementiaBank. Εξετάσαμε επίσης τη χρήση φωνητικών χαρακτηριστικών τα οποία υπήρχαν στα δείγματα ομιλίας του συνόλου δεδομένων DementiaBank, τα οποία μετατράπηκαν σε μια πιο ερμηνεύσιμη μορφή την οποία ονομάσαμε CHA tokens και ύστερα δημιουργήσαμε εξηγήσεις από τρεις διαφορετικές μεθόδους επεξηγησιμότητας, τις LIME, Transformers-Interpret και Anchors. Παρατηρήσαμε ότι οι LIME και Transformers-Interpret ήταν πιο αποτελεσματικές στην παροχή εύκολα ερμηνεύσιμων εξηγήσεων, ενώ η μέθοδος Anchors δεν είχε την ίδια απόδοση. Επίσης, παρατηρήσαμε ότι τα CHA tokens κατάφεραν να βελτιώσουν την ερμηνευσιμότητα του μοντέλου παρέχοντας πιο κατανοητές εξηγήσεις που συμφωνούσαν με την υπάρχουσα βιβλιογραφία σχετικά με τα συμπτώματα της άνοιας. Ένα βασικό εύρημα της εργασίας μας ήταν ότι ο συνδυασμός διαφορετικών εξηγήσεων οι οποίες δημιουργήθηκαν από διαφορετικούς transformers για την τελική δημιουργία μίας ενιαίας εξήγησης, μια μορφή ensemble learning δηλαδή αλλά στα πλαίσια της εξήγησης, μπορεί να μειώσει προκαταλήψεις των μοναδικών εξηγήσεων και να προσφέρει πιο αξιόπιστες ερμηνείες. Αυτά τα ευρήματα προήλθαν από μια κλινική μελέτη που πραγματοποιήσαμε για να αξιολογήσουμε τη χρησιμότητα και την αποτελεσματικότητα των μεθόδων επεξηγησιμότητας σε ένα πραγματικό περιβάλλον. Οι επαγγελματίες υγείας που συμμετείχαν στην κλινική μελέτη πιστεύουν ότι οι εφαρμογές που βασίζονται στο XAI όπως το DEMET μπορεί να είναι χρήσιμες στην ανίχνευση και παρακολούθηση της άνοιας, ιδιαίτερα για μη κλινικές, εμπορικές εφαρμογές. Αυτά τα εργαλεία μπορούν να παρέχουν σε ασθενείς ένα μη επεμβατικό, οικονομικό μέσο αξιολόγησης της κατάστασής τους, και τελικά να τους οδηγήσουν στην επικοινωνία με κάποιον ειδικό. Τα ευρήματα από αυτήν την έρευνα συμβάλλουν στις συνεχείς προσπάθειες του έργου COMFORTAGE για την ανάπτυξη αποτελεσματικών λύσεων για τη φροντίδα της άνοιας, με το DEMET να παίζει έναν πιθανό ρόλο στην επίτευξη αυτών των στόχων.

Λέξεις Κλειδιά: XAI, XAI Evaluation, Ensemble Learning, Dementia Detection, Speech Analysis, Ensemble XAI, CHA tokens, Transformers, LIME, Anchors, Transformers-Interpret, COMFORTAGE

Abstract

This study focuses on the field of Explainable Artificial Intelligence (XAI) and its application in the development of machine learning models for dementia detection through speech analysis. The research was conducted as part of the COMFORTAGE project, which aims to develop tools for monitoring and diagnosing dementia patients. The study involved a comprehensive review of XAI methods and their metrics, highlighting the advantages and disadvantages of various approaches. We also delved into the ethical and legal considerations surrounding the use of AI in healthcare, emphasizing the importance of transparency, accountability, and fairness in model development, and the role that XAI stands to play in this ever-evolving landscape. Our research culminated in the development of an XAI-driven cognitive assessment tool, named DEMET, which leveraged ensemble learning techniques in order to combine classifiers and transformers to detect dementia from spontaneous speech samples. DEMET demonstrated that ensemble models can significantly improve performance, achieving over 97% accuracy on the DementiaBank dataset. We explored the use of phonological features derived from the speech samples of the DementiaBank dataset, which were converted into a more interpretable format which we called CHA tokens and generated explanations from three different explainable methods, namely LIME, Transformers-Interpret and Anchors, to assess each method's performance on both qualitative and quantitive metrics. Our results showed that LIME and Transformers-Interpret were more effective in providing interpretable explanations, while Anchors did not perform as well. We also found that CHA tokens managed to enhance model interpretability by providing more understandable explanations which were aligned with existing literature on dementia symptoms. A key finding of our work was that combining explanations from different models could reduce potential biases of singular explanations and offer more reliable interpretations. These findings were derived from a clinical study we conducted to evaluate the usability and effectiveness of the explainability methods in a real-world setting, and although the clinical study involved a small sample size, feedback from healthcare professionals indicated that XAI-driven approaches like DEMET could be valuable in dementia detection and monitoring, particularly for non-clinical, commercial applications. These tools could provide individuals with a non-invasive, cost-effective method of self-assessment, potentially leading to professional consultation. The findings from this research contribute to the ongoing efforts of the COMFORTAGE project to develop effective solutions for dementia care, with DEMET playing a potential role in achieving these goals.

Keywords: XAI, XAI Evaluation, Ensemble Learning, Dementia Detection, Speech Analysis, Ensemble XAI, CHA tokens, Transformers, LIME, Anchors, Transformers-Interpret, COMFORTAGE

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή μου, Δρ. Εμμανουήλ Πρωτονοτάριο, για την ευκαιρία που μου έδωσε να εργαστώ σε αυτό το έργο και την επακόλουθη διπλωματική εργασία. Η υποστήριξή του και η πίστη στις ικανότητές μου υπήρξαν ανεκτίμητες για μένα, και είμαι ευγνώμων για την καθοδήγησή του. Επίσης, θα ήθελα να ευχαριστήσω τον ερευνητή Σωτήρη Μεσσίνη που αφιέρωσε το χρόνο και τη γνώση του για να με βοηθήσει στην εκπόνηση αυτής της εργασίας. Θέλω να ευχαριστήσω τις αδερφές μου, Δανάη και Κυριακή, που πάντα πίστευαν σε μένα και με στήριζαν σε κάθε μου εγχείρημα αλλά και τους γονείς μου για την υποστήριξή τους. Τέλος ευχαριστώ τη γάτα και συγκάτοικό μου, Pixie, που μου κρατούσε συντροφιά όλες εκείνες τις ώρες εργασίας.

Το έργο αυτό είναι αφιερωμένο στην Κική.

Contents

1 Εκτεταμένη Περίληψη στα Ελληνικά **1**
1.1 Εισαγωγή . . . 1
1.1.1 Κίνητρα για τη μελέτη . . . 2
1.1.2 Ερευνητικά ερωτήματα και στόχοι . . . 2
1.2 Θεωρητικό υπόβαθρο . . . 3
1.2.1 Μέθοδοι XAI . . . 3
1.2.2 Μετρικές XAI . . . 5
1.3 Μεθοδολογία . . . 5
1.3.1 Σχεδιασμός Έρευνας . . . 5
1.3.2 Δεδομένα . . . 6
1.3.3 Μοντέλα . . . 7
1.3.4 Μεθοδολογία Εξήγησης . . . 8
1.4 Αποτελέσματα . . . 10
1.4.1 Αποτελέσματα Μοντέλων . . . 10
1.4.2 Αποτελέσματα Εξήγησης . . . 11
1.5 Ηθικά Ζητήματα . . . 12
1.6 Συμπεράσματα . . . 13

2 Introduction **14**
2.1 Background and overview of AI in healthcare . . . 15
2.2 Importance of explainability in AI . . . 16
2.3 Motivation for the study . . . 17
2.4 Research questions and objectives . . . 18

3 Literature Review **19**
3.1 Evolution and Current State of XAI . . . 19
3.2 Explainable AI in Health Care . . . 20
3.3 Mental Health and XAI . . . 20
3.4 Existing Frameworks and models for Explainability . . . 22
3.4.1 Interpretable Models . . . 22
3.4.2 Model Agnostic Methods . . . 24
3.5 Applications of Explainable AI models in healthcare . . . 30
3.6 State-of-the-Art Models and Techniques . . . 32
3.6.1 Speech Processing . . . 32
3.6.2 Text Processing . . . 34
3.6.3 Multi-modal . . . 36

4 Evaluation of Explainable AI models **43**
4.1 Scoring the Explainability of XAI Models . . . 44

5 Methodology **47**
5.1 Research design . . . 47
5.2 Data collection methods . . . 48
5.2.1 Collection and Processing . . . 48
5.2.2 Analysis . . . 50
5.3 Model Development . . . 54
5.3.1 Selection Criteria . . . 54
5.3.2 Model Architecture . . . 56
5.4 Selection Criteria for explainable methods . . . 59
5.4.1 Selection Criteria . . . 59
5.4.2 Providing Explainability . . . 60
5.5 Ethical considerations . . . 66

6 The COMFORTAGE Project **67**
6.1 Dementia and Frailty . . . 67
6.2 Problem Objectives and Outcomes . . . 68
6.3 DEMET: Our Contribution to the COMFORTAGE Project . . . 69
6.4 Algorithms and Models: Implementation and Results . . . 71
6.4.1 Model Evaluation . . . 72
6.4.2 Explainability Evaluation . . . 75
6.5 Discussion . . . 77

7 Ethical and Legal Considerations **78**
7.1 Roadmap of ethical concerns . . . 78
7.2 Patient privacy and consent . . . 80
7.2.1 Core principles of privacy and consent . . . 81
7.2.2 Research and development under privacy regulations . . . 82
7.2.3 Privacy preserving AI techniques . . . 83
7.3 Bias and fairness in AI algorithms . . . 83
7.3.1 Core concepts of bias and fairness . . . 84
7.3.2 Sources of bias in AI algorithms . . . 84
7.3.3 Mitigating bias in AI algorithms . . . 85
7.3.4 Frameworks to reduce bias in healthcare . . . 86
7.4 Compliance with healthcare regulations . . . 88
7.4.1 AI Act . . . 88
7.4.2 Corporate Adaptation . . . 89
7.4.3 Regulatory Challenges . . . 90

8 Future Directions and Challenges **92**
8.1 Emerging trends in explainable AI for healthcare . . . 92
8.2 Anticipated challenges and potential solutions . . . 94
8.3 Recommendations for future research . . . 96

9 Conclusion **98**
9.1 Summary of key findings and contributions of the thesis . . . 98
9.1.1 CHA Tokens . . . 98
9.1.2 Ensemble Models . . . 99
9.1.3 Ensemble Explainer . . . 99
9.1.4 User-Centric Design . . . 99
9.2 Implications for the healthcare industry . . . 100
9.2.1 Enhancing trust in AI technologies . . . 100

9.2.2 Patient-centric approach 100
9.2.3 Breaking new ground in Explainability 101
9.2.4 Conclusion 101
9.3 Concluding remarks 102

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Εισαγωγή

Στο χώρο της υγειονομικής περίθαλψης, οι αναδυόμενες τεχνολογικές καινοτομίες καταγράφουν μια ταχεία μετάβαση προς μια προσέγγιση που θέτει στο επίκεντρο τον ασθενή, σε σύγκριση με προσεγγίσεις του παρελθόντος [26]. Αυτή η αλλαγή έχει ονομαστεί ως η τέταρτη επανάσταση υγειονομικής περίθαλψης ή Healthcare 4.0, η οποία προωθεί τη χρήση τεχνολογιών που βασίζονται σε μεγάλο όγκο δεδομένων, εφαρμογές blockchain, υπολογιστικά νέφη και συσκευές με αισθητήρες για τη βελτίωση των αποτελεσμάτων και των εμπειριών των ασθενών [21]. Στον πυρήνα αυτής της επανάστασης βρίσκεται η επιτάχυνση της ιατρικής καινοτομίας και έρευνας, και η ταυτόχρονη παροχή των απαραίτητων εργαλείων και πόρων για την επίτευξη της καλύτερης δυνατής φροντίδας των ασθενών. Η Healthcare 4.0 υιοθετείται πλέον από πολλούς παρόχους υγειονομικής περίθαλψης και αναμένεται να αποτελέσει το νέο πρότυπο στον κλάδο.

Η τεχνητή νοημοσύνη (AI), μαζί με άλλες τεχνολογίες, στη ραγδαία ανάπτυξή της έχει δείξει σημαντικά αποτελέσματα στον τομέα της υγειονομικής περίθαλψης, οδηγώντας την έτσι σε μια νέα και βελτιωμένη εποχή, όπου η έμφαση δίνεται πλέον στην ευημερία του ασθενή και την ποιότητα της φροντίδας που λαμβάνει, ενώ παράλληλα μειώνει τα κόστη και βελτιώνει την αποτελεσματικότητα των συστημάτων. Αυτή η εποχή ονομάζεται Healthcare 5.0 και περιλαμβάνει την ενσωμάτωση της τεχνητής νοημοσύνης με εκατομμύρια συσκευές IoT, οι οποίες θα είναι διασυνδεδεμένες και θα μπορούν να επικοινωνούν μέσω δικτύων 5G. Αυτές οι συσκευές θα συνδυαστούν με εξελιγμένους αλγόριθμους τεχνητής νοημοσύνης για να παρέχουν εξατομικευμένη περίθαλψη στους ασθενείς [139].

Ωστόσο, η τεχνητή νοημοσύνη δεν είναι αλάνθαστη, και η εκτεταμένη χρήση της στην υγειονομική περίθαλψη μπορεί να έχει σοβαρές επιπτώσεις στην ασφάλεια των ασθενών [112]. Είναι προφανές ότι, προκειμένου να επιτευχθεί η συμμόρφωση των συστημάτων αυτών με τις ηθικές αρχές και τις απαιτήσεις που θέτει το συνεχώς μεταβαλλόμενο τοπίο των τεχνολογικών καινοτομιών, είναι απαραίτητη η στροφή προς ένα σύστημα που είναι υπεύθυνο και διαφανές. Αυτή η ανάγκη έχει οδηγήσει στη δημιουργία της Υπεύθυνης Τεχνητής Νοημοσύνης (Responsible AI), η οποία είναι μια διεπιστημονική και δυναμική διαδικασία, που υπερβαίνει τις τεχνικές πτυχές της ανάπτυξης της τεχνητής νοημοσύνης και περιλαμβάνει τα ηθικά, νομικά και κοινωνικά πρότυπα που είναι απαραίτητα για διαφανή και υπεύθυνα συστήματα τεχνητής νοημοσύνης [54].

Η Επεξηγηματική Τεχνητή Νοημοσύνη (Explainable AI, XAI) είναι ένας υποτομέας της Υπεύθυνης Τεχνητής Νοημοσύνης και αφορά την ανάπτυξη συστημάτων τεχνητής νοημοσύνης ικανών να παρέχουν εξηγήσεις για τις αποφάσεις και τις ενέργειές τους. Με την πρόοδο στην τεχνητή νοημοσύνη, τα μοντέλα γίνονται ολοένα και πιο πολύπλοκα, καθιστώντας έτσι δύσκολη,

ακόμη και για τους ειδικούς, την κατανόηση της διαδικασίας λήψης αποφάσεων αυτών των μοντέλων [160]. Πιο απλά μοντέλα, όπως η γραμμική παλινδρόμηση ή τα δέντρα αποφάσεων, είναι εύκολα κατανοητά στις αποφάσεις τους, καθώς βασίζονται σε μαθηματικούς κανόνες και μπορούν να αναπαρασταθούν οπτικά. Αυτά τα μοντέλα θεωρούνται εγγενώς ερμηνεύσιμα, δηλαδή ερμηνεύσιμα από τον σχεδιασμό τους. Αντίθετα, τα μοντέλα βαθιάς μάθησης, όπως τα νευρωνικά δίκτυα, δεν τηρούν αυτήν την ιδιότητα, καθώς η κλίμακα και η πολυπλοκότητά τους καθιστούν δύσκολο να μπορέσει κανείς να κατανοήσει πως τα δεδομένα εισόδου, συμβάλλουν στη μεταβολή των παραμέτρων του μοντέλου και στην τελική του απόφαση. Το XAI αποσκοπεί στην επίλυση αυτού του ζητήματος παρέχοντας εξηγήσεις και ερμηνείες για τις αποφάσεις που λαμβάνονται από αυτά τα πιο περίπλοκα μοντέλα. Στον τομέα της υγειονομικής περίθαλψης, όπου οι αποφάσεις που λαμβάνονται μπορεί να είναι κρίσιμες για την ευημερία των ασθενών, η ανάγκη για διαφανή και υπεύθυνα εργαλεία υποστήριξης είναι ουσιαστική. Πλέον η λύση ενός "μαύρου κουτιού" δεν είναι είναι επαρκής, σε ένα χώρο όπου η εμπιστοσύνη και η διαφάνεια είναι απαραίτητες. Τα XAI συστήματα έχουν την δυνατότητα να αποσαφηνίσουν αυτά τα μοντέλα "μαύρου κουτιού" σε ερμηνεύσιμα και κατανοητά εργαλεία υποστήριξης.

1.1.1 Κίνητρα για τη μελέτη

Το κίνητρο για αυτήν τη μελέτη προέρχεται από την ανάγκη να γίνει μια ολοκληρωμένη ανασκόπηση του XAI και της εφαρμογής του στη βιομηχανία της υγειονομικής περίθαλψης, ιδιαίτερα στον τομέα της ψυχικής υγείας και των νευροεκφυλιστικών ασθενειών. Αυτή η μελέτη στοχεύει στη συλλογή και ανάλυση της τρέχουσας κατάστασης του XAI και στην παροχή μιας λεπτομερούς επισκόπησης των διαφόρων μεθόδων και τεχνικών που χρησιμοποιούνται στον τομέα. Η μελέτη θα εξετάσει επίσης τα πιθανά οφέλη και τις προκλήσεις της εφαρμογής του XAI στην ψυχική υγεία μέσα από την μελέτη σχετικών έργων στον τομέα. Επίσης θα διερευνήσει τις ηθικές και νομικές επιπτώσεις της χρήσης αυτών των συστημάτων στην κλινική. Τέλος, αυτή η έρευνα στοχεύει να συμβάλει στην ανάπτυξη συστημάτων τεχνητής νοημοσύνης που είναι διαφανή, υπεύθυνα και αξιόπιστα. Κάνοντάς το αυτό, επιδιώκει να αντιμετωπίσει το κρίσιμο χάσμα μεταξύ των δυνατοτήτων της τεχνητής νοημοσύνης και της πραγματικής εφαρμογής αυτών των συστημάτων στην υγεία.

1.1.2 Ερευνητικά ερωτήματα και στόχοι

Αυτή η διπλωματική εργασία καθοδηγείται από τα ακόλουθα ερευνητικά ερωτήματα:

1. Πώς μπορούν οι μέθοδοι XAI να εφαρμοστούν για την ανάπτυξη αξιόπιστων μοντέλων τεχνητής νοημοσύνης για την ψυχική υγεία;

2. Ποια είναι τα τρέχοντα υπερσύγχρονα μοντέλα τεχνητής νοημοσύνης στην ψυχική υγεία τα οποία χρησιμοποιούν επεξηγήσιμες μεθόδους, και πόσο αποτελεσματικά είναι στην ταξινόμηση ή πρόβλεψη διάφορων προβλημάτων ψυχικής ή νευροεκφυλιστικής υγείας;

3. Ποιοι είναι οι βασικοί δείκτες για την αξιολόγηση της επεξηγησιμότητας των μοντέλων τεχνητής νοημοσύνης, και πώς επηρεάζουν αυτοί οι δείκτες την εμπιστοσύνη και την αξιοπιστία των αποφάσεων που προέρχονται από την τεχνητή νοημοσύνη;

4. Ποιοι ηθικοί κίνδυνοι πρέπει να ληφθούν υπόψη κατά την ανάπτυξη και την εφαρμογή μοντέλων τεχνητής νοημοσύνης στην ψυχική υγεία, και πώς μπορεί το XAI να συμβάλει στην αντιμετώπιση αυτών των ηθικών προκλήσεων;

Για την αντιμετώπιση αυτών των ερευνητικών ερωτημάτων, οι στόχοι αυτής της μελέτης είναι:

1. Να γίνει ανασκόπηση του τρέχοντος τοπίου των μεθόδων XAI και των εφαρμογών τους στην ψυχική υγεία.

2. Να εντοπιστούν υπερσύγχρονα μοντέλα τεχνητής νοημοσύνης που χρησιμοποιούν επεξηγήσιμες μεθόδους για τη διάγνωση, τη θεραπεία και την παρακολούθηση καταστάσεων ψυχικής υγείας, επισημαίνοντας τα δυνατά τους σημεία, τις αδυναμίες τους και τους τομείς για βελτίωση.

3. Να αναπτυχθεί ένα πειραματικό μοντέλο τεχνητής νοημοσύνης χρησιμοποιώντας εξηγήσεις από το μοντέλο LIME για να αποδειχθεί η πρακτική εφαρμογή του XAI στην ψυχική υγεία και συγκεκριμένα στην άνοια.

4. Να διερευνηθούν οι ηθικές διαστάσεις της τεχνητής νοημοσύνης στην υγεία, με ιδιαίτερη έμφαση στον ρόλο του XAI στην ενίσχυση της διαφάνειας, της υπευθυνότητας και της εμπιστοσύνης μεταξύ παρόχων υγειονομικής περίθαλψης και ασθενών.

Μέσω αυτής της έρευνας, η μελέτη στοχεύει να συνεισφέρει πολύτιμες γνώσεις και πρακτικά εργαλεία στον τομέα της τεχνητής νοημοσύνης για την ψυχική υγειονομική περίθαλψη, με έμφαση στην σημασία της επεξηγησιμότητας, της ηθικής και του ανθρωποκεντρικού σχεδιασμού στην ανάπτυξη εφαρμογών τεχνητής νοημοσύνης.

1.2 Θεωρητικό υπόβαθρο

Αρχικά εισήχθη από τους Van Lent et al [88] το 2004, το XAI στόχευε να καταστήσει τα συστήματα τεχνητής νοημοσύνης κατανοητά από τους ανθρώπους. Τα επόμενα χρόνια, καθώς τα μοντέλα μηχανικής μάθησης (ML) και βαθιάς μάθησης (DL) προόδευαν, η έμφαση μετατοπίστηκε στην επεξηγησιμότητα αυτών των μοντέλων. Ο όρος XAI έγινε συνώνυμος με την υπεύθυνη τεχνητή νοημοσύνη, γεφυρώνοντας το χάσμα μεταξύ της φύσης "μαύρου κουτιού" των μοντέλων τεχνητής νοημοσύνης και της ανάγκης για διαφάνεια και ανθρώπινη κατανόηση. Το 2017, η DARPA [57] [1] ξεκίνησε το πρόγραμμα XAI, για την ανάπτυξη ενός συστήματος τεχνητής νοημοσύνης για καλύτερη κατανόηση της διαδικασίας λήψης αποφάσεων από τους τελικούς χρήστες. Έχουν επίσης διεξαχθεί μελέτες από αξιόπιστους οργανισμούς για το XAI και το αντίκτυπο που έχει στην κοινωνία και την επιστήμη.

Σήμερα, το XAI χρησιμοποιείται σε διάφορους τομείς και σε ποικίλες εφαρμογές και ενισχύει τα προϋπάρχοντα μοντέλα τεχνητής νοημοσύνης, σε μια εποχή διαφάνειας και υπευθυνότητας. Η χρήση του εκτείνεται απο την ανάλυση συναισθημάτων για καλύτερες προτάσεις διαφημίσεων στο ηλεκτρονικό εμπόριο, μέχρι πραγματικού χρόνου εφαρμογές σε στρατιωτικές επιχειρήσεις. Άλλες εφαρμογές περιλαμβάνουν τη χρήση του XAI στην υγειονομική περίθαλψη, στις μεταφορές, στα οικονομικά, στη δικαιοσύνη, στις επιχειρήσεις και σε πολλά άλλα.

1.2.1 Μέθοδοι XAI

Η επεξηγησιμότητα μπορεί να είναι τοπική (local), που σημαίνει ότι ενδιαφερόμαστε για μια συγκεκριμένη περίπτωση πρόβλεψης και την εξήγηση του μοντέλου για αυτή τη συγκεκριμένη περίπτωση, ή συνολική (global), που σημαίνει ότι ενδιαφερόμαστε για το πώς λειτουργεί το μοντέλο σε έναν πληθυσμό περιπτώσεων και πώς συμπεριφέρεται γενικά το μοντέλο. Η τοπική επεξηγησιμότητα μπορεί να διατυπωθεί ως το ερώτημα του γιατί το μοντέλο πήρε μια συγκεκριμένη

[1] Η DARPA είναι μια υπηρεσία που αναπτύσσει πρωτοποριακή τεχνολογία και καινοτομίες με στόχο την εθνική ασφάλεια των Ηνωμένων Πολιτειών.

απόφαση για μια δοσμένη περίπτωση, και η συνολική επεξηγησιμότητα μπορεί να διατυπωθεί ως το ερώτημα του πώς το μοντέλο θα συμπεριφερθεί γενικά όσον αφορά την επεξήγηση, όταν πρόκειται για μια σειρά περιπτώσεων. Στην παρούσα μελέτη, θα επικεντρωθούμε κυρίως στην τοπική ερμηνευσιμότητα και επεξηγησιμότητα.

Υπάρχουν δύο γενικοί τύποι μεθόδων για την επεξηγησιμότητα [110]. Αρχικά, υπάρχει η μέθοδος που είναι συγκεκριμένη για το μοντέλο (model specific), η οποία είναι κυρίως προσανατολισμένη προς μοντέλα μηχανικής μάθησης όπως τα Decision Trees, Random Forests και άλλα. Η μέθοδος αυτή αναφέρεται στην ικανότητα της μεθόδου να έχει προηγούμενη εκτεταμένη γνώση της εσωτερικής δομής και των εσωτερικών λειτουργιών του μοντέλου, όπως τις παραμέτρους του, τις συναρτήσεις ενεργοποίησης και ενδεχόμενως βελτιστοποιήσεις του μοντέλου. Η δεύτερη μέθοδος ονομάζεται ανεξάρτητη από το μοντέλο (model agnostic). Ο τρόπος που λειτουργεί αυτή η μέθοδος είναι ότι παρέχει επεξηγησιμότητα χωρίς καμία γνώση σχετικά με το πώς λειτουργεί το μοντέλο στην πραγματικότητα, αλλά αντλεί όλες τις εξηγήσεις εξετάζοντας την είσοδο και την έξοδο του μοντέλου. Αυτή η μέθοδος χρησιμοποιείται κυρίως με δίκτυα βαθιάς μάθησης, συνελικτικά νευρωνικά δίκτυα και άλλα μοντέλα όπου είναι εξαιρετικά δύσκολο να κατανοηθεί πώς υπολογίζουν τις εξόδους τους και λαμβάνουν αποφάσεις [69].

Υπάρχουν και άλλοι τρόποι με τους οποίους μπορεί να κατηγοριοποιηθεί η επεξηγησιμότητα, όπως οι ante-hoc και οι post-hoc μέθοδοι. Η ante-hoc μέθοδος αναφέρεται σε μοντέλα που είναι εγγενώς ερμηνεύσιμα, πράγμα που σημαίνει ότι ο σχεδιασμός τους ευνοεί την επεξηγησιμότητα με κάποιο τρόπο, κυρίως λόγω της χρήσης κανόνων απόφασης και οπτικοποιήσεων. Η post-hoc αναφέρεται σε μοντέλα που είναι τύπου "μαύρο κουτί", όπως τα νευρωνικά δίκτυα και οι transformers. Εδώ η επεξηγησιμότητα προκύπτει με την αξιολόγηση της εισόδου και της εξόδου αυτών των μοντέλων.

Η τεχνητή νοημοσύνη έχει σημειώσει μεγάλες επιτυχίες τα τελευταία χρόνια, αλλά η φύση των μοντέλων "μαύρου κουτιού" δεν έχει επιτρέψει τη μεγάλη ενσωμάτωσή τους σε πολλούς τομείς όπου η λήψη αποφάσεων είναι κρίσιμη για την ασφάλεια και την προστασία ανθρώπων και οργανισμών. Τα μοντέλα που σημειώνουν την υψηλότερη ακρίβεια είναι στην πραγματικότητα, στις περισσότερες περιπτώσεις, τα πιο πολύπλοκα, γεγονός που καθιστά τους επιστήμονες και τους επαγγελματίες επιφυλακτικούς στη χρήση τους. Οι κανονισμοί που θέτονται από τα κράτη και τους κυβερνητικούς οργανισμούς επίσης αποτελούν εμπόδια για τη χρήση τέτοιου τύπου μοντέλων. Οι επεξηγηματικές μέθοδοι post-hoc στοχεύουν να προσδώσουν εμπιστοσύνη σε αυτά τα μοντέλα και να διευκολύνουν την ενσωμάτωσή τους σε περιβάλλοντα που απαιτούν αυτή τη διαφάνεια.

Ante-Hoc

Τα εγγενώς ερμηνεύσιμα μοντέλα παρέχουν, από τον σχεδιασμό τους, ρητούς κανόνες και διαισθητικές αναπαραστάσεις ώστε οι χρήστες να μπορούν να κατανοήσουν τη διαδικασία λήψης αποφάσεων τους. Θεωρούνται ο ευκολότερος και ταχύτερος τρόπος για την παραγωγή εξηγήσεων, αλλά περιορίζονται σε εκείνο το μοντέλο για το οποίο έχουν σχεδιαστεί.

Post-Hoc

Οι μέθοδοι post-hoc χρησιμοποιούνται για να εξηγήσουν τις προβλέψεις πιο πολύπλοκων μοντέλων. Χρησιμοποιούνται για να παράγουν εξηγήσεις χωρίς καμία γνώση της εσωτερικής δομής του μοντέλου. Είναι πιο ευέλικτες και προσαρμόσιμες, καθώς οι ερμηνείες μπορούν να λειτουργήσουν με οποιοδήποτε μοντέλο μηχανικής μάθησης. Είναι επίσης πολύ ευέλικτες ως προς τη μορφή της εξήγησης, με μορφές όπως κανόνες, δέντρα αποφάσεων, παραδείγματα και άλλα. Παραδείγματα τέτοιων μεθόδων είναι το LIME, το SHAP, το Anchors, το SHAPley Additive Explanations και άλλα.

1.2.2 Μετρικές XAI

Όταν ασχολούμαστε με συστήματα XAI που παράγουν εξηγήσεις για να βοηθήσουν τη λήψη αποφάσεων για εργασίες σε ένα επαγγελματικό περιβάλλον, τίθεται γρήγορα το ερώτημα σχετικά με την εγκυρότητα των εξηγήσεων που παρέχονται από τα συστήματα αυτά [66]. Τι κάνει λοιπόν μια εξήγηση έγκυρη; Το μεγαλύτερο μέρος της έρευνας που έχει γίνει στο θέμα της επεξηγησιμότητας έχει επικεντρωθεί έντονα στην ανάπτυξη νέων μεθόδων και τεχνικών για την παραγωγή εξηγήσεων, ενώ προσπαθεί να αυξήσει την απόδοση της διαδικασίας. Υπάρχει ένα θεμελιώδες κενό στη βιβλιογραφία όταν πρόκειται για τη σύνδεση της επεξηγησιμότητας με συγκεκριμένες περιπτώσεις χρήσης, και λίγη έρευνα έχει γίνει για να αξιολογηθεί η ποιότητα των εξηγήσεων που παράγονται από τα XAI. Συνεπώς, προκύπτει η ανάγκη για την ποσοτικοποίηση της ποιότητας των εξηγήσεων του XAI, προκειμένου να είναι δυνατή η αξιολόγηση της απόδοσης μιας μεθόδου και η σύγκριση ή αντιπαράθεση διαφορετικών μεθόδων πάνω σε συγκεκριμένες εργασίες.

Ένα ζήτημα στο πλαίσιο της αξιολόγησης είναι ότι οι εξηγήσεις που παρέχονται από αυτές τις μεθόδους είναι συχνά εξαιρετικά υποκειμενικές και εξαρτώνται από το πλαίσιο πάνω στο οποίο ενεργούν, καθιστώντας δύσκολη την αξιολόγηση της ποιότητάς τους, δεδομένης της δυσκολίας ορισμού του τι πραγματικά είναι τελικά μια καλή εξήγηση. Δεν υπάρχει ακόμη κάποιο καθολικά αποδεκτό μέτρο ποιότητας εξηγήσεων από τους ερευνητές στον τομέα. Η ειδική φύση των εξηγήσεων για κάθε τομέα χωριστά καθιστά δύσκολη τη γενίκευση της διαδικασίας αξιολόγησης σε διαφορετικές εφαρμογές, καθώς η ποιότητα μιας εξήγησης εξαρτάται σε μεγάλο βαθμό από το πλαίσιο στο οποίο χρησιμοποιείται. Σύμφωνα με το Alan Turing Institute [72], οι εξηγήσεις τοποθετούνται σε 6 κατηγορίες, τις rationale, responsibility, data, fairness, safety και impact εξηγήσεις. Κάθε μια από αυτές τις κατηγορίες αναφέρεται σε διαφορετικό τμήμα της εξήγησης και της αξιολόγησης των μοντέλων XAI.

Οι μέθοδοι για να αξιολογήσει κανείς την ποιότητα των εξηγήσεων που παράγονται από τα XAI μπορούν να χωριστούν σε τρεις κατηγορίες, συγκεκριμένα τις application-grounded μεθόδους, που αναφέρονται στην αξιολόγηση των εξηγήσεων μέσω έρευνας σε ανθρώπους, πάνω σε συγκεκριμένες εφαρμογές, τις human-grounded μεθόδους, που αναφέρονται στην αξιολόγηση των εξηγήσεων και πως αυτές εξυπηρετούν στην κατάκτηση απλών στόχων από ανθρώπους χωρίς τεχνικές γνώσεις και τις functionally-grounded μεθόδους, που αναφέρονται στην αξιολόγηση των εξηγήσεων μέσα από σαφείς ορισμούς των στόχων που έχουν οριστεί για να είναι μια εξήγηση καλή. Οι κυρίαρχες κατηγορίες των μετρικών για την αξιολόγηση εξηγήσεων είναι αυτές που μετρούν την ποιότητα (qualitative) και αυτές που μετρούν αντικειμενικές πληροφορίες σχετικά με την εξήγηση (quantitative). Οι μετρικές ποιότητας είναι πιο υποκειμενικές και αξιολογούν την ποιότητα της εξήγησης από την άποψη του χρήστη, ενώ οι μετρικές ποσότητας είναι πιο αντικειμενικές και αξιολογούν την ποιότητα της εξήγησης μέσω συγκεκριμένων μετρικών.

1.3 Μεθοδολογία

1.3.1 Σχεδιασμός Έρευνας

Ο σχεδιασμός της έρευνας που χρησιμοποιησάμε στη παρούσα μελέτη είναι μια υβριδική προσέγγιση μεταξύ της ποιοτικής και της ποσοτικής έρευνας. Το πρώτο κομμάτι της μελέτης αφορά την ανάλυση της βιβλιογραφίας και την καταγραφή των πιο σύγχρονων μεθόδων και τεχνικών XAI, καθώς επίσης και τις μετρικές και εφαρμογές τους μέσα από συστηματική έρευνα. Το δεύτερο κομμάτι αφορά την ανάλυση και επεξεργασία δεδομένων ομιλίας από ασθενείς με άνοια για την εξαγωγή χαρακτηριστικών ομιλίας και την εκπαίδευση μοντέλων. Τα εξαγώμενα χαρακτηριστικά μετατράπηκαν σε μια πιο κατανοητή μορφή ονόματι CHA tokens τα οποία χρησιμοποιήθηκαν για την

παραγωγή πιο ερμηνεύσιμων εξηγήσεων. Η τρίτη φάση ήταν η εκπαίδευση transformer μοντέλων και η αρχική παραγωγή εξηγήσεων με τρεις διαφορετικές μεθόδους, το LIME, το Transformers-Interpret και το Anchors. Οι παραγώμενες εξηγήσεις δόθηκαν σε επαγγελματίες ιατρούς για να αξιολογηθούν ως προς διαφορετικές μετρικές ποιότητας, με γνώμονα τη χρήση των μεθόδων αυτών σε πραγματικά κλινικά περιβάλλοντα. Τέλος, η τελευταία φάση αφορά την δημιουργία μιας εφαρμογής XAI που θα επιτρέπει την εύκολη χρήση των μεθόδων που αναπτύχθηκαν στην παρούσα μελέτη από ειδικούς και μη ειδικούς χρήστες.

1.3.2 Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν για την έρευνα αυτή προέρχονται από το σύνολο δεδομένων DementiaBank [87] [53], το οποίο περιλαμβάνει γραπτές ηχογραφήσεις συνεδριών κλινικής αξιολόγησης ασθενών με άνοια. Λειτουργήσαμε με το σύνολο δεδομένων σύμφωνα με τον οδηγό CHAT που προσφέρει το DementiaBank. Έγινε επεξεργασία πάνω σε αυτά τα δεδομένα ώστε να εξαχθούν χαρακτηριστικά ομιλίας, τα οποία μετατράπηκαν σε CHA tokens. Μετά έγινε ένας έλεγχος για το ποιο μέγεθος κειμένου ήταν το βέλτιστο για την εκπαίδευση και αξιολόγηση των μοντέλων μεταξύ τριών διαφορετικών μεγεθών, 5, 20 και 50 λέξεων ανά κείμενο. Ονομάσαμε αυτά τα μεγέθη 'μικρα', 'μεσαία' και 'μεγάλα' αντίστοιχα και βέλτιστο μέγεθος βρέθηκε το 'μεσαίο' καθώς συνδυάζει καλή διαχείριση του μηχανισμού attention των transformers και ένα ικανοποιητικό τελικό μέγεθος dataset για την εκπαίδευση και επαλήθευση των μοντέλων. Παρακάτω φαίνεται ο πίνακας μετατροπής των CHAT συμβόλων σε CHA tokens για κάποια από τα συχνότερα σύμβολα που χρησιμοποιούνται στο DementiaBank.

CHAT Symbol	**CHA Token**
`[/]`	[CHA REPETITION]
`[//]`	[CHA RETRACING]
`(.)`	[CHA SHORT PAUSE]
`(..)`	[CHA MEDIUM PAUSE]
`(...)`	[CHA LONG PAUSE]
`+...`	[CHA TRAILING OFF]
`&+`	[CHA PHONOLOGICAL FRAGMENT]
`&*`	[CHA INTERPOSED WORD]
`&-`	[CHA FILLER]
`text(text)text`	[CHA NON COMPLETION OF WORD]
`+..?`	[CHA TRAILING OFF QUESTION]
`+/.`	[CHA INTERRUPTION]
`+/?`	[CHA INTERRUPTION OF QUESTION]
`+//.`	[CHA SELF-INTERRUPTION]
`+//?`	[CHA SELF-INTERRUPTED QUESTION]

Table 1.1: Μετατροπή CHAT σε CHA tokens

Μετά την μετατροπή και την επεξεργασία των δεδομένων, έγινε ανάλυση ώστε να εξαχθούν διαφορετικά συμπεράσματα για τα δεδομένα και τα χαρακτηριστικά ομιλίας που χρησιμοποιήθηκαν. Συγκεκριμένα είδαμε ότι υπήρχαν συγκεκριμένα χαρακτηριστικά ομιλίας που εμφανίζονταν συχνά σε ασθενείς με άνοια, το οποίο συμφωνεί με την βιβλιογραφία που υπάρχει για το θέμα. Πολλά από αυτά τα χαρακτηριστικά ομιλίας εμφανίζονταν συχνά ως ζευγάρια στο ίδιο κείμενο, το οποίο ενισχύει την δυναμική τους ως ενδείξεις για την παρουσία άνοιας. Από τα παραπάνω αντιλαμβανόμαστε πόσο σημαντικά μπορεί να είναι τα CHA tokens για την εξαγωγή ερμηνεύσιμων εξηγήσεων από τα μοντέλα XAI και πάνω σε αυτό το φαινόμενο ακριβώς θα στηριχτούμε για την επεξηγημότητα των μοντέλων μας.

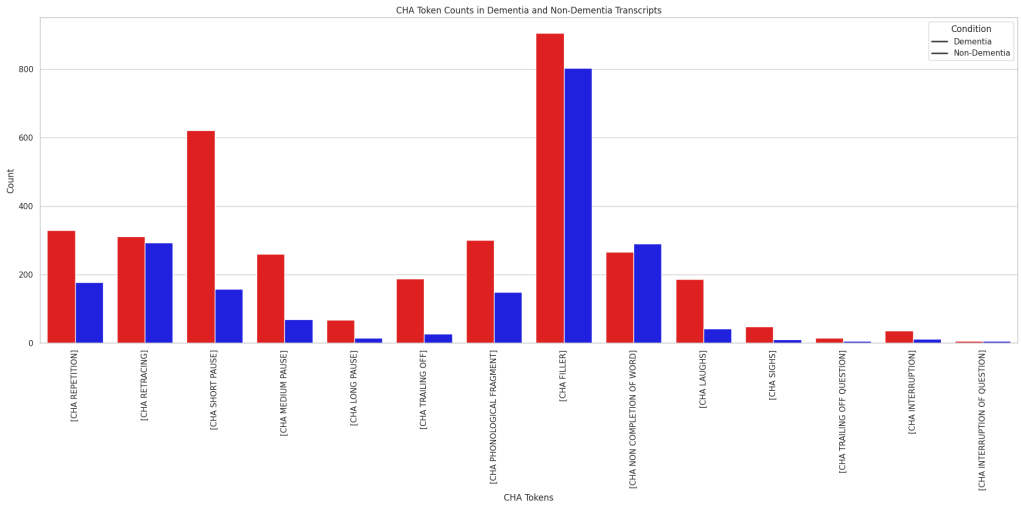


Figure 1.1: Κατανομή των CHA tokens στο σύνολο των δεδομένων

Παρακάτω φαίνεται ο πίνακας των CHA tokens ως προς την συχνότητα εμφάνισης ζευγαριών στο ίδιο κείμενο σε μορφή heatmap. Είναι φανερό πως κάποια tokens τα οποία αποτελούν ισχυρή ένδειξη εμφάνισης άνοιας εμφανίζονται συχνά μαζί στο ίδιο κείμενο όπως για παράδειγμα τα [CHA SHORT PAUSE] και [CHA LONG PAUSE] ή τα [CHA TRAILING OFF]. Είναι σημαντικό να τονίσουμε ότι επιβάλλεται εκτενέστερη ανάλυση για να δούμε πως τα CHA tokens μπορεί να επηρεάσουν την πρόβλεψη ενός transformer μοντέλου για την επιλογή μιας κλάσης μεταξύ ασθενών με άνοια και χωρίς.

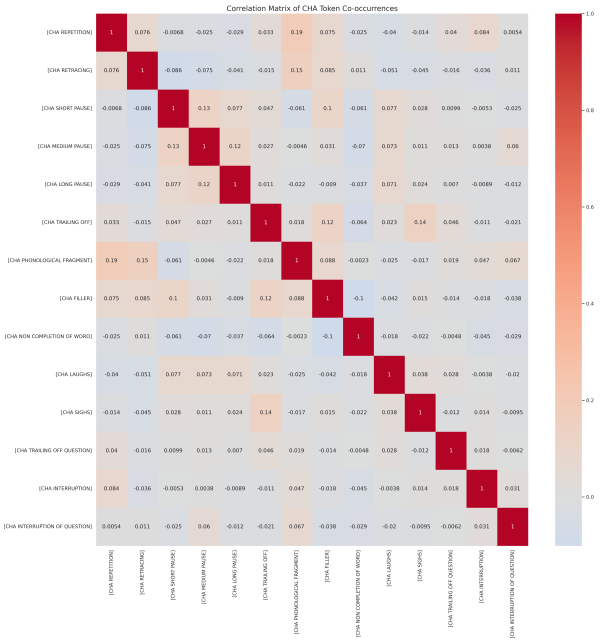


Figure 1.2: Κατανομή των CHA tokens ως προς τη συχνότητα εμφάνισης ζευγαριών στο ίδιο κείμενο

1.3.3 Μοντέλα

Για την εφαρμογή των μεθόδων XAI στα δεδομένα μας, επιλέξαμε να χρησιμοποιήσουμε μοντέλα transformer, τα οποία έχουν αποδειχθεί ότι είναι εξαιρετικά αποτελεσματικά για εργασίες επεξεργασίας φυσικής γλώσσας. Οι λόγοι για τους οποίους τα επιλέξαμε είναι ο μηχανισμός attention

που διαθέτουν, ο οποίος επιτρέπει στα μοντέλα αυτά να εστιάζουν σε συγκεκριμένα τμήματα του κειμένου και στον τρόπο που αυτά αλληλεπιδρούν μεταξύ τους για την παραγωγή της πρόβλεψης, το transfer learning που προσφέρουν, το οποίο εξυπηρετεί στην εκπαίδευση μοντέλων για πολύπλοκες εργασίες με χρήση μικρής σχετικά ποσότητας δεδομένων και τέλος την δυνατότητα παραγωγής εξηγήσεων μέσω της χρήσης των σχέσεων που προσδίδει ο μηχανισμός attention στα δεδομένα. Επιλέξαμε να χρησιμοποιήσουμε αρχικά τρεις transformers με διαφορετική αρχιτεκτονική, το BERT, το RoBERTa και το DistilBERT, να τους εκπαιδεύσουμε στα δεδομένα και ύστερα να εκπαιδεύσουμε μια σειρά από classifiers πάνω στα embeddings που παρήγαγαν τα μοντέλα αυτά. Επαναλάβαμε ύστερα το πείραμα για ακόμη δύο transformers, τους ClinicalBERT και BioBERT για να δούμε αν μπορούμε να βελτιώσουμε την απόδοση των classifiers και είδαμε πως με αυξημένη πολυπλοκότητα των transformers, αυξάνεται και η απόδοση τους. Τέλος, επιχειρήσαμε να χρησιμοποιήσουμε τις εξόδους όλων των classifiers για μια τελική πρόβλεψη μέσω της μεθόδου του ensemble learning, όπου είδαμε ότι η απόδοση των μοντέλων μας δεν βελτιώθηκε τελικά, πιθανότατα λόγω του μικρού αριθμού των transformers που χρησιμοποιήσαμε και το γεγονός ότι κάθε classifier είχε τους ίδιους transformers στην βάση του. Παρακάτω φαίνεται η συνολική αρχιτεκτονική που χρησιμοποιήσαμε για την εκπαίδευση των μοντέλων μας.

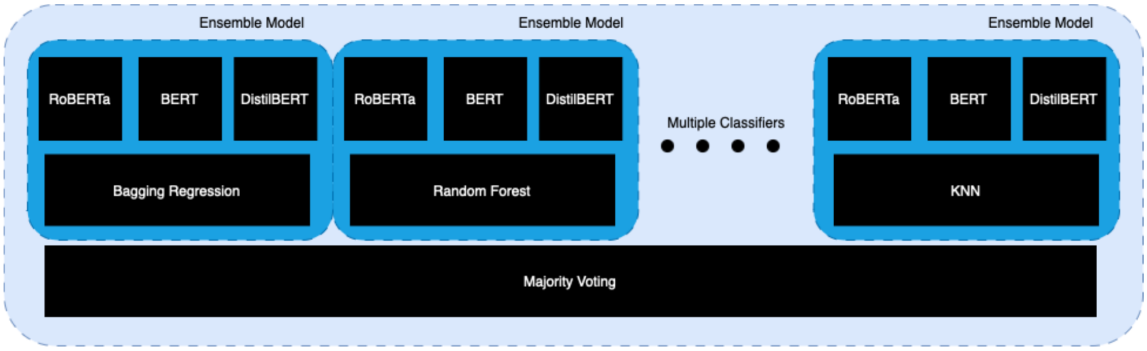


Figure 1.3: Συνολική αρχιτεκτονική μοντέλων

1.3.4 Μεθοδολογία Εξήγησης

Μετά την εκπαίδευση των αρχικών transformer μοντέλων, εφαρμόσαμε τρεις διαφορετικές μεθόδους εξήγησης, το LIME, το Transformers-Interpret και το Anchors. Τα αποτελέσματα των εξηγήσεων δόθηκαν σε επαγγελματίες ιατρούς για να αξιολογηθούν ως προς την ποιότητα τους. Μετά την ανάλυση των αποτελεσμάτων, η οποία λόγω μικρής δειγματοληψίας δεν ήταν στατιστικά σημαντική, επιχειρήσαμε να χρησιμοποιήσουμε την κυρίαρχη μέθοδο εξήγησης, το LIME, για να εξηγήσουμε τις προβλέψεις των μοντέλων μας. Το LIME επιλέχθηκε λόγω της απλότητας των εξηγήσεων που παρήγε και της σημασίας που έδωσε στα CHA tokens. Το Transformers-Interpret είχε παρόμοια απόδοση με το LIME αλλά δεν ήταν τόσο αρεστό στους επαγγελματίες ιατρούς, ενώ το Anchors δεν είχε καλή απόδοση καθώς δεν κατάφερε να παράγει εξηγήσεις οι οποίες να είναι εύκολα ερμηνεύσιμες από τους ειδικούς. Με την δημιουργία των εξηγήσεων είδαμε πως κάθε μοντέλο εστίαζε σε διαφορετικά τμήματα του κειμένου για να παράγει την πρόβλεψη του. Αυτό μας οδήγησε στο συμπέρασμα ότι μπορούμε να δημιουργήσουμε συνολικές εξηγήσεις οι οποίες χρησιμοποιούν τις εξηγήσεις από διαφορετικά μοντέλα για την δημιουργία μιας εξήγησης η οποία δεν είναι προκατειλημμένη από το μοντέλο που την παράγει για συγκεκριμμένες λέξεις ή tokens. Η εξήγηση χρησιμοποιεί weighted average των εξηγήσεων από τα τρία μοντέλα που εκπαιδεύσαμε με γνώμονα τις μετρικές των transformers. Για την αξιολόγηση των εξηγήσεων χρησιμοποιήσαμε διάφορες qualitative και quantitative μετρικές, όπως το fidelity, ο χρόνος που χρειάστηκε για την παραγωγή της εξήγησης, η σημασία που δόθηκε στα CHA tokens και η ευκολία που είχαν οι επαγγελματίες ιατροί στην ερμηνεία των εξηγήσεων.

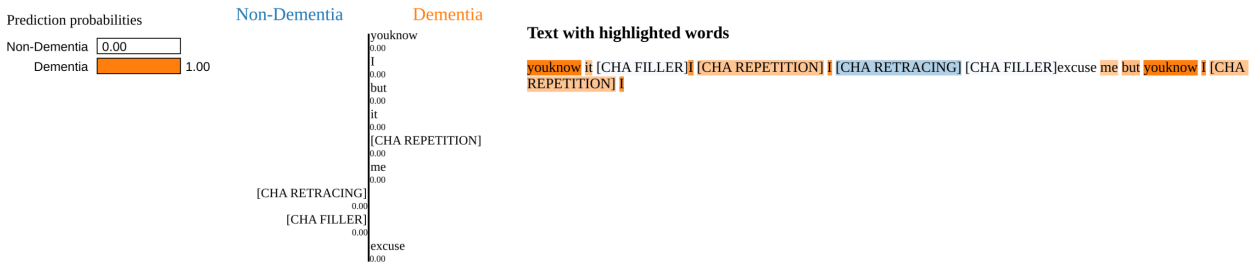


Figure 1.4: BERT LIME

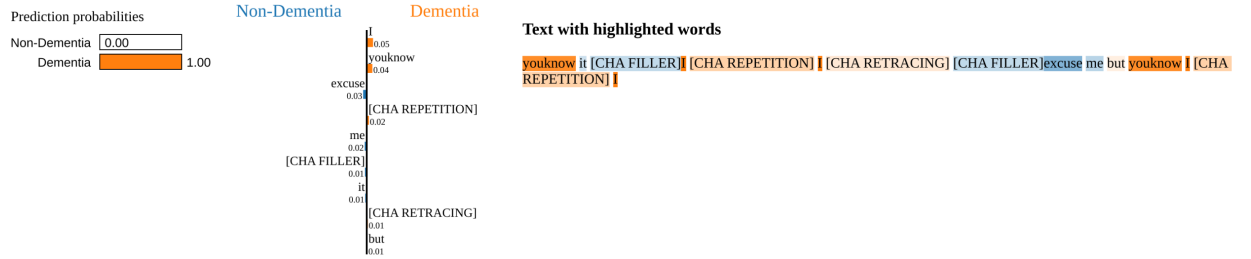


Figure 1.5: RoBERTa LIME

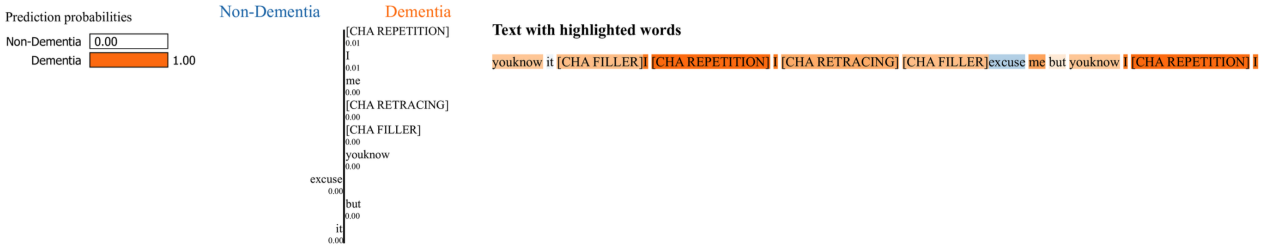


Figure 1.6: DistilBERT LIME

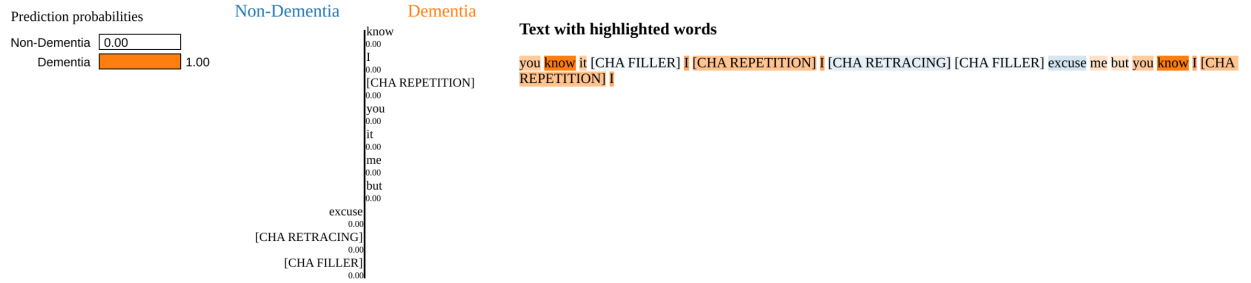


Figure 1.7: Συνολική εξήγηση

1.4 Αποτελέσματα

Η μελέτη αυτή έγινε στα πλαίσια ενός ευρωπαϊκού προγράμματος, του COMFORTAGE, που ασχολείται με την ανάπτυξη εργαλείων για την παρακολούθηση και διάγνωση ασθενών με άνοια. Στόχος της μελέτης ήταν να εξετάσει την αποτελεσματικότητα των μεθόδων XAI στην εξήγηση προβλέψεων και η δημιουργία ενός πειραματικού περιβάλλοντος για την εξήγηση των προβλέψεων των μοντέλων μηχανικής μάθησης για την διάγνωση ασθενών με άνοια το οποίο να είναι εύχρηστο και αξιόπιστο. Η συνεισφορά μας στο COMFORTAGE είναι το DEMET, μια εφαρμογή XAI που επιτρέπει την εύκολη χρήση των μεθόδων που αναπτύξαμε στην παρούσα μελέτη από ειδικούς και μη ειδικούς χρήστες μέσω μιας γραφικής διεπαφής. Το DEMET χρησιμοποιεί τα ensemble μοντέλα που εκπαιδεύσαμε και τις εξηγήσεις που παράγαμε για την παραγωγή μιας συνολικής εξήγησης και πρόβλεψης για το αν ένας ασθενής πάσχει πιθανότατα από άνοια μέσω ομιλίας. Το DEMET βρίσκεται σε πρώιμο στάδιο ανάπτυξης και αποτελεί μια προσπάθεια για την ενσωμάτωση των μεθόδων XAI σε κλινικά περιβάλλοντα.

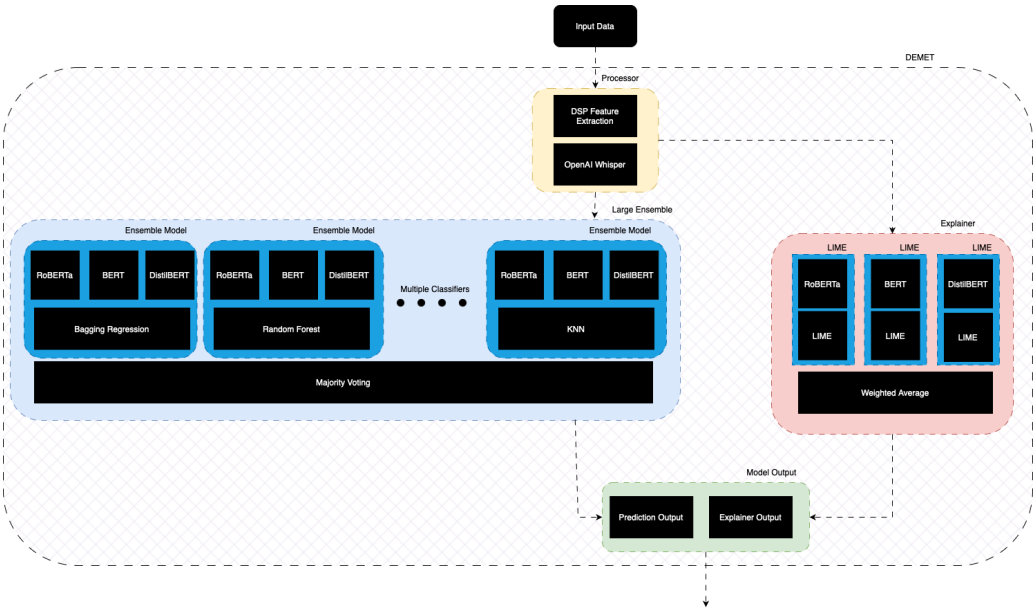


Figure 1.8: DEMET Αρχιτεκτονική

1.4.1 Αποτελέσματα Μοντέλων

Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση τόσο των transformers όσο και των classifiers που εκπαιδεύσαμε είναι οι εξής: accuracy, precision, recall και F1-score. Για την επιλογή αρχικά του μεγέθους των δεδομένων τρέξαμε το πείραμα με τα τρία διαφορετικά μεγέθη, 5, 20 και 50 λέξεων ανά κείμενο και είδαμε ότι το μεσαίο μέγεθος είχε την καλύτερη απόδοση. Το μεγάλο μέγεθος είχε καλύτερα αποτελέσματα φαινομενικά αλλά οδηγούσε το dataset σε ένα πολύ μικρό μέγεθος, ενώ το μικρό μέγεθος είχε την χειρότερη απόδοση. Για την εκπαίδεση των μοντέλων αρχικά εκπαιδεύσαμε τρεις transformers, το BERT, το RoBERTa και το DistilBERT, και ύστερα επεκτείναμε την μελέτη μας στα ClinicalBERT και BioBERT. Τα αποτελέσματα των μοντέλων μας φαίνονται στο παρακάτω πίνακα.

Model	Accuracy	Precision	Recall	F1-score
BERT	0.85	0.87	0.86	0.86
RoBERTa	0.86	0.90	0.81	0.85
DistilBERT	0.83	0.80	0.86	0.83
ClinicalBERT	0.81	0.91	0.71	0.79
BioBERT	0.82	0.85	0.81	0.83

Table 1.2: Αποτελέσματα μοντέλων

Ύστερα χρησιμοποίησαμε τις εξόδους των transformers ως εισόδους σε classifiers για την παραγωγή των τελικών προβλέψεων. Επιχειρήσαμε το πείραμα δύο φορές, την πρώτη με τρεις transformers και την δεύτερη με πέντε για να δούμε αν μπορούμε να βελτιώσουμε την απόδοση των μοντέλων μας με ενισχυμένη πολυπλοκότητα. Τα αποτελέσματα για τα ensemble μοντέλα με τρεις transformers φαίνονται στον παρακάτω πίνακα.

Classifier	Accuracy	Precision	Recall	F1-score
Bagging Regressor	0.9616	0.9433	0.9609	0.9792
Random Forest	0.9666	0.9533	0.9662	0.9794
Gradient Boosting	0.9683	0.9533	0.9678	0.9828
Support Vector Machine	0.9616	0.9433	0.9609	0.9792
Logistic Regression	0.9600	0.9533	0.9597	0.9662
K-Nearest Neighbors	0.9633	0.9500	0.9628	0.9760
Decision Tree	0.9533	0.9366	0.9493	0.9623
Majority Voting	0.9666	0.9533	0.9662	0.9794

Table 1.3: Αποτελέσματα classifiers με τρεις transformers

Τα ensemble μοντέλα με πέντε transformers είχαν καλύτερα αποτελέσματα, το οποίο μας οδήγησε στο συμπέρασμα ότι απαιτείται η εύρεση του βέλτιστου αριθμού transformers και classifiers για την επίτευξη της μέγιστης απόδοσης. Λίγοι transformers μπορεί να αποδόσουν χαμηλά αποτελέσματα, και πολλοί transformers μπορούν να εισάγουν θόρυβο, επομένως υπάρχει χώρος σε αυτό το σημείο για βελτιστοποίηση.

Classifier	Accuracy	Precision	Recall	F1-score
Bagging Regressor	0.9683	0.9533	0.9678	0.9828
Random Forest	0.9766	0.9766	0.9766	0.9766
Gradient Boosting	0.9683	0.96	0.9681	0.9763
Support Vector Machine	0.9716	0.9633	0.9714	0.9796
K-Nearest Neighbors	0.9783	0.9733	0.9782	0.9831
Logistic Regression	0.9716	0.9633	0.9714	0.9796
Decision Tree Classifier	0.9633	0.9633	0.9633	0.9633
Majority Voting	0.9733	0.9666	0.9731	0.9797

Table 1.4: Αποτελέσματα classifiers με πέντε transformers

1.4.2 Αποτελέσματα Εξήγησης

Για την αξιολόγηση των εξηγήσεων χρησιμοποίησαμε τις εξής μετρικές, απλότητα στην εξήγηση, ικανοποίηση του χρήστη ως προς την εξήγηση, οπτική ποιότητα και ικανοποίηση του χρήστη, συμπεριφορά εξήγησης σε σχέση με τα CHA tokens, ευκολία χρήσης, χρόνος παραγωγής εξήγησης

και πόσο καλά η εξήγηση κατάφερε να προσεγγίση το μοντέλο. Τα αποτελέσματα των μεθόδων εξήγησης φαίνονται στον παρακάτω πίνακα.

Type	Metric	LIME	T-I	Anchor	**DEMET**
Qlt	Simplicity	4	3	1	4
	Human Evaluation	4	3	1	4
	Visual Appeal	4	3	1	4
	CHA Token Importance	4	4	3	4
	Ease of Use	1	1	1	5
Qnt	Fidelity	0.73	1	0.9	1
	Time of Generation (GPU + 83 GB RAM)	55.47s	1.05s	0.11s	167.43s
	Time of Generation (CPU + 16 GB RAM)	**FAILED**	59.19s	9.43s	**FAILED**

Table 1.5: Αποτελέσματα μεθόδων εξήγησης

Τα αποτελέσματα έδειξαν ότι η μέθοδος LIME ήταν η πιο αρεστή στους επαγγελματίες ιατρούς. Η μέθοδος Anchors δεν κατάφερε να παράγει εξηγήσεις οι οποίες να είναι εύκολα ερμηνεύσιμες από τους ειδικούς, ενώ το Transformers-Interpret είχε παρόμοια απόδοση με το LIME αλλά δεν ήταν τόσο αρεστό όσο το LIME. Λόγω της υπολογιστικής πολυπλοκότητας της μεθόδου LIME, προτείνουμε, στην περίπτωση όπου η εξήγηση των προβλέψεων είναι απαραίτητο να γίνει σε σύντομο χρονικό διάστημα και με μικρότερη υπολογιστική πολυπλοκότητα, να γίνει χρήση της μεθόδου Transformers-Interpret. Το DEMET, όντας βασισμένο στην μέθοδο LIME, είχε αντίστοιχη απόδοση σε όλες τις μετρικές με το LIME, πέρα από την ευκολία χρήσης στην οποία ήταν η βέλτιστη επιλογή.

1.5 Ηθικά Ζητήματα

Ένα από τα κυριότερα ηθικά ζητήματα στην έρευνά μας είναι η ιδιωτικότητα και η εμπιστευτικότητα των δεδομένων που χρησιμοποιούνται. Τα δεδομένα που παρέχονται από το DementiaBank είναι εξαιρετικά ευαίσθητα και περιέχουν προσωπικές πληροφορίες για άτομα και ασθενείς. Η διαχείριση τέτοιων δεδομένων απαιτεί αυστηρή συμμόρφωση με τους νόμους και τους κανονισμούς προστασίας δεδομένων. Οι ερευνητές πρέπει να διασφαλίζουν ότι αυτά τα δεδομένα είναι ασφαλισμένα και ανώνυμα, ώστε να αποτρέπεται οποιαδήποτε πιθανή παραβίαση της ιδιωτικότητας των ατόμων αυτών. Ένα άλλο σημαντικό ηθικό ζήτημα είναι η προκατάληψη που εισάγεται πιθανά στα δεδομένα και στα μοντέλα και η διατήρηση της δικαιοσύνη στις αποφάσεις τους. Η εκπαίδευση των μοντέλων τεχνητής νοημοσύνης σε δεδομένα που είναι προκατειλημμένα για μια συγκεκριμένη μερίδα ανθρώπων μπορεί να οδηγήσει σε προκατειλημμένα και άδικα αποτελέσματα για τους ασθενείς. Ένα παράδειγμα προκατάληψης είδαμε και στις εξηγήσεις της μεθόδου LIME, όπου ορισμένες λέξεις και σύμβολα υπογραμμίζονταν ως ισχυρές ενδείξεις άνοιας, ακόμα κι αν δεν συνδέονταν διαισθητικά με τη νόσο. Τέτοιες προκαταλήψεις μπορούν να εισαχθούν κατά τη διάρκεια της εκπαίδευσης των μοντέλων, λόγω του τρόπου με τον οποίο συλλέγονται και επεξεργάζονται τα δεδομένα, ή ακόμα και κατά τη διαδικασία μεταγραφής της ομιλίας των ασθενών. Γνωρίζουμε ότι τα πρότυπα ομιλίας μπορούν να διαφέρουν δραστικά ανάμεσα σε διαφορετικές δημογραφικές ομάδες, και τα μοντέλα πρέπει να αξιολογούνται ώστε να διασφαλίζεται ότι δεν επηρεάζουν δυσανάλογα κάποια συγκεκριμένη ομάδα. Η διαφάνεια και η λογοδοσία σε αποφάσεις υψηλής επικινδυνότητας και σημασίας είναι επίσης μια πτυχή που πρέπει να ληφθεί υπόψιν κατά την ανάπτυξη μοντέλων τεχνητής νοημοσύνης για εφαρμογές υγειονομικής περίθαλψης. Η έλλειψη διαφάνειας μπορεί να οδηγήσει σε κακή χρήση ή υπερβολική εξάρτηση από τα συστήματα αυτά, φαινόμενο το οποίο ενδέχεται να έχει επιβλαβείς συνέπειες για τους ασθενείς και οργανισμούς που

τα χρησιμοποιούν. Απαιτούνται αυστηρές κατευθυντήριες γραμμές και κανονισμοί για να διασφαλιστεί το ποιος οφείλει να λογοδοτεί για τις αποφάσεις που λαμβάνονται από τα μοντέλα αυτά όταν εκείνα χρησιμοποιούνται για διαγνωστικούς σκοπούς. Είναι επίσης ουσιώδες να αναφέρουμε ότι αυτά τα εργαλεία χρησιμοποιούνται ως συμπληρωματικά εργαλεία στην κλινική, και όχι ως υποκατάστατα της ανθρώπινης κρίσης και εξειδίκευσης. Τέλος, πρέπει να εξεταστεί η ευρύτερη ηθική χρήση της τεχνητής νοημοσύνης στην υγειονομική περίθαλψη. Τα εργαλεία τεχνητής νοημοσύνης πρέπει να χρησιμοποιούνται για να βελτιώνουν την ποιότητα ζωής και φροντίδας των ασθενών, και να εφαρμόζονται με συνεπή και ηθικό τρόπο.

1.6 Συμπεράσματα

Στα πλαίσια της παρούσας μελέτης καταφέραμε να αναλύσουμε εκτενώς το πεδίο του XAI και των μεθόδων εξήγησης προβλέψεων των μοντέλων μηχανικής μάθησης. Μιλήσαμε για τις μεθόδους αυτές και τις μετρικές τους, καθώς επίσης και για τα πλεονεκτήματα και τα μειονεκτήματά τους. Τα αποτελέσματα της δουλειάς μας στο DEMET έδειξαν πως η χρήση ensemble μοντέλων μπορεί να καταφέρει σημαντικές βελτιώσεις από τα απλά transformer μοντέλα, με αποτελέσματα που ξεπερνούν το 97% στο σύνολο δεδομένων του DementiaBank. Είδαμε πως η χρήση φωνολογικών χαρακτηριστικών ως CHA tokens μπορεί να βελτιώσει την επεξηγησιμότητα των μοντέλων μας προσφέροντας μια πιο ερμηνεύσιμη εξήγηση η οποία στηρίζεται στη βιβλιογραφία σχετικά με τα συμπτώματα της άνοιας. Σημαντικό εύρημα ήταν ο συνδυασμός διαφορετικών εξηγήσεων μιας ίδιας μεθόδου εξήγησης, για την δημιουργία μιας νέας, η οποία δεν πλήττεται από πιθανές προκαταλήψεις των επιμέρους εξηγήσεων και παρέχει μια πιο αξιόπιστη εξήγηση. Οι έρευνα που πραγματοποιήσαμε με επαγγελματίες ιατρούς, παρά το μικρό αριθμό των συμμετεχόντων, έδειξε πως η συγκεκριμένη κατεύθυνση μπορεί να επιφέρει σημαντικά αποτελέσματα στον τομέα της ανίχνευσης και παρακολούθησης της άνοιας, όχι τόσο για κλινική χρήση, αλλά για εμπορικές εφαρμογές και υπηρεσίες, παρέχοντας σε άτομα μια μη επεμβατική, οικονομική και αποδοτική μέθοδο αυτοαξιολόγησης η οποία στη συνέχεια μπορεί να οδηγήσει σε συνάντηση με έναν ειδικό. Στα πλαίσια της έρευνάς μας, τονίστηκε η πολυδιάστατη φύση της άνοιας και πως δεν αρκούν μόνο τα φωνολογικά χαρακτηριστικά για την ανίχνευση της νόσου, αλλά απαιτούνται και άλλα είδη δεδομένων και σίγουρα η κλινική εμπειρία και γνώση ενός ειδικού. Το έργο του COMFORTAGE στοχεύει στην ανάπτυξη και παροχή λύσεων για το πρόβλημα της άνοιας, και ελπίζουμε ότι το DEMET μπορεί να προσφέρει κάποια βοήθεια στην επίτευξη των στόχων αυτών.

Chapter 2

Introduction

In the healthcare industry, emerging technological advances have been documented in the shift towards a more patient centric approach to treatment and diagnosis, compared to the hospital centric approach that has been prevalent in the past [26]. This shift has been coined as the fourth healthcare revolution or Healthcare 4.0, promoting the use of data-driven technologies, blockchain applications, fog and cloud computing, cyber-physical systems and sensor enabled devices to improve patient outcomes and experiences [21]. At the core of this revolution is the accelaration of medical innovation and research, and the simultaneous provision of necessary tools and resources to achieve the best possible patient care. Healthcare 4.0 is now being adopted by many healthcare providers, and is expected to be the new standard in the industry.

AI, along with other technologies much like the ones mentioned above, in its rapid development has shown great promise in the healthcare industry, driving healthcare to a new and improved paradigm shift, where the focus is deeply rooted in the patient's well-being and the quality of care they receive, while also reducing costs and improving systems efficiency much like Healthcare 4.0 has promised. This shift is termed Healthcare 5.0, and it involves the integration of millions of IoT devices, which will be interconnected and have the ability to communicate through networking infrastuctures like 5G. These devices will be combined with state-of-the-art AI algorithms to provide personalized and precise healthcare to patients [139].

But AI is not without fault, and the extensive use of AI in healthcare can have serious implications on patient safety [112]. It is apparent that in order for these systems to adhere to the ethical principles and regulatory requirements set by the ever changing landscape of technological innovations, a shift towards a system that is accountable and transparent is necessary. This need has led to the creation of Responsible AI, which is an interdiscplinary and dynamic process, that goes beyond the technical aspects of AI development, and includes the ethical, legal and societal standards that are necessary for transparent and accountable AI systems [54].

Explainable AI (XAI) is a subfield of Responsible AI, and is concerned with the development of AI systems capable of providing explanations for their decisions and actions. With the advancements in AI, models are becoming all the more complicated and convoluted, making it difficult for experts even to understand the decision making process of these models [160]. Simpler models, like linear regression or decision trees, are easily interpretable in their decisions, as they are based on mathematical rules and can be visually represented. These models are considered to be inherently interpretable, meaning interpretable by design. On the other hand, deep learning models, like neural networks, do not adhere to this property, since their scale and complexity makes it difficult to understand how the data in the model contributes to the shifting of the parameters in the model, and ultimately the decision that is made. Explainable

AI aims to solve this issue by providing explanations and interpretations for the decisions made by these more complicated models. In the field of healthcare, where the decisions made by clinicians and healthcare professionals can be crucial to the well-being of patients, the need for transparent and accountable support tools is essential. It is also essential for patients to be able to receive explanations about their diagnosis and treatment, and so it is apparent that a black box solution will not suffice. XAI has the potential to demystify these black box models into interpretable and understandable support tools, promoting trust and accountability in the healthcare industry.

2.1 Background and overview of AI in healthcare

AI is already being utilised in the industry, with applications in a variety of fields. Some of these fields include medical operations and data management, drug discovery, medical diagnosis through image recognition, and personalized treatment plans [119].

Drug Discovery
AI has been a catalyst for pharmaceutical companies in the discovery of new drugs, speeding up their processes and allowing for new ways to identify potential drug candidates [84]. AI's ability to analyze large datasets and identify patterns has been instrumental in processes of drug repurposing and minimising repeated clinical trials and tests. Pfizer, for example, has used IBM's Watson to accelerate immuno-oncology research and find potential treatments [121]. Sanofi has also been investing in AI, collaborating with Exscientia, to produce an AI-based pipeline for drug discovery focusing on oncology and immunology. Genentech, a subsidiary of Roche, in collaboration with GNS Healthcare in Cambridge, has been using AI to identify new drug targets for cancer treatment. There is skepticism surrounding these practices, but proponents argue that AI can assist in faster, cheaper and more efficient drug developement automating a lot of the processes involved in drug discovery, allowing for better understanding of the biological components of the target, optimising drug structure design.

Diagnosis
AI has also been used in the diagnosis of diseases, enabling healthcare professionals to make more accurate and timely decisions. In the field of cardiology, AI has been used for screening and diagnosis of heart diseases like cardiac contractile dysfunction [1] and arrhythmias [2] using electrocardiogram (ECG) data [175] [165]. In the field of pediatrics, scientists have developed AI models to distinguish between control and hydrocephalic groups [3] using MRI data [123]. MRI data has also contributed to the development of AI applications that are able to diagnose children with hypoxic-ischemic encephalopathy [4] [136] using classification and machine learning models. In our field of interest which includes mental health and neurodegenerative diseases, which we will delve deeper into in the next sections, AI has been utilised to diagnose and predict the progression of neurodegenerative diseases like Alzheimer's and Parkinson's, dementia and different types of mental disorders and diseases [74] [15] [101]. Dentistry is another field that has been revolutionised by AI, with systems capable of assessing bone quality for osteoporosis [5] or providing tooth segmentation solutions [77] with architectures like CNN models.

[1] Decrease in contractile function of the heart muscle and prolonged relaxation.
[2] Irregularity in the heart's rhythm and beat.
[3] Hydrocephalus is a condition where there is an accumulation of cerebrospinal fluid in the brain.
[4] Brain injury caused by lack of oxygen before or shortly after birth
[5] A condition where bones become weak and brittle

Surgery and Operating Room

AI has been a prevalent tool in the operating room as well, assisting in intraoperative operations like pharmacotherapy, hemodynamic optimization [6], neuromuscular block monitoring [7], and anesthesia depth assessment [43]. AI stretches its reach to applications in bariatric surgery [8] [30] and the management operations of the operating room [157]. Researchers have also demonstrated that the use of AI can facilitate in the prediction of postoperative complications and mortality, as well as operation success or failure, allowing for clinicians to assess their approach and make necessary adjustments for better patient outcomes as early as 2002 [59].

Patient Care

AI has also been used to fasttrack operations in patient care, namely in remote patient monitoring, where it allows for cost effective and efficient monitoring, optimising in this way hospitilisation and assisting in the avoidance of complications by early detection of deteriorating health [2]. Extensive research has been conducted in the field of robotics and AI, for the creation of exoskeletons and wearables to assist patients in rehabilitation and physical therapy [143] [65]. Chatbots, as we will later discuss in detail, have been used to provide mental assistance and support to clinically depressed patients, allowing for a more accessible and affordable approach to mental health care [3]. Lastly with the use of wearable devices and IoT, AI is at the palm of our hands, with applications able to monitor and track our health and fitness, in real time, providing us with insights and recommendations for a healthier lifestyle [44]. All this data can later be collected and analyzed to be used in the development of new AI applications and models, that can further improve patient care.

It is evident that AI is here to stay. The applications of AI in healthcare can be a cornerstone of innovation and development in the industry, allowing for better patient outcomes and providing clinicians and healthcare professionals with additional support tools for more informed and accurate decisions. The afformentioned applications do not only highlight the capabalities that AI possesses, in augmenting and improving the healthcare industry beyond what was once possible, but also lay the foundation for the importance of further research into ways to make these systems more available for clinical practice.
In order for that to be a possibility however, we need to go back to the importance of Responsible AI, and the need for transparency in these systems and models. The rapid development of AI systems has led to the increase of black box models, and that deters both professionals and patients alike from trusting and using these systems in their practices. The need for explainability in AI is essential in order to combat this issue and promote trust and accountability in the healthcare industry.

2.2 Importance of explainability in AI

The importance of explainability is prevalent in healthcare, where the decisions made by clinicians and healthcare professionals are directly influenced by the support tools they use. AI systems and deep learning models can contribute immensely to the decision making process, but they are not always correct or reliable in their outputs. We must not forget that these models are trained systems based on mathematical foundations, but can statistically fail in

[6] adequate oxygen delivery for tissue needs during operations

[7] stimulation of nervers for muscle movement measuring

[8] Surgery performed on the stomach or intestines to induce weight loss.

their predictions. Not to say that clinicians don't fail in their diagnosis or treatment plans, but the issue arises when the model fails and accountability has to be attributed to the parties involved. The patient also has the right to know why a certain decision was made in order to understand the reasoning behind their diagnosis or treatment and to be able to assess whether or not they want to proceed with a specific plan of action. Multiple cases of misjudgement have been documented where the model has failed to provide adequate results in their task due to different reasons. Such reasons include the lack of diversity in the training data, where the model provides biased decisions towards a specific group of people and hence fails to provide accurate results for those specific groups. Another reason can be the weathering of the training data over time, where due to changes in environment or settings of deployment, the model fails to predict correctly, making it unreliable for the task at hand. Research has shown that the extensive unfiltered use of Artificial Intelligence in the workplace can lead to de-skilling of professionals, where the reliance on the model's decisions can be taken for granted, thus producing a lack of critical thinking and reasoning in the way professionals approach their tasks. This can lead to a substantial decrease in the quality of services provided by healthcare professionals which can have serious implications in the way patients are treated and diagnosed.

The solution cannot be to simply abolish the use of AI in healthcare. That would certainly prove to be detremental to the progress in strides that researchers have made in the field. We have already highlighted the importance of AI in industry and the potential for it is vast and promising. The industry requires for a solution that provides context, a solution that assists both professionals and patients alike and produces in turn, a better environment in which healthcare is provided. Explainability has proven to be one such solution. In this instance, the model stops being a blackbox-like uninterpretable machine, and start to become a dynamic and transparent tool that provides essential insight in its own process of creating an output, providing clinicians with necessary information required for them to make informed decisions and assess a situation accordingly. Explainability, in the process of failure in a model's prediction, can produce outcomes that facilitate the understanding of failure by professionals and thus the mitigation and misjudgement. This process can also provide patients with the necessary information they need to proceed with their own decisions as to whether or not they allow for any procedure or treatment. Individuals have an inherent need and human right to understand the reasoning behind a decision made in their regard, especially when it comes to their health or health services provided to them, and in order for them to be able to trust and accept these decisions, they need to be provided with the necessary explanations as to why a certain decision was made. Transparency allows for accountability and the ability for backtracking a failed decision in order to address who is at fault in a legal standing. Explainability can be the gateway to a more trustworthy and accountable system that can be used widely in clinical practice. These models have the ability to provide the necessary foundations for ethical and legal use of AI in healthcare which is a matter of utmost importance in the industry today.

2.3 Motivation for the study

The motivation for this study is driven by the need to provide a concrete and comprehensive review of explainable AI and its application in the healthcare industry, particularly in the field of mental health and neurodegenerative diseases. This study aims to collect and analyze the current state of the art in XAI, and to provide a detailed overview of the different methods and techniques used in the field. The study will also investigate the potential benefits and challenges of implementing XAI in mental health by reviewing related work in the field, and will explore the ethical and legal implications of using these systems in clinical practice. Lastly, this

research aims to contribute to the development of AI systems that are transparent, accountable and trustworthy. In doing so, it seeks to address the critical gap between the capabilities of AI and the actual implementation of these systems in clinical practice.

2.4 Research questions and objectives

This thesis is guided by the following research questions:

1. How can explainable AI methods be applied to develop transparent and trustworthy AI models for mental health care?
2. What are the current state-of-the-art AI models in mental health care that employ explainable methods, and how effective are they in classifying or predicting various mental health conditions?
3. What are the key metrics for evaluating the explainability of AI models in the context of mental health, and how do these metrics impact the trust and reliability of AI-driven decisions?
4. What ethical considerations must be taken into account when developing and implementing AI models in mental health care, and how can explainable AI contribute to addressing these ethical challenges?

To address these research questions, the objectives of this study are:

1. To review and critically assess the current landscape of explainable AI methods and their applications in mental health care.
2. To identify and evaluate the state-of-the-art AI models that utilize explainable methods for diagnosing, treating, and monitoring mental health conditions, highlighting their strengths, limitations, and areas for improvement.
3. To develop an experimental AI model using transformers and Local Interpretable Model-Agnostic Explanations (LIME) to demonstrate the practical application of explainable AI in mental health care and how it can be used to predict dementia.
4. To explore the ethical dimensions of AI in mental health care, with a particular focus on the role of explainable AI in enhancing transparency, accountability, and trust between healthcare providers and patients.

Through this research, the study aims to contribute valuable insights and practical tools to the field of AI in mental health care, with a particular emphasis on the critical importance of explainability, ethics, and human-centered design in the development of AI systems.

Chapter 3

Literature Review

3.1 Evolution and Current State of XAI

Initially introduced by Van Lent et al [88] in 2004, XAI aimed to make AI systems understandable by humans. Over the coming years, as Machine Learning (ML) and Deep Learning (DL) models progressed, the focus shifted towards the explainability of these models in an assortment of sectors. The term XAI became synonymous with responsible AI, bridging the gap between the black-box nature of AI models and the need for transparency and human comprehension. In 2017, the Defense Advanced Research Projects Agency (DARPA) [57] [1] launched their XAI program, to develop an AI system for better understanding of decision making by end-users. There have also been studies conducted by reputable organisations on XAI and their impact.

Today Explainable AI is used in various domains and applications and enhances traditional preexisting AI models, in a much needed time of transparency and accountability in AI systems due to their increasing prevalence in our daily lives. Their use stretches from e-commerce product reviews with improved sentiment analysis and emotion classification, to military applications with real-time sensor data in military operations, which require data assurance. Other applications include the use of XAI in healthcare as we have already established, transportation, finance, justice, bussiness and many more.

There are specific desired properties which are widely used in research and development of XAI models.

1. **Interpretability:** It refers to a sense of understanding how the model works.

2. **Explainability:** It explains as to how a decision was made.

3. **Transparency:** It assesses the information that is available to the user.

4. **Justifiability:** It refers to the facts that support the decision.

5. **Contestability:** It refers to the ability to challenge the decision by the user.

[1] DARPA is an agency that develops cutting-edge technology and innovations geared towards the national security of the United States.

Explainability can be local, meaning that we are interested in a single instance of a prediction and the model's explanation for that specific instance, or global, meaning that we are interested in how the model works in a population of instances and how the model behaves in general. Local explainability can be formed as a question as to why the model made a specific decision for a specific instance, and global explainability can be formed as a question as to how the model will behave generally in terms of explaining, when it comes to a series of instances. In this particular study, we will focus mostly on local interpretability and explainability.

3.2 Explainable AI in Health Care

XAI in healthcare is of paramount importance. It is used in various clinical assessments of models, data management, diagnosis, reducing sensor bias, disease classification and critical object separation in medical images. These models help in easier error correction and better performance as they explain the results of their decisions. They also have the ability to provide explanation for the entire model and its architecture, as well as for each prediction separately, while also adapting to the needs and conditions of the patient, thus maintaining a high level of trust and reliability.

XAI is being intergated with clinical knowledge to enhance the accuracy and reliability of health diagnostics and predictions. New methodologies leverage multi-modal and multi-center data fusion, utilizing case studies in areas like COVID-19 classification and ydroccephalus[2] segmentation. Research is focusing on the interpretability and explainability of machine learning algorithms, on personalised healthcare services and practical scenarios like ECG data analysis, surgical planning and COVID-19 diagnosis. A generalised taxonomy for XAI is being proposed to address current challenges and guide future research. Frameworks are being developed as well, to collect data for XAI applications. Lastly, cloud-centric systems are being proposed for multi-modal data analysis. [139] Explainable AI plays a pivotal role for the future of healthcare and the vision for Healthcare 5.0.

Healthcare 5.0 is a vision for the future of healthcare that strives for the personalisation of medical interventions on patients based on genetic makeup, lifestyle, and specific health needs. In this vision of the future, the patient and his needs are placed in the center of the decision making process of treatment. The implimentation of advanced technological appliances such as IoT devices, 5G networks and Explainable and Responsible AI models is at the forefront of Healthcare 5.0, as well as the ethical implications of involving such advanced technologies in healthcare.

3.3 Mental Health and XAI

We will focus our attention on the application of XAI in the field of mental health. Mental health, like a plethora of other fields, has been impacted by the rise of AI and digital technologies. The use of AI in mental health has been a topic of interest for researchers and practitioners alike, as it has the potential to revolutionise the way we combat mental health issues. The introduction of digital mental health has established new tools and applications such as personal sensing or digital phenotyping, natural language processing of clinical texts

[2]Buildup of fluid in cavities deep within the brain

and social media content and also chatbots for the betterment of the conditions of patients. [36]

1. **Personal Sensing:** Personal sensing or digital phenotyping is the use of digital devices to collect data on an individual's behaviour and environment. This data can be used to monitor and predict mental health conditions. Research in the area as been particularly involved in the identification of signs of depression and anxiety through indicators such as physical activity and smartphone interactions. Early studies suggest the potential detection of conditions like schizophrenia by monitoring changes in communication patterns. Additionally smartphones serve as a platform for Ecological Momentary Assesments (EMA), enabling real time mental health monitoring through questionnaires. This concept can be extended with the concept of Ecological Momentary Interventions which offers timely psychological interventions based on the data collected from the user, with the end goal being the development of AI-driven applications that provide highly personalised and contexual relevant therapy recommenndations.

2. **Natural Language Processing:** Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. NLP is used to analyse clinical texts and social media content to predict mental health conditions. Characteristics indicative of language disturbance, such as limited vocabulary, lack of semantic coherence, and simple syntax are very telling of severe mental health problems like schizophrenia. These characteristics can be quantified with the use of NLP techniques, and subsequently fed into ML or DL models for mental health prediction and classification with notable accuracy. Additionally the sheer amount of data available online allows for the development of models able to predict mental health issues like depression and phychosis, as far as suicidal tendencies from users' posts, with high accuracy. This research corroborates long-standing observations concering the connection between language and mental health, and offers an accurate, scalable and efficient way of detection and intervention of mental health issues.

3. **Chatbots and other agents:** Agents are designed to simulate conversation with end users, employing a range of simple rule-based interactions, conservational classifications all the way to more advanced and complicated NLP techniques. The origins of said agents in mental health can be traced back to the 1960s, with the development of ELIZA, a computer program that emulated a Rogerian psychotherapist. Chatbots have since been evolved to address various mental health conditions with general positive user satisfaction. Modern day chatbots such as Woebot, Wysa and Tess are designed to provide mental health support and interventions, using techniques such as cognitive behavioural therapy (CBT) and motivational interviewing. These Chatbots show their potential when it comes to users hesitant to seek traditional therapy by professionals, due to stigma or accessibility issues. Further research is needed concerning ethical considerations, handling of emergencies and limitations of these agents.

3.4 Existing Frameworks and models for Explainability

There are two types of general methods to explainability [110]. First, there is the model specific method, mainly geared towards Machine Learning models like Decision Trees,Random Forests and others. Model specific refers to the ability of the method to have prior extended knowledge of the internal structure and inner workings of the model, such as parameters, activation functions and optimisations to the model. The second method is called model agnostic. The way this method works is it provides explainability without any knowledge concerning the way the model actually works but derives all explanations from examining input and output of the model. This method is used predominanlty with Deep Neural Networks, Convolutional Neural Networks and other models where it's extremely difficult to understand how they compute their outputs and make decisions [69].

There are other ways that explainability can be categorised, such as ante-hoc and post-hoc. Ante-hoc refers to models that are inherently interpretable, meaning that their design favours explainability in some sense, mostly due to the usage of decision rules and visualisations. Post-hoc refers to models that are black-box like, like neural networks and transformers, and the explainability is derived by assessing input and output of these models.

Artificial Intelligence has had great success over the past few years, but the black-box nature of models up until recent years has not allowed for major integration in numerous fields where decision making is critical to safety and security. The models that score the highest in accuracy and precision are in fact, in most cases, the most complex, meaning that scientists and professionals alike are hesitant to use them. Regulations also produce a hurdle when it comes to using none transparent models. These explainable methods aim to enstill trust in these models and facilitate this integration.

3.4.1 Interpretable Models

Inherently interpretable models provide explicit rules, feature importance and intuitive representations so that users can understand their decision making process. They are deemed the easiest and fastest way to produce explanations but are limited to the one model that they are designed upon.

1. **Linear Regression** [105] is a statistical model that examines the response of an outcome variable to a set of predictor variables. This regression can be simple or multiple, for a single predictor variable or multiple ones respectively. The output of this model is a function of the sum of weighted predictor variables. These weights are the coefficients of the model and are used to estimate the relationship between the predictor variables and outcome variable. The model is easy to interpret and understand, and is used in a variety of fields such as economics, biology, and psychology. The relationship between a specific predictor variable and the outcome variable can be estimated by holding all other predictor variables fixed, changing the specific predictor variable and seeing how the model reacts to the changes. The importance of a feature can be derived from the magnitude of the coefficient of the variable in the model. These features produce inherent interpretability. The model is limited in its use cases, since it can only model linear relationships between variables, and is not suitable for complex data.
2. **Logistic Regression** [120] is an extension of linear regression for classification tasks. The way logistic regression works is by using the logistic function to model the proba-

bility of a linear equation between 0 and 1. This approach helps combat the problem of linear regression and its inability to model non-linear relationships and perform classification tasks properly. Interpretability is obtained in a different way in the case of logistic regression, where weights don't impact the probability linearly. In this case, the weights are formatted as odds, and the exponent of the weights is used to calculate the odds ratio. The odds ratio is then in turn used to measure the impact of a feature on the outcome. This process can be done for numerical, binary and categorical features, by using any encoding to transform them into binary features. Each feature's odds ratio changes differently depending on the type of feature, but outcomes are easily interepretable.

3. **Decision Trees** [133] are a non-parametric supervised learning method used for classification and regression tasks. The model works by splitting the data into subsets based on the value of a feature. It then continues to split the data into smaller and smaller subsequent datasets until the data is homogenous or a stopping criterion is met. The model is easy to understand and interpret, and visualisation is intrinsicly available due to the model's architectural design. Decisions trees are robust to outliers and missing data, and have the ability to handle numerical and categorical data. Decision Trees can be prone to overfitting and introduce instabilities in the way the model classifies different datasets, meaning a small change in the data can in fact produce different trees. A Decision Tree can be interpreted by simply following the path of the constructed tree and observing the edges that explain the decision as to why each step was taken, until a leaf is reached and the decision is made. Feature importance can be measured using the Gini index to indentify the information gain each time a feature was used for a split.

4. **Decision Rules** [68] [89] are simple IF-THEN statements that contain a condition (IF) and a statement or outcome (THEN). Multiple of these rules can be combined similarly to a Decision Tree to predict a model's decision to a specific input. These rules strongly resemble natural language making them one of the most interpretable solutions for explainability. Each condition can only use one feature, but can be combined with other conditions with AND statements. Decision rules have two metrics to measure quality of a rule, support and accuracy. Support is the percentage of data that the rule applies to, and accuracy is the percentage of data that the rule predicts correctly. There can be a trade-off between the two metrics, where adding more features to a rule can increase the accuracy but decrease the support. In order to get the most out of decision rules, in order to classify a decision, multiple rules should be used to cover one's bases. This can introduce issues and complexity, and these issues are combated by using decision lists and decision sets. Decision lists allow for order in the evaluation of each rule and decision sets allow for majority voting between rules to produce a decision, where some rules can be more important than others. Some methods for extracting these rules include OneR, which learns rules from a single feature, sequential covering, which removes data covered by each rule created sequentially and Bayesian rule lists which use Bayesian statistics [166] to create decision lists, much like the name suggests.

5. **RuleFit** is an algorithm introduced by Friedman and Popescu [49] in 2008, which learns a sparse linear model by combining the original features and a set of decision rules. These decision rules extract the connection between the original features and are generated by a tree ensemble model. Each path through a tree can essentially be translated as a decision rule. Feature importance can be calculated by the coefficients of the linear model. RuleFit also includes partial dependence plots, which are used to visualise the effect of a change of a feature on the model's prediction. Interpretation for linear features in RuleFit is the same as with linear regression, but for the decision rules, the interpretation differs.

Decision Rules are binary rules, which are equal to 1 if the condition is met, and 0 if the condition is not met. An advantage of RuleFit is the induction of interactions of features to the linear model, without having to manually add each one separately. It is used for classification and regression and is very interpretable and easy to understand. A drawback can be detected when it comes to the linear features which are fixed to assess importance and reactions to a singular feature. Performance also can be unstable in some cases, and the model can be prone to overfitting.

6. **Naive Bayes** is a well known family of classifiers which use Bayes conditional probability [38]. It is a supervised and statistical learning method that assumes a probabilistic model and allows for the capturing of uncertainty by calculating the probability of an outcome [159]. Interpretable is derived in Naive Bayes by the clarity in which we can assess why a decision was made based on a single feature by its contribution towards a prediction of a certain class, due to its conditional probability.

7. **K-nearest neighbors** or KNN [16] is a non-parametric method used for classification and regression. The model works by finding the K-nearest neighbors of a data point and classifying it based on the majority class of its neighbors [86]. To explain the way a decision was made by a KNN model, it is advised to collect the K-nearest neighbors of the data point in question and observe the class of each neighbor. The majority decision of the neighbors is the decision of the model.

3.4.2 Model Agnostic Methods

Model agnostic methods are used to explain the predictions of complex models. They are used to produce explanations without any knowledge of the model's internal structure. They are more versatile and flexible since the interpretations can work with any machine learning model. They are also very flexible in terms of explanation format, since the explanations can be in the form of rules, feature importance, or even visualisations.

1. **Local Interpretable Model Agnostic Explanations**, or LIME in sort, originally introduced by Marco Tulio Ribeiro et al. [129] in 2016 is a model agnostic method that explains the predictions of complex models. This is achieved by generating samples of input and evaluating them by the original model. After this evaluation process, LIME tries to approximate the original model using a linear function which is easier to interpret. Essentially LIME produces a surrogate model of the original, a reduction of sorts, to produce explanations. This can be very computationally expensive in some cases, since the quality of explainations are dependant upon the quality of the reduction to this surrogate model, which in turn requires extensive sampling to be obtained. Sampling also introduces uncertainty in explanations and it has shown to produce different explanations for the same input. This instability is prevelant when compared to other model-agnostic XAI methods. [96]

2. **Anchors**, which was also introduced by Ribeiro et al. [130] in 2018, describes human readable simple rules that describe conditions in which a model's prediction is subject to change. These rules are essentially "anchors" that explain the decision boundary of a model for a specific instance. The resulting explanations come in the form of IF-THEN statements, which define the decision boundary region for each anchor. In order to extract these rules reinforcement learning is used, combined with beam search and heuristic search algorithms by cycling through different candidate anchors, filtering out the best ones and finally extending the anchor rules. Anchors are evaluated using two metrics, precision and

coverage. Ribeiro et al. define precision as the accuracy in which the anchor can predict the model's behaviour. Coverage describes the amount of data that an anchor's decision rule applies to. Anchors produce understandable and human readable explanations but can be computationally intensive in tuning hyper-parameters and dealing with constant calls to the original model. Instances close to the decision boundaries of different anchors require extensive feature extraction.

3. **GraphLIME** [70] by Huang et al. is a model agnostic method that explains the predictions of GNN (Graph Neural Network) models and is based on LIME as its name suggests. This particular variation of LIME tries to find a surrogate model based on a non linear feature selection method called Hilbert-Schmidt Independence Criterion (HSIC) Lasso to explain a specific node on a graph. It borrows from the training process of the GNN to extract representative embeddings from each node, which is used in node classification. GraphLIME samples in an N-op neighborhood of nodes to collect features for node prediction, these features are in turn used to train the HSIC Lasso, which is kernel based and thus explainable.

4. **LRP** or Layer-wise Relevant Propagation was original introduced by Bach et al. [22] in 2015. This XAI method requires the internal structure of the model to be explained. It uses the network's parameters and architecture to redistribute the explanable factors of the network from the output layer all the way to the input layer, step by step through every layer using back propagation. This way LRP breaks the original model down into smaller pieces which are easier to explain.

5. **The Deep Taylor Decomposition** or DTD [111] is a propagation-based explanation technique which uses decomposition in order to explain a neural network's decision. Similarly to LRP, it redistrutes the output values to the input values, layer by layer in combination with the first-order Taylor expansion, which is used to extract the relevance of the lower layer while this redestribution process is taking place. This method is highly based on mathematical and theoretical efficiency, making it computationally sound and trustworthy. Researchers have used this method to provide software for further research in explainability with iNNvestigate [14], an interface in Python build on Keras and TensorFlow 2.0, with out of the box implementations for various models, and Zennit [76], a highly customisable framework also in Python, build on Torch.

6. **Prediction Difference Analysis** or PDA [177] is a continuation of the work of Robnik-Sikonja et al. [132] and measures the importance of a feature when the feature is omitted from the prediction, either by actually omitting it, flagging it as unknown or by margilazing it. The relevance of this feature is then measured by the difference in the prediction when the feature is present and when it is not to draw a conclusion on the importance of the feature.

7. **Testing with Concept Activation Vectors** or TCAV [81], is a method of explainability where the amount of influence of a concept is measured in the decision of a model concerning classification. This is done by combining two different datasets, one in which the concept is prevelant and one which is not. A logistic regression is then used on the final dataset to assess whether the concept is present on an image sample or not. The coefficients of the regression used are then formatted as a vector, called concept activation vector, or CAV in short. CAVs are finally used to calculate the conceptual sensitivity of the model, which is a measure of how strongly the presence of a specific concept contributes to classification. The formula for conceptual sensivity is the dot product of the CAV and the gradient of the model's output. TCAV is easy to use and does not require

any extensive knowledge about the model's architecture. The main objective for this method to work is the gathering of the datasets, which a lot of times requires domain specific knowledge and expertise. TCAV can also be used to improve the robustness of a model and eliminate biases or overfitting. Drawbacks can emerge when the concept is not well defined, or is extremely abstract, making difficult to gather data, or detect. Another drawback is that TCAV is only applicable to image data and not text or speech.

8. **Explainable Graph Neural Networks** [167] is an explainability solution for graph neural networks and is part of DIG [92] (Deep Graph Library), a Python library for graph neural networks. Yuan et al. explain [168] XGNN as a method that generates graphs based on nodes of the preexisting graph, to maximise a certain prediction of the model. The graph generation is done via reinforcement learning, where in each step of the generation an edge is added to the original graph. The trained GNN provides information that is used to guide the generation of the graph with a policy gradient method. XGNNs can be used to identify issues with the trained GNN and provide solutions for improvement.

9. **SHapley Additive exPlanations** or SHAP for sort, is a family of explainability methods for explaining machine learning models. SHAP is based on cooperative game theory and Shapley Values, where each player is assessed in terms of their contribution to the outcome of the game [145]. This idea is applied to machine learning with each feature representing a player of sorts, and their contribution to the models prediction is measured through the Shapley Values. This Shapley Value is the average marginal contribution of a feature for an instance across all possible coalitions. SHAP is model agnostic and can be used for any model, making it extremely versatile. Lundberg and Lee introduced a unified framework for explainability in 2017 [94]. SHAP draws from deep mathematical theory, which can prove accurate and trustworthy in its capabilities and outcomes, but a drawback to this is its complexity and computational intensity. When SHAP is used with deep neural networks and high dimensional input use cases, the computations can prove time consuming beyond practicality. Efforts by Lundberg have been made to produce interpretability for tree based machine learning models [95] like random forests, decision trees, and gradient boosted trees which stray towards the model specific side of explainability.

10. **Assymetric Shapley Values** or ASV, is a method of explainability originally introduced by Christopher Frye et al. [50]. The researchers found that the Shapley Value was very restrictive, especially in regards to the causal structure of data. This is due to the symmetry in SHAP, meaning that if the effect of two different variables on a model is the same, they will take identical values, and thus identical attributions. But in the case where a variable as a causal effect on another, the entirety of attributions should be given to the causal variable. This is what ASV can achieve by describing the causal relationship between variables, using a causal graph. When the nodes of this causal graph are not connected, the graph is reduced to Shapley Values. Using this causal graph, the ASV values are calculated as the average effect of adding a variable to other variables similarly to the way SHAP does, but with the added condition that the variable is not a cause of the other variables. This method is particularly useful when doing model fairness analysis, which is a process where a model is assessed in fairness and equality in its predictions by capturing effects on variables by other variables.

11. **Break-Down** [27][28][3] is based on variable contribution analysis, which produces an order in which variables contribute to a model's prediction. The Break-Down method analyses

[3]The book can be found in this link `https://pbiecek.github.io/ema/`

these orders to identify and visualise various interactions between variables inside the model. To make attributions on variables, a greedy euristic function is used to produce a final ordering of variables, which will determine the attributions to be made on each variable. This method is very useful when a model's behaviour in its prediction is not additive, where SHAP would fail [55].

12. **Shapley Flow** [162] is similar to ASV in that it allows for dependency evaluation between variables. Another similarity is that it uses causal graphs to extract these dependencies. The main difference is that attributions are now distributed to the edges of the graph, and not the nodes. The nodes represent variables and the edges represent the relationships between those variables. An added property to this modified graph is that the boundaries hold the original Shapley Values for each explanation. The edge most case of boundary reduces to the ASV values. It is apparent that Shapley Flow contains a lot of information concerning the causal relationship between variables and also the explanation boundaries between variables. This information can make explanations far more robust. A downside to the use of this method is the need for a causal graph, which in turn limits the use of Shapley Flow to only models that have a causal structure. The complexity of the graph also grows exponentially with the number of variables, making the method computationally expensive, and not interpretable by humans for large amounts of variables.

13. **Textual Explanations of Visual Models** is a method of explainability that produces sentences explaining an image sample. This was originally done by combining a CNN model for processing the input image, and an RNN model for learning textual representations. This method produces an easier way to analyse and evaluate data, than attribution maps do. It is apparent that this method requires extensive modification to produce explainability since, these sentences describe images and don't actually explain the model's decision process. Hendricks et al. [64] introduced a method that produced sentences which contain the unique attributes that differentiate images to their specific classes. In essence, the sentences contained those attributes that the model used to produce a prediction and thus explainability is accomplished. This is done through a use of a discriminative loss function to generate sentences which contain class discriminative attributes while also using the relevance loss which produces sentences relevant to the prediction. Backpropagation is done through reinforcement learning with the help of REINFORCE [164]. It is important for validation purposes, to consult experts in the field of the images' domains, to ensure the validity of the explainations proposed by these methods, since there is no other way to check whether they are correct or not.

14. **Integrated Gradient** [151] is a method used widely in deep neural networks and differentiable models. It is based on sensitivity and implementation invariance. Sensitivity assigns attributions to inputs that differ slightly but produce different outputs. Implementation invariance asserts that if the behaviour of two different models is identical, than the attributions of inputs must also be identical. The approach consists of aggregating the gradients of the model's output with respect to the input, over a path from a baseline input to the input of interest. It is essential to produce a proper baseline observation, which is a reference point for the model to compare the input to. This method can only be used with differentiable models, the path from baseline to interest can sometimes not be properly defined or covered. Also the gradient shattering problem can produce issues to the explainability [23] [138] of this method.

15. **Meaningful Perturbations** is a perturbation-based model agnostic explainability method originally introduced by Fong and Vedaldi [45]. A perturbed input sample is inserted into

the model to produce a prediction, and the explainability is derived solely from this prediction due to the perturbation. The model measures the relevance between the perturbed sample and the original one to produce a sparse occlusion map. This method formats the explanation problem as a meta-prediction task and aims to solve this task with optimisation techniques. Drawbacks to this method is the genesis of image artifacts by moving the sample of natural image manifold, and the computational demands, which are far larger than propagation-based methods. Agarwal and Nguyen proposed a method to alleviate the production of artifacts using generative models [5].

16. **EXplainable Neural-Symbolic Learning** or X-NeSyL [39] is a method which combines CNN and SHAP to provide insight for feature relevance in the decision making process of a model. After this process a graph is build that contains information concerning the relationship between features and the constraints that are present in those relationships. A designated loss function is used to punish non overlap between these two axioms. This loss function has shown to help with explainability and performance of the model. It is based upon Neural-Symbolic Learning [52] which derives from prior human knowledge like concept learning, to provide highly explainable and intrepretable solutions like mathematical equations and domain specific languages [103]. This method is designed to produce explainability but lacks in the fact that it needs domain specific knowledge to function properly which is not always available.

17. **Grad-CAM** [142] is a method that produces visual explanations for CNN models. It is based on the gradient of the model's output with respect to the feature maps of the last convolutional layer. The gradients are then used to produce a heatmap that highlights the regions of the image that the model used to make a decision. This method is very useful for image classification tasks, but is limited to CNN models and is not applicable to other types of models.

Concluding this review of state-of-the-art explainable methods, it is important to note that there is space for valuable contribution, especially in the realm of human intervation concerning the validity of explainability predictions. There is a need for software development and research in the field of human-AI interaction, so that experts who are not able to understand the inner workings of AI models, can provide domain specific knowledge and intergration can be achieved so that human experience and knowledge and be enstilled in these models for explainability. Further research in the field of variable casaulity is also needed since it has been shown that these relationships produce more robust and trustworthy explanations.

Researchers	**Method**	**URL**
Marco Tulio Ribeiro et al.	LIME	https://github.com/marcotcr/lime
Qiang Huang et al.	GraphLIME	https://github.com/WilliamCCHuang/GraphLIME
Marco Tulio Ribeiro et al.	Anchors	https://github.com/marcotcr/anchors
Christopher J. Anders et al.	Zennit (LRP)	https://github.com/chr5tphr/zennit
Maximilian Alber et al.	iNNvestigate (LRP)	https://github.com/albermax/innvestigate
Luisa Zintgraf et al.	PDA	https://github.com/lmzintgraf/DeepVis-PredDiff
Been Kim et al.	TCAV	https://github.com/tensorflow/tcav
Hao Yuan et al.	XGNN	https://github.com/divelab/DIG/tree/dig-stable/benchmarks/xgraph
Lundberg and Lee	SHAP	https://github.com/slundberg/shap
Nickalus Redell [126]	SHAP and ASV (experimental)	https://github.com/nredell/shapFlex
Przemyslaw Biecek et al. [27] [24]	Break-Down	https://github.com/ModelOriented/DALEX
Jiaxuan Wang et al.	Shapley Flow	https://github.com/nathanwang000/Shapley-Flow
Hendricks et al.	Textual Explanations of Visual Models	https://github.com/LisaAnne/ECCV2016
Ankur Taly et al.	Integrated Gradients	https://github.com/ankurtaly/Integrated-Gradients
Fong and Vedaldi	Meaningful Perturbations	https://github.com/ruthcfong/perturbexplanations
Jules Sanchez et al.	X-NeSyL	https://github.com/JulesSanchez/X-NeSyL
Jules Sanchez et al.	X-NeSyL	https://github.com/JulesSanchez/MonuMAI-AutomaticStyleClassification
Ramprasaath R. Selvaraju et al.	Grad-CAM	https://github.com/ramprs/grad-cam/

Table 3.1: Model Agnostic XAI Methods

3.5 Applications of Explainable AI models in healthcare

We have already discussed some of the applications of AI in healthcare and established the importance of integrating these technologies into clinical practices. These applications range from drug discovery, to clinical diagnosis of diseases, to patient care and treatment. In this section, we will discuss the integration of explainability in AI applications and how these integrations can be applied to clinical practice in order to improve patient outcomes and facilitate in the adoption of AI in a more transparent and ethical manner in the healthcare industry.

Essentially, the applications of explainable AI aim to improve the afformationed applications of AI in healthcare by providing the addition of interpretability in a model's prediction and output. In that regard, we won't steer away too much from what has already been mentioned in terms of the broad spectrum of applications in the industry, but rather review similar applications in a different light, namely how diagnosis, treatment and patient outcomes can be improved by the use of explainable AI models.

Diagnostic Decision Support

In the field of diagnostic decision support researchers have been making strides in developing systems that integrate explainablity in their models with the core objective of providing clinicians with additional information on what features are important and how these features contribute to the model's prediction. One such research comes from researchers Amoroso et al. [17] who developed an explainable framework that extracts the most important clinical features concerning patients' profiling when it comes to breast cancer treatments using clustering and dimensionality reduction techniques. Through their work, it is apparent that data-driven models can be used to identify features correlated with breast cancer in a more personalised and patient-centric manner. Dindorf et al. [40] developed an explainable classifier based on SVM and RF for spinal posture classification. For this task, researchers employed LIME for explainability. Peng et al. [75] created an explainable framework for the assistance of doctors in prognosing hepatitis patients by comparing between different intrinsicly interpretable models such as regression based, kernel based and decision rule based models to see which one would perform the best in the task. Sarp et al. [140] developed a CNN model capable of classifying different chronic wound types and used LIME to provide clinicians with visual queues and representations on the data for explainability. Another group of researchers, namely Tan et al. [153] used logical neural networks on temporal high-resolution computed tomography [4] (HRCT) bone slices for fenestral otosclerosis diagnosis. XAI was used in order to provide visualisations on the most important features derived for the LNNs. Rucco et al. [134] propose an XAI solution for diagnosing glioblastoma [5]. The researchers computed the local feature importance and relevance in the test set using LIME to produce explainability. Another proposal is derived by Meldo et al. [107] who developed a computer-aided system on diagnosing lung cancer. The first part of the system is using LIME to extract the relevant features from lung segmentations and the second part is responsible for the transformation of the extracted features into natural language explanations. A system for prognostic and diagnostic analysis for traumatic brain injury (TBI) was developed using clustering methods by researchers Yeboah et al. [37] with capabilities of combining data analysis and medical expert knowledge. Interpretations were provided by the system by analyzing how features contributed to creating clusters or discriminating between them. Wang et al. [85] developed COVID-NET, a CNN network for diagnosing COVID-19 from chest X-ray imaging. The researchers employed GSInquire as an explainability

[4] A type of computed tomography that enhances image resolution

[5] the most aggressive and most common type of cancer that originates in the brain, and as very poor prognosis for survival

tool to assess the predictions of their network.

Treatment recommendation systems

Explainable AI can facilitate immensly in the development of treatment recommendation systems, allowing clinicians to assess the best possible plan for a specific patient. The computational ability of AI allows for extensive analysis on patient data and medical records to assess proper parameters for drug selection and dosage. Clinicians, through the explanations given by these models, can understand model reasoning and take into consideration patient genetic makeup, their medical history and drug response. The healthcare industry is rapidly crossing over to a more patient-centric approach to healthcare, with the integration of AI and XAI models. **Personalised Treatment Plans:** Healthcare professionals are now able to produce personalised treatment to patients based on analysis done to their genetics, biomarkers and treatment responses of similar patients. **Predictive Analysis for better Outcomes:** Predictive analysis on this data allows for the identification of specific subgroups of population that can possibly respond negatively to a treatment plan. This obviously allows for further and better adjustements in treatment regimens and dosages. XAI as also been used in infectious disease treatment and management, with applications in selecting antibiotics based on pathogen identification and resistance. Especially in the realm of mental health and psychiatry, explainable AI has been at the forefront of treatment recommendations. XAI models have been used to predict mental disorders and recommend treatment plans such as professional intervention or medication. These recommendations are accompanied by explanations for both clinicians and patients to enhance understanding and trust in the model's decisions [19]. AI models can also efficiently analyze medical imagery and derive critical information for treatment and diagnosis, resulting in the reduction of misdiagnosis and improved patient outcomes. **Treatment Optimisation:** These treatment plans can be optimised, by AI, for patient preferences, resource allocation, cost effectiveness and clinical guidelines and regulations. AI also plays a crucial role in developing precision therapy by identifying and targeting specific biomarkers and genetic mutations associated with diseases, infections and viruses.

Predictive modeling and health risk assessment

XAI has been utilised in health risk assessment and predicting wellness outcomes. In the field of mental health, researchers have developed models able to predict risk of suicide in depressed individuals through tabular data using explainable methods like SHAP and RF models [58]. Another approach to predicting risk was taken by researchers Akter et al. [6] who set out to develop an explainable model for predicting the possibility of a patient suffering a stroke incident. Their approach consisted of an ensemble of classifiers for predicting these incidents. Important features where highlighted using explainability methods and tools. As we have already mentioned in this literature review, researchers have also made great efforts in assessing the risk of individuals progressing into MCI, Alzheimer's and Parkinson's disease. Other applications in this space include predicting the risk of heart attacks or cancer. A team of researchers utilised genetic algorithms, neural networks and fuzzy logic to assess risk of eart attack in patients. The explainability is derived in the form of graphs detailing feature importance [106]. Another team used an explainable approach to assess the risk of skin cancer in patients through 2D imagery [93]. Their impressive scores provided a proof-of-concept for the use of AI in similar tasks. In the topic of cancer, research has been conducted regarding the risk of breast cancer and its developing stages in order to assist in the development of smart sensors capable of detecting the disease in its early stages. The research focuses on the primary identifiers of breast cancer and provides an interpretable solution for feature importance and explanation [73]. Another study by researchers in the field of cancer treatment was done in order to combat the lethality of lung cancer by predicting its risk. The researchers proposed a multi-modal expainable method to

assist clinicians in assessing their cancer patients [144].

Regulatory compliance and legal implications
The integration of AI in healthcare has raised concerns regarding the ethical and legal implications of the use of these technologies in a clinical setting. The European Union as well as the United States have implemented strict regulations regarding the use of AI in healthcare. XAI has the potential to provide transparency and accountability in AI models, in order to combat the regalutory and ethical concerns surrounding these practices. **Patient Privacy and Consent:** The development of AI models and systems should be in accordance with regulation regarding patient privacy and consent, especially when it comes to collecting medical data and patient information 7.2.2. In this regard, research has been conducted in order to approach the matter in a way that is both ethical and legal. Researchers and corporations in the field have developed different frameworks and guidelines in order to assure the preservation of privacy of patients. Different approaches in the matter include collaborative learning, federated learning and synthetic data generation all of which can be of great help in the development of AI models 7.2.3 with ethicacy in mind. **Bias and Fairness:** The development of AI models should be done in a way that is fair and unbiased. Data collection and model development is predominanlty done by humans and is thus subject to certain human biases and errors. There have been a significant amount of instances where bias has influenced the outcome of AI model prediction and thus has produced unfair results for patients of specific demographics 7.3.2. There is a number of ways in order to mitigate bias in AI algorithms and researchers and corporations alike have made great efforts in the development of methods and tools to ensure fairness in AI algorithms. Specific ways of mitigating bias include tools that produce metrics for model evaluation concerning fairness and bias, model simulations under specific conditions and environmental changes in order to evaluate sway in data and model predictions, and lastly frameworks that assess data for bias and fairness across different demographics 7.3.4.

3.6 State-of-the-Art Models and Techniques

In the course of this literature review, a systematic search was conducted across three prominent databases: IEEE Xplore, Science Direct, and PubMed. The search query employed was structured to include a combination of key terms related to explainability and artificial intelligence, as well as their applications in the mental health domain, specifically targeting dementia, Alzheimer's, and general mental health. The terms used were: (explainability OR XAI) AND (mental health OR dementia OR Alzheimer's) AND (deep learning OR machine learning OR transformer). A total of 27 articles were meticulously reviewed, forming the foundation of the insights and discussions presented in this review concerning state of the art research in the field of explainable AI in mental health.

3.6.1 Speech Processing

In the realm of Alzheimer's Disease and Related Dementia (ADRD) detection, several notable advancements have emerged. One such innovation is ADscreen, a speech-processing based screening algorithm for identification of patients with ADRD (Alzheimer's Disease and Related Dementia) by Maryam Zolnoori et al.[178] It does noise reduction and extensive linguistic analysis of patients' recorded speech, and uses Joint Mutual Information Maximization (JMIM) to extract important features in speech. By fusing these features with three different ML architectures (DistilBERT, BiLSTM, CNN) the researchers attained a better understanding of the

correlation between these features and ADRD presence in patients. For the classification task, the researchers employed different ensemble, gradient boosting and kernel based ML classifiers to see how each one performed. They achieved F1-score of 84.64% and AUC-ROC of 92.53%.
Bahman Mirheidari et al. [108] focused their efforts on developing a cognitive impairment assessment system for stroke survivors which predicts the MoCA scores of patients after responding to prompts from the Intelligent Virtual Agent (IVA), a chatbot utility developed for the CognoSpeak system. Data collection was done on site by speaking to patients in person. The model proposed firstly transforms the data collected from speech to transcript, features are then extracted from the transformed data where regression is done for the prediction of MoCA scores and classification for cognitive impairment identification. The results yielded an F1-score of 0.74%, a Specificity of 0.73% and a Sensitivity of 0.75%.
Erik Edwards et al. [42], in response to the ADReSS Challenge, explored the significance of phoneme representations in AD classification. First they used acoustic features extracted from the data by discarding highly correlated features using Correlation Feature Selection (CFS) followed by Recursive Feature Selection with Cross Validation (REFCV) where feature importance is evaluated and removed if it doesn't improve in repetition. These features where used for classifications by a multitude of ML models and yielded an accuracy of 0.74%. They continued by processing transcript data and extracting linguistic features. These features where used to train different Deep Random Forest models, fine tune pre-trained transformer models or train models from scratch. The feature extraction was done by word embeddings like Word2Vec, GloVe and Sent2Vec. The best performing model in this experiment was a pre-trained model using Word2Vec, scoring a 0.926% for accuracy and a 0.923% for F1-score. Lastly the researchers used phoneme representations of data by transcribing the segment text into phoneme written text with the help of CMUDict. Text classifiers (Fast2Text, Sent2Vec, StarSpace) where then trained on this data and yielded results of up to 0.9352% by combining Word2Vec, phonemes and audio.
Junghyun Koo et al. [82] responded to the ADReSS Challenge with a multi-modal feature approach, using acoustic and textual features extracted from the ADReSS Dataset using embeddings which are later set as input to a Scaled Dot-Product Attention Layer followed by a one dimensional CNN. Lastly the CNN outputted embeddings are placed into a BiLSTM that outputs the final classification of AD and regression for the MMSE patient score. The model manages an accuracy of 0.8125% when using ensembled output features.
Ilias Loukas and Askounis Dimitris [74] used the ADReSS Challenge Dataset and utilised several transformer-based models like BERT, BioBERT, BioClinicalBERT, ConvBERT, RoBERTa, ALBERT, and XLNet to process transcriptions. There approach consisted of two different methodologies, the first one was a single task method to classify whether a transcription from the dataset belonged to a person with dementia or not, and the second one was a multi task method where the first task was to classify the above and the second was to identify the MMSE (Mini Mental State Examination) score of each person. Both approaches where evaluated using different metrics such as Accuracy, Precision, Recall, F1-score, and Specificity. They showed that the BERT pretrained model outperforms all other proposed models for the single task and scores 81.66% for Recall, 86.73% for F1-Score and 87.50% for Accuracy. The multi task models performed an average Precision of 73.62%, average Recall of 69.16% and an average F1-Score 64.75 with the MTL-BERT model, which outperforms other multi task models for these particular task in the literature. They employed LIME to produce an interpretable solution for the decision making of their best performing model, the pretrained BERT model.
Changye Li et al. [90] hypothesized that speech impairments characteristic of dementia could lead to systematic errors in Automatic Speech Recognition(ASR)-generated transcripts, which might be leveraged to improve the accuracy of dementia classification models. The study focused on evaluating the impact of ASR errors on the performance of these models, both in

processing imperfect transcripts and during the fine-tuning phase with ASR-generated language samples. They went on to utilise two datasets: the Alzheimer's Dementia Recognition through Spontaneous Speech (ADReSS) and recordings from the Wisconsin Longitudinal Study (WLS), both featuring the "Cookie Theft" picture description task, where they were shown a picture stimuli and where asked to describe what they saw. They pre-processed the audio and transcripts by removing artifacts and resampling audio to match ASR training requirements, and divided the recordings into smaller segments. The ADReSS dataset was used for dementia classification, while the WLS dataset helped adapt ASR models and train a language model. Data was split into training and test portions for model evaluation. The study assessed two ASR models, Wav2Vec2 and HuBERT, alongside a BERT model for classifying dementia in ASR-generated speech transcripts. The models used advanced decoding techniques for generating transcripts and were adapted at three levels: unmodified pre-trained, fine-tuned with task-specific data, and fine-tuned BERT for classification. The evaluation process was done in two phases, the first phase was generating the transcripts with pre-trained and task-specific ASR models, and the second phase was fine-tuning BERT to classify ASR-derived transcripts from the first phase. By repeating the evaluation process 100 times they managed to achieve a best overall performance of 0.863% for accuracy and 0.882% for AUC with the domain specific wav2vec2-large-960h model. In this research, explainability is achieved through the use of SHapley Additive exPlanations (SHAP) during the text classification stage, specifically after the ASR-generated transcripts have been obtained and used to fine-tune the BERT model for classifying between dementia patients and healthy controls. SHAP is applied to analyse how individual input features (words, phrases, or tokens from the ASR-generated transcripts) contribute to the model's classification decisions. By calculating Shapley values for these input features, the researchers can quantify the contribution of each feature to the likelihood of a transcript being classified as indicating dementia or being from a healthy control. This approach provides a detailed explanation of the model's decision-making process, ighlighting which features are most influential in its predictions and offering insights into the linguistic characteristics associated with dementia in spontaneous speech.

3.6.2 Text Processing

Anshu Malhotra's et al. [102] proposed system to help prevent suicide attempts involves extracting text from social media, preprocessing it, fine-tuning pretrained BERT models with mental health datasets, and then applying SHAP and LIME for post-hoc explainability. Additionally, BERTopic is used for unsupervised topic modelling to identify themes and issues discussed by users, which can help in assessing mental health risks and prioritizing urgent cases for treatment and intervention. Preprocessing of user generated data is done before using it for training, cleaning and standardizing it for the transformer's input structure. After this process the data is funnelled to pretrained BERT like models for detecting depression and suicidal behaviours in users. The researchers employed post-hoc XAI techniques, specifically LIME and SHAP, to explain and interpret the decisions of Transformer Language Models trained for detecting depression and suicide from social media posts.
Kumar, Abhinav et al. [83] deployed a methodology where they combined various machine learning techniques such as Random-Forest, KNN, Naïve Bayes, Gradient Boosting and Decision Trees and deep learning models in order to detect depression in English and Arabic speaking social media posts. The deep learning models consisted of BERT-based models combined with LSTM, BiLSTM and GRU models. They managed to provide explainability to this methodology by evaluating the importance of the words inside each tweet and applying a color

to each word, thus created a heatmap that allows for explainability in the decision making process of the model, similarly to what others researchers have done in the space. The researchers evaluated the performance of their models for detecting depressive indicators in tweets using various metrics such as Precision (P), Recall (R), F1-score (F1), confusion matrices, and AUC-ROC curves. For training and testing the models they used three datasets: two in Arabic (D1 and D2) and one in English (D3). For the conventional machine learning models, Random Forest (RF) was found to be the most effective across all datasets. Specifically, for the Arabic dataset D1, RF achieved high scores in precision, recall, and F1-score (0.98), with its performance visualized in confusion matrices and ROC curves. Similarly, for the Arabic dataset D2, RF again outperformed other conventional models with scores of 0.84 (precision), 0.83 (recall), and 0.82 (F1-score). For the English dataset D3, RF achieved precision, recall, and F1 scores of 0.62, 0.59, and 0.52, respectively. In the exploration of deep learning models, the researchers found that for the Arabic datasets (D1 and D2), the Arabic- camelBERT + Bi-LSTM model outperformed both conventional and other deep learning models. For D1, this model achieved perfect scores (1.00) in precision, recall, and F1-score. For D2, it also performed best with scores of 0.82 for both precision and F1-score, and 0.83 for recall. For the English dataset D3, the fine-tuned RoBERTa model was the most effective among the deep learning approaches, achieving scores of 0.61 in precision, recall, and 0.60 in F1-score.

Yeldar Toleubay et al. [155] developed a model based on Logical Neural Networks (LNNs) to classify mental disorders. Their LNN model differs slightly from standard neural networks, primarily because its neurons operate under the constraints of logical gate truth functions and handle both upper and lower bounds for logical predicates or sub formulas, making them more complex than typical dense neurons. In their approach, they designed the LNN with four AND logic gates, each serving as a binary classifier for a different class of mental disorder. The inputs to these gates are predicates derived from patient utterances, with the training data consisting of samples that represent truth values for these predicates. After training, the model assigns weights to each predicate and outputs scores for an input, which are the averaged lower and upper bounds for each mental disorder class. The effectiveness of each logic gate, acting as a binary classifier, is assessed using True Positive Rate (TPR) and False Positive Rate (FPR) metrics, alongside the generation of ROC curves to visualize performance across different thresholds. This setup allows the researchers to evaluate the model's ability to accurately identify instances of each mental disorder. The researchers implemented three pruning methods to reduce training time which was found to grow exponentially as predicates grew in numbers. By grouping together similar predicates (similarity pruning) the number of predicates was reduced by half, they removed predicates that were not unique to a single class and also eliminated those that appeared only once (exclusive pruning) which showed a significant variation in the number of predicates amongst different classes and lastly they prioritised predicates which appeared more frequently then others (frequency pruning) which showed that a large portion of predicates appeared only once, and also did not significantly impact the model's scores in any way. To evaluate the model they employed a pretrained BERT model fine-tuned for the task at and showing AUC scores above 0.72 for all classes treated as binary classifiers but only 58% accuracy in a multi-class setting. The baseline LNN managed to score an AUC score of 0.76 for anxiety, and the exclusive LNN managed to score an AUC score of 0.79 for depression, though the DL model outperformed all LNNs for binary classifications.

Elma Kerz et al. [80] approach consisted of creating three different types of models for mental health detection using text based data from the Self-reported Mental Health Diagnoses (SHMD), Dreaddit , GoEmotions, and MBTI Kaggle datasets. The first type of model is a Bidirectional LSTM either on GLFs(General Linguistic Features)(model A), LTFs(Lexicon-Based Features)(model B) or both(model C). These features provide a comprehensive approach to understanding and analyzing language in the context of mental health and provide inter-

pretability. The second type is a pre-trained fine-tuned mentalRoBERTa model and the third type is a multi-task fusion model utilizing the mentalRoBERTa model being trained on two (MHC + emotion recognition, MHC + personality detection) or three tasks (MHC + emotion + personality) with the predicate that the model will use all knowledge about emotion and personality to perform better for the task at hand. For type 1 models the researchers used SP-LIME (Submodular Pick Lime) for interpretability and for the transformer based models they used LIME and AGRAD a self-explaining method based on attention gradients. Type 1 models' performance ranged from 57.14% to 70.78% with the highest performance being in detecting stress. Detecting stress also scored highest for the Type 2 models with an F1 score of 81.62%.

Ahmed H. Alkenani et al. [15] contributed significantly in the field by utilizing patient language samples describing the cookie theft picture, which were later transformed into transcripts using the CHAT transcript protocol. They used a feature space consisting of lexicosyntactic and n-gram vocabulary features which were unified and later filtered to extract the most prevalent and important features for the task. Feature selection was done using Pearson's correlation. They researchers employed 10 fold cross validation to train different base ML models which were after combined into an ensemble fusion model. The fusion model scored an AUC of 98.1% and accuracy and F1-score of 95% for spoken data and an AUC of 99.47% and accuracy and F1-score of 97% for written data.

3.6.3 Multi-modal

David Ortiz-Perez et al. [118] used DementiaBank audio and text data to produce a multimodal classification model for dementia. To create feature vectors for audio they produced a Mel Spectogram for each MP3 file and later set it as input to a CNN which outputs the audio feature vectors. For the creation of text feature vectors, they fine-tuned a BERT model with CHAT encoded text, which in turn outputs word embeddings into an LSTM that lastly outputs into a dense layer that produces the final text feature vectors. They tested each model separately, and later concatenated the two for the multimodal approach. The best result was obtained with the text based approach with an accuracy score of 90.36%. To explain the transformer's decisions, the researchers used the Transformers Interpret software proposed by Charles Pierse in 2021 .

The main objective that Pavan Rajkumar Magesh et al. [101] set out to accomlish is the creation of a deep learning model capable of diagnosing early stage PD using SPECT DaTSCANs, provide an comprehensive analysis of the results of the model and provide an intrepretable solution using LIME. DaTSCANS and other imagery as been used in the past by researched for PD classifications with great success. Namely Towey et al used the Naïve Bayes classifier and Principal Component Analysis for such task, Oliveira et al used Support Vector Machines for the same task and Martinez-Murzia et al proposed the use of CNNs for PD classification. The researchers preprocessed SPECT DaTSCAN images from the Parkinson's Markers Initiative (PPMI) dataset by performing attenuation correction, reconstruction, and spatial normalization to a standard coordinate system, followed by cropping, contour detection, and intensity normalization to standardize the images and enhance feature visibility, particularly in the putamen and caudate regions, before resizing them for compatibility with the VGG16 neural network architecture that they used. The model achieved an accuracy of 92.0%, specificity of 81.8%, sensitivity of 97.5%, precision of 90.9%, with a Cohen's Kappa score of 0.81% and an F1 score of 0.94%. The researchers improved their model scores by optimizing the threshold value for classifying the predicted probabilities into PD and non-PD classes by changing the

original threshold of classification to a more optimal value.The final scores after optimizing the threshold were an accuracy of 95.2%, specificity of 90.9%, precision of 95.2%, with a Cohen's Kappa score of 0.89% and an F1 score of 0.96%. The researchers used LIME to ighlight the specific regions within the brain images, particularly the putamen and caudate regions, which the model deemed most influential in making its predictions. By applying LIME, they could visually demonstrate how the model distinguishes between healthy controls and Parkinson's disease (PD) patients based on the appearance of these regions in the SPECT scans. This visual explanation made it easier for non-experts to understand the basis of the model's diagnoses, by showing which parts of the image were most important for the model's decision-making process.
Das et al. [9] addressed the challenge of diagnosing and treating Alzheimer's Disease (AD) in a cost-effective and interpretable manner, especially given the limitations of current diagnostic tools like cerebrospinal fluid (CSF) tests and neuroimaging (MRI, PET) which are accurate but often expensive, invasive, and not widely available. Their approach was to develop a computer-aided diagnosis (CAD) framework that leverages large and diverse datasets from studies like the Alzheimer's Disease Neuroimaging Initiative (ADNI) to assist medical practitioners in making more informed decisions about AD diagnosis and treatment. SHIMR is part of a multistage diagnostic process, where it serves as an initial screening tool using less expensive and more accessible biomarkers, such as blood tests. This is the first step before potentially moving on to more advanced and costly diagnostics like CSF analysis or MRI, depending on the outcomes of the initial stage. A distinctive feature of SHIMR is its rejection option, which allows the model to withhold making a decision on cases that are difficult to classify, i.e., patients who are close to the decision boundary of the model. This ensures that only patients for whom the model as high confidence are classified at this stage, thereby reducing the risk of misdiagnosis. The model employs interval conjunction rules to make its decisions. These rules are sets of "if-then" statements that are both interpretable and accurate, making it easier for medical practitioners to understand the basis of the model's decisions. The interval conjunction rules contribute to the interpretability of SHIMR by providing clear and concise decision sets. These decision sets allow practitioners to see exactly why the model made a particular diagnosis, including which features were most influential in the decision-making process. SHIMR showed strong performance in AD vs. NC classification, with an AUC (Area Under the Curve) of 0.86% during internal cross-validation, high sensitivity of 0.84%, and a specificity of 0.69%, indicating its reliable accuracy in distinguishing between the two groups. Compared to a Decision Tree, which is a known interpretable, SHIMR not only provided better interpretability but also achieved a igher classification AUC of 0.86 compared to the DT classifier's AUC of 0.73. This suggests that SHIMR can maintain a balance between interpretability and accuracy more effectively than traditional DT classifiers. Further experimentation showed that increasing interpretability (e.g., by simplifying the model) could lead to a slight reduction in classification accuracy. However, SHIMR was able to maintain a reasonable accuracy (AUC = 0.79%) even with a highly interpretable model (rule length 18), surpassing DT's accuracy (AUC = 0.73%).
C.R. Aditya et al. [4] sought to develop a computational method to distinguish between non-demented (ND) subjects and those with Alzheimer's Disease (AD) using data from the Open Access Series of Imaging Studies (OASIS) (https://www.oasis-brains.org). The researchers employed MAA, which involves generating a Multifactor Affiliation Table (MAT) that simplifies the complex relationships between multiple data features. This method reduced high-dimensional data into a more manageable form, capturing the inter-feature relationships important for distinguishing between ND and AD subjects. The study further analyzed individual subjects by calculating their "AD affiliation distances," a measure of how closely a subject's data aligns with AD characteristics compared to the ND group. This analysis used Euclidean distance calculations and the MAT to assess the degree of affiliation. The AD affiliation distances were refined using a weightage factor 'W', adjusted according to the subjects' Clinical

Dementia Rating (CDR) scores to enhance the differentiation between ND and AD subjects. After the refinement process there was a clear differentiation between ND and AD subjects into two separate groups. The explainability of this research lies in its methodical approach to breaking down complex data into more understandable and interpretable components.
Ahmad Wisnu Mulyadi et al. [113] proposed the XADLiME (Explainable Alzheimer's Disease Likelihood Map Estimation) and ADPEN (Alzheimer's Disease Progressive Engaging Network) models to assess the progression of Alzheimer's disease and to provide explainable likelihood maps for the disease. XADLiME integrates clinically relevant information with neuroimaging data to create explainable likelihood maps. These maps are designed to visually represent the likelihood of Alzheimer's disease. ADPEN leverages machine learning techniques to analyze MRI data and predict AD progression. They used 2 different datasets for training and evaluation, ADNI (Alzheimer's Disease Neuroimaging Initiative) and GARD (Gwangju Alzheimer's and Related Dementia). Overall, XADLiME demonstrates strong performance, particularly in distinguishing between CN and AD cases (AUC of 0.9492%, balanced accuracy of 0.9183%, and an F1-Score of 0.9182%), as indicated by the high AUC, balanced accuracy, and F1-Score in both the ADNI and GARD scenarios. However, its performance appears to be varied across different classification tasks and datasets, with some scenarios showing lower scores.
El-Sappagh et al. developed a model consisting of two main layers, each with its own oracle classifier based on Random Forest (RF) and 22 explainers and trained on the ADNI dataset and other EHR data. For interpretability, they used Decision Trees (DT), SHAP and Fuzzy Rule-Based Systems (FRBS) as explainers to interpret the oracle decisions. These models were chosen due to their accuracy and explainability. The model outputs diagnosis and prognosis decisions along with a multitude of explanations concerning the model's decision making in various forms such as visualisations and fuzzy natural language. In detecting AD patients the model scored a multi class accuracy score (MCA) of 93.33% and a multi class F1 score of 93,82% using specific modulaties. In predicting whether MCI patients would progress to AD the model scored 88% for precision and accuracy and 87% for F1 score. Compared to other classifiers the model showed igher performance overall.
Mahmud et al. [13] used established convolutional neural networks to produce ensemble models of VGG and DenseNet models respectively using MRI imagery of AD patients. The ensemble models outperformed the standalone models and scored up to 95% for accuracy, 91% for precision, 90% for recall and 89% for f1-score (Ensemble 2). To improve upon those finding and produce interpretable solutions to AD classifications they proposed a novel model which utilises saliency maps [6] and grad-CAM. The model achieved impressive results with accuracy reaching 96%. For the proposed model the researchers used EfficientNet for feature extraction and proceeded to create a CNN for classification.
Jahan et al. [11] set out to create an AD predictive model using a multimodal dataset. The dataset was created after fusing clinical, phychological and MRI data. The researchers used the KNN algorithm to fill in missing values and used statistical analysis to evaluate feature relevance. Feature selection was done using the Pearson's correlation to identify and retain important features while removing highly correlated redundant features. To address class imbalance in their dataset they applied Synthetic Minority Oversampling Technique (SMOTE) before model training. The model used was a Random Forest with 10-fold cross validation. Explainability was achieved using SHAP and the model achieved 98.81% accuracy.

[6] an image that ighlights the regions in which an eye focuses first

Researchers	AI	Results	XAI
Maryam Zolnoori et al.	Speech processing, JMIM, DistilBERT, BiLSTM, CNN	F1-score: 84.64%, AUC-ROC: 92.53%	-
Ahmed H. Alkenani et al.	CHAT transcript protocol, lexicosyntactic features, ensemble fusion model	Spoken data: AUC 98.1%, F1-score: 95% Written data: AUC 99.47%, F1-score: 97%	-
Bahman Mirheidari et al.	Intelligent Virtual Agent (IVA), regression, classification	F1-score: 0.74%, Specificity: 0.73%, Sensitivity: 0.75%	-
Erik Edwards et al.	Acoustic features, CFS, REFCV, Deep Random Forest, Word embeddings (Word2Vec, GloVe, Sent2Vec), CMU-Dict, Text classifiers (Fast2Text, Sent2Vec, StarSpace)	Accuracy: 92.6%, F1-score: 92.3%, Combined Accuracy: 93.52%	-
Junghyun Koo et al.	Scaled Dot-Product Attention Layer, one-dimensional CNN	Accuracy: 81.25%	-
Anshu Malhotra et al.	BERT, BERTopic	D1(MentalBERT): 0.888% Accuracy, D2(PHSBERT): 0.967% Accuracy, D3(PHSBERT): 0.98% Accuracy	SHAP, LIME
Abhinav Kumar et al.	ML classifiers, BERT-based transformers, BiLSTM, GRU	camelBERT + BiLSTM: Perfect Scores for precision, recall, F1-score (D1) 0.82% for precision and F1-score, 0.83 for recall (D2)	Word Coloring Heatmap
Yeldar Toleubay et al.	Logical NN, BERT, Pruning	Single-Class: AUC 0.72%, Multi-Class: AUC 0.58%, Baseline LNN: AUC 0.76% for Anxiety, AUC 0.79% for Depression	LNN
Changye Li et al.	Wav2Vec, HuBERT, BERT	Accuracy: 0.863%, AUC: 0.882%	SHAP
Ilias Loukas and Dimitris Askounis	BERT Variants, XLNet	Single-Task: 81.66% Recall, 86.73% F1-Score, 87.50% Accuracy	LIME

Table 3.2: Related Work

Researchers	**AI**	**Results**	**XAI**
Elma Kerz et al.	BiLSTM, mental-RoBERTa	81.62% F1-score (Stress Detection)	SP-LIME
David Ortiz-Perez et al.	CNN, BERT, CHAT encoding, LSTM	90.36% Accuracy	Transformers Interpret
Pavan Rajkumar Magesh et al.	VGG16	95.2% Accuracy and Precision, 0.96%	LIME
Das et al.	SHMIR	0.86% AUC	SHMIR
C.R. Aditya et al.	MAA	-	MAA
Ahmad Wisnu Mulyadi et al.	ADPEN	0.9492% AUC, 0.9183% Accuracy, 0.9182% F1-score	XADLIME
El-Sappagh et al.	RF, DT	93.33% Accuracy (MCA), 93.82% F1-score (MCA), 88% Precision and Accuracy (MCI to AD progression)	SHAP, DT, FRBS
Mahmud et al.	VGG ensemble, DenseNet ensemble, EfficientNet	96% Accuracy	Saliency Map, grad-CAM
Jahan et al.	KNN, Pearson's corellation, SMOTE, RF, 10-fold validation	98.81% Accuracy	SHAP

Table 3.3: Related Work (Continued)

Researchers	Datasets	Type	Availability
Maryam Zolnoori et al.	English Pitt Databank	Speech	-
Erik Edwards et al.	Provided by ADReSS Challenge	Speech	-
Junghyun Koo et al.	Provided by ADReSS Challenge	Speech	-
Changye Li et al.	ADReSS, Wisconsin Longitudinal Study	Speech	Available
Ilias Loukas and Dimitris Askounis	ADReSS	Speech	Available
Bahman Mirheidari et al.	On site recordings	Speech	Not Available
Ahmed H. Alkenani et al.	Dementia Bank, Alzheimer's Disease Blog Corpus	Speech (DB), Text (ADBC)	Available
Anshu Malhotra et al.	Rezazabeh's Twitter Depression Detection Dataset, Komati's Kaggle Suicide and Depression Detection Dataset, Murarka's Reddit scraped Dataset, Haque's Reddit scraped Dataset	Text	Partially Available
Abhinav Kumar et al.	Modern Standard Arabic mood changing and depression dataset, Depression Detector **??**	Tweets	Available
Yeldar Toleubay et al.	Counseling and Psychotherapy Transcripts	Text	Available
Elma Kerz et al.	Self-reported Mental Health Diagnoses (SHMD), Dreaddit , GoEmotions, MBTI Kaggle	Text	Available
David Ortiz-Perez et al.	DementiaBank	Audio and Text	Available
Pavan Rajkumar Magesh et al.	Parkinson's Progression Markers Initiative	DaTSCAN	Available
Das et al.	ADNI Dataset	Multi-modal	Available after Application Review
C.R. Aditya et al.	OASIS Data	Neuroimaging	Available
Ahmad Wisnu Mulyadi et al.	ADNI, GARD	Multi-modal	Available after Application Review
El-Sappagh et al.	ADNI	Multi-modal	Available after Application Review
Mahmud et al.	Alzheimer MRI Preprocessed Dataset from Kaggle	MRI Imagery	Available
Jahan et al.	OASIS-3 Dataset	Multi-modal	Available

Table 3.4: Datasets used by Researchers

This literature review has traversed the evolving landscape of Explainable Artificial Intelligence (XAI), with a focused examination of its application within healthcare, including mental health. It has highlighted the progression from foundational models and frameworks to state-of-the-art techniques in areas such as speech and text processing, and multi-modal approaches. The review underscores the critical role of interpretability and transparency in AI models, which is paramount in sensitive sectors like healthcare where decisions can significantly impact human lives.

Despite considerable advances, the integration of XAI in healthcare still faces significant challenges especially concering ethical implications when it comes to practical and clinical use. Furthermore, specific areas such as mental health are still in the early stages of employing XAI, signaling a vast arena for future research.

Future studies should focus on developing more robust, interpretable models that do not compromise performance for transparency. These systems should be geared towards the ethical application of XAI, in order to promote their usage in clinical practice. There is also a pressing need to create standardized frameworks for evaluating the efficacy of XAI systems in real-world medical settings. As AI continues to permeate healthcare, the ethical implications of explainability become increasingly important. Addressing these concerns will not only improve patient outcomes but also enhance the trust and acceptance of AI systems by healthcare professionals.

Through a diligent exploration of existing models and emergent techniques, this review lays a foundation for further inquiry and practical experimentation within the field. The continuous advancement in XAI promises to bridge the gap between AI capabilities and human understanding, ultimately leading to more informed and ethical decisions in healthcare.

Chapter 4

Evaluation of Explainable AI models

When dealing with Explainable Artificial Intelligence that produces explanations in order to assist decision makers in their tasks and pursuits, a question quickly arises to the validity of the explanations provided by the model [66]. So what makes an explanation valid? Most of the research done on the topic of explainability has been strongly focused on the development of new methods and techniques to produce explanations while trying to increase performance in the process. There is a fundamental gap in the literature when it comes to connecting explainability to specific use cases and little research has been done to evaluate the quality of explanations produced by XAI. Consequently, a need arises for the quantification of the quality of XAI explanations in order to be able to evaluate method performance and to compare and contrast different methods in relation to specific tasks.

Now an issue in the context of evaluation, is that explanations provided by these methods are often highly subjective and context dependent, making it difficult to evaluate their quality given the difficulty of defining what a good explanation really is. To add to the complexity, there is still no agreed upon metric of explanation quality by researchers in the field. The domain-specific nature of explanations makes it difficult to generalize the evaluation process across different applications, as the quality of an explanation is highly dependent on the context in which it is used. We will now attempt to provide an overview of the evaluations metrics and criteria surrounding the explainability of XAI models as provided by the literature using as base the roadmap on metrics provided by Zhou et al. [174].

According to the Alan Turing Institute [72], explanations can be categorised into six types which are listed below:

- **Rationale explanation** These explanations answer as to why a decision was taken, providing reasons and rationale for that decision. This answer is provided to the user in a non-technical manner, with simplicity in mind. This specific explanation, should it fail to meet the expectations of developers, can be used to assess the model's reasoning and correct flaws.
- **Responsibility explanation** These explanations answer as to who is responsible for the development, management and implementation of the AI system, this way it provides a clear pathway of communications for inquiry about the model's decision.
- **Data explanation** These explanations provide insight on the data used for a specific decision made. These insights include the data sources, and how the data was used to support decision making. These types of explanations are important in order for users to understand how data influences the model's decision.

- **Fairness explanation** These explanations provide insight on the steps taken during design and implementation of the model to ensure fairness in the model's decisions. These explanations ensure that the model's decisions are not biased and can act with equity towards all users.

- **Safety and performance explanation** These explanations answer as to how the model provides maximum performance while ensuring safety, security and reliability across decisions. These explanations are especially important when it comes to regulatory compliance and ensuring that the model is safe to use.

- **Impact explanation** These explanations provide insight on the impact of the model's decisions on the user, society and the environment. These explanations are important in order to understand the consequences of the model's decisions and to ensure that the model is used responsibly.

A common perception on the matter is that the main factors of measuring understandability in the context of XAI are the features of the system on which the explanations need to be provided upon and the user's cognitive abilities of understanding the explanations provided. There are three main methods of evaluating explainability according to a widely cited paper by Doshi-Velez and Kim [41].

- **Application-grounded evaluation** This kind of evaluation requires conducting human experiments in the context of an application. It is used to test the effectiveness of the explanation in relation to the application and its performance. Whether the explanation performs well or not is direct evidence of the success of the explanation. The effectiveness of the explanation is measured by how it helps the user complete the task at hand and is widely used in decision making tasks.

- **Human-grounded evaluation** This kind of evaluation requires conducting human experiments in the context of a simpler task that is strongly related to the application, and thus keeps the essence of it without the complexity. These experiments do not require domain expertise and can be conducted by any user. The effectiveness of the explanation is measured by how well the user understands the explanation.

- **Functionally-grounded evaluation** This kind of evaluation does not require human experiments but is based on a formal definition of an explanation as proxy to evaluate the explanation quality of a model's decision.

4.1 Scoring the Explainability of XAI Models

The first two categories depend strongly on the pool of users selected for evaluating the explanations of a model. The usage of these types of evaluation methods can provide strong support in evaluating the quality of explanations but can be expensive and time consuming since they require the gathering of human subjects and the conducting of experiments. In addition to these drawbacks, the evaluations are prone to subjectivity and bias and the conducted experiments require thorough planning and execution in order to provide reliable results. On the other hand, functionally-grounded evaluation can be less expensive and time consuming, provided that the proxy used to evaluate the explanation quality is well defined and reflects the quality of the explanation.

Application-grounded and Human-grounded evaluation

The evaluation of the quality of explanations can be done using a variety of metrics and criteria. Firstly, the two main categories of metrics used to evaluate the quality of explanations in human-grounded and application-grounded evaluation are as follows:

- **Subjective or Qualitative metrics** These metrics are based on the user's perception of the explanation quality. They include user trust, confidence and satisfaction with the explanation. Hoffman et al. [66] provided a list of subjective metrics much like the ones mentioned. The literature suggests that these types of metrics are in the center of evaluating the quality of explanations. Zhou et al. [173] found that providing the correlation between features and target variables affected user confidence in the explanation and subsequent decision.

- **Objective or Quantitative metrics** These metrics are based on objective information concerning the task at hand or the human behavior when interacting with a decision or explanation. These metrics include physiological, behavioral and psychological indicators provided by humans and task completion success and length. Zhou et al. [172] investigated how physiological signals such as Galvanic Skin Response (GSR) and Blood Volume Pulse (BVP) of users interacting with XAI can be used as objective metrics to evaluate the quality of explanations and user trust. Again it was highlighted that presenting how features correlate with target variables in the explanations enhances user trust in the model.

Functionally-grounded evaluation

The evaluation metrics of explanations when it comes to functionally-grounded evaluation are placed into three broad categories of quantitative metrics and are as follows:

- **Model-based metrics** These metrics use the model to evaluate the quality of explanations. Various metrics such as model size and runtime operations counts are included in this category of metrics. Model size such as amount of nodes or number of rules in a decision tree, and boolean or arithmetic operations can be used to provide insight about the model's complexity which has shown to be directly correlated to the explainability of the model as we have discussed in earlier sections. Reducing complexity in the model can greatly impact its interpretability. Another metric included in the category is the measure of agreement between the model's decision and the explanation provided by the XAI model, which can express the level of clarity and correctness in the explanation.

- **Attribution-based metrics** These metrics quantify the importance of features in the model's decision making process and their explanatory capabilities. Metrics in this category include monotonicity, non-sensitivity, and effective complexity for the assessment of explanation qualities with individual features all proposed by Nguyen and Martinez [116]. The researchers propose a method to assign feature importance through feature attributions. These attributions are then used to measure the direction of the feature's influence on the model explanations, which describes the monotonicity metric. Non-sensitivity on the other hand, ensures that zero-importance is assigned to attributions that do not influence the model's explanations. Effective complexity is used to measure the effects of non-important features on the model's explanations. If a series of attributions are non-monotonic, it is an indication that these attributions do not provide proper feature importance. If effective complexity appears to be low for a series of attributions, it is indicative of the fact that some features can be emitted or altered without affecting the model's explanations.

- **Exampled-based metrics** Example-based explanations summarise a model by providing representative examples or high-level concepts. Nguyen and Martinez [116] provide quantitative metrics for example-based explanations. The researchers propose non-representativeness to measure the amount of representativeness in the examples. They also define diversity to measure the degree of integration of the explanation. Lastly they define simplicity as the number of examples needed to compute non-representativeness and diversity. The smaller the number of examples, the simpler the explanation and the easier it is for a human to understand it.

Chapter 5

Methodology

In the introduction of this thesis, we discussed the importance of explainability in AI models, especially in healthcare. In this chapter, we address one of the main questions of this thesis, which is the application of explainable AI in mental health in order to ensure trust and transparency in AI-driven clinical practice. In order to address this question, we will develop and deploy explainable AI systems and evaluate their performance and applicability.

5.1 Research design

Our research design employed a hybrid approach, combining both qualitative and quantitative research methods. The first phase of our research included a comprehensive literature review of State-of-the-Art explainable AI models in healthcare and mental health, as well as explainable metrics and applications. This was done with a systematic review of the literature, in order to identify the most relevant and recent studies. The second phase of our research included data collection,processing and statistical analysis of spontaneous speech data from patients with dementia. The data collected was processed and analyzed in order to extract phonological features, which were then used to develop and evaluate explainability. These features were transformed into a more interpretable format, which allows for easier interpretation and understanding by clinicians and other users. The third phase was the initial model development and evaluation of explainable methods using spontaneous speech and phonological features derived from the collected data. Once the model was developed and the explanations from different explainable methods were generated, a survey was conducted to evaluate the explanations. The survey's participants were clinicians, including mental health professionals and neurologists. They were chosen to assess whether the explanations could be useful in clinical practice, and provide feedback as to whether the model could be trusted and used in clinical practice. Once the feedback was collected, we used the best performing explainable method in order to develop an ensemble explainer and an ensemble model, which was then evaluated using the same dataset. The final phase of our research was the development and deployment of an explainable interface, which was designed to be user-friendly, practical and accessible to clinicians. The tool's purpose is to provide explanations for predictions made, and streamline the decision-making process in clinical practice. It is important to note that these tools are not meant to replace clinicians, but rather to assist them in making more informed decisions. We will now delve deeper into each phase of our research design.

5.2 Data collection methods

5.2.1 Collection and Processing

The data collected were spontaneous speech samples in written transcript form from patients with dementia and control. The data were collected from DementiaBank [87], which is a well-established and reputable database containing real-world data from interviews with patients [53]. The interviews were conducted by trained professionals and were in the form of different well known tests, such as verbal fluency, story recall, and picture description. For the processing and analysis of the data, we created a Processor Object, which was used to to create three different datasets from the transcripts for training and testing the transformers. We initialised the Processor Object with configuration details, including paths to input files, an output path, and patterns for text replacement. For each text file from DementiaBank, specific patterns like timestamps and speaker labels are removed or replaced to clean the text. A very important replacement was made to introduce greater interpretability and simplicity by transforming specific CHAT symbols to more interpretable tokens. These tokens are phonological features derived from patients. We will later discuss as to how these tokens can be used as evidence of dementia or non dementia. These tokens are called CHA tokens and are very helpful in the later stages of producing explainability to our models. The table 5.1 includes all the specified patterns and their replacements in a clear and structured format. For the conversion, the official CHAT Transcript Format was used, provided by DementiaBank. Due to automated approach of the Processor Object, a large number of CHAT symbols were not found on the data provided by the corpuses and thus will not be present in the analysis of the data.

After the cleaning and transformation process of the data, the Processor examines each file and splits the transcripts into segments based on the specific markers that correlate with the participant of the interview, meaning that we only took into account the patient and not the clinician. The segments were then split into three different sizes to test which size was best for training our transformers. The sizes were as follows, 'short' for a median segment size of 5 words, 'medium' for a median segment size of 20 words, and 'large' for a median segment size of 50 words. For each segment, we provided a ground truth, the value 1 for Dementia and value 0 for Control patients. The ground truth was provided by the initial data. The processed segments and corresponding labels are then saved into csv files, balanced and shuffled for training and evaluation. The Processor Object is also able to function as a minimal digital signal processor for audio files, detecting and condensing silences, removing repeating words, and fillers to produce a cleaner transcription based on the CHA tokens. This entire process ensured that the data collected from DementiaBank, is cleaned, organised and ready to be utilised for further analysis and model training and evalution. The format provided by the CHA tokens is more interpretable and easier to understand, which will prove to be very useful in the later stages of our research, specifically in the explainability of the models. Providing a more interpretable format for the data and the phonological features derived from it, will allow for a more comprehensive understanding of the patient's cognitive state, which is crucial in the diagnosis and treatment of dementia. We will now delve deeper into the analysis of the data and the phonological features derived from it.

CHAT Symbol	CHA Token
[/]	[CHA REPETITION]
[//]	[CHA RETRACING]
(.)	[CHA SHORT PAUSE]
(..)	[CHA MEDIUM PAUSE]
(...)	[CHA LONG PAUSE]
+...	[CHA TRAILING OFF]
&+	[CHA PHONOLOGICAL FRAGMENT]
&*	[CHA INTERPOSED WORD]
&-	[CHA FILLER]
text(text)text	[CHA NON COMPLETION OF WORD]
&=belches	[CHA BELCHES]
&=hisses	[CHA HISSES]
&=grunts	[CHA GRUNTS]
&=whines	[CHA WHINES]
&=coughs	[CHA COUGHS]
&=hums	[CHA HUMS]
&=roars	[CHA ROARS]
&=whistles	[CHA WHISTLES]
&=cries	[CHA CRIES]
&=laughs	[CHA LAUGHS]
&=sneezes	[CHA SNEEZES]
&=whimpers	[CHA WHIMPERS]
&=gasps	[CHA GASPS]
&=moans	[CHA MOANS]
&=sighs	[CHA SIGHS]
&=yawns	[CHA YAWNS]
&=groans	[CHA GROANS]
&=mumbles	[CHA MUMBLES]
&=sings	[CHA SINGS]
&=yells	[CHA YELLS]
&=growls	[CHA GROWLS]
&=pants	[CHA PANTS]
&=squeals	[CHA SQUEALS]
&=vocalizes	[CHA VOCALIZES]
+..?	[CHA TRAILING OFF QUESTION]
+/.	[CHA INTERRUPTION]
+/?	[CHA INTERRUPTION OF QUESTION]
+//.	[CHA SELF-INTERRUPTION]
+//?	[CHA SELF-INTERRUPTED QUESTION]

Table 5.1: Transformation of CHAT symbols into CHA tokens

5.2.2 Analysis

Our approach to the analysis of the data was to extract phonological features from the transcripts, and see how they correlate with the patient's cognitive state.

CHA Token Distribution

In our analysis process, we first explored the frequencies of the tokens derived from the processing stage inside the segments. By iterating through the dataset, we counted the occurrences of the tokens within both groups of dementia and non dementia transcripts. To visualise our findings we generated a bar plot which displayed the counts of tokens inside both groups. Each group for color coded for better differentiation between the groups. The plot highlighted distinct differences in token frequencies between dementia and non-dementia transcripts, providing insights into the phonological markers associated with dementia. This plot underlines the potential of CHA tokens as valuable indicators of dementia in patient's speech patterns and requires for further investigation.

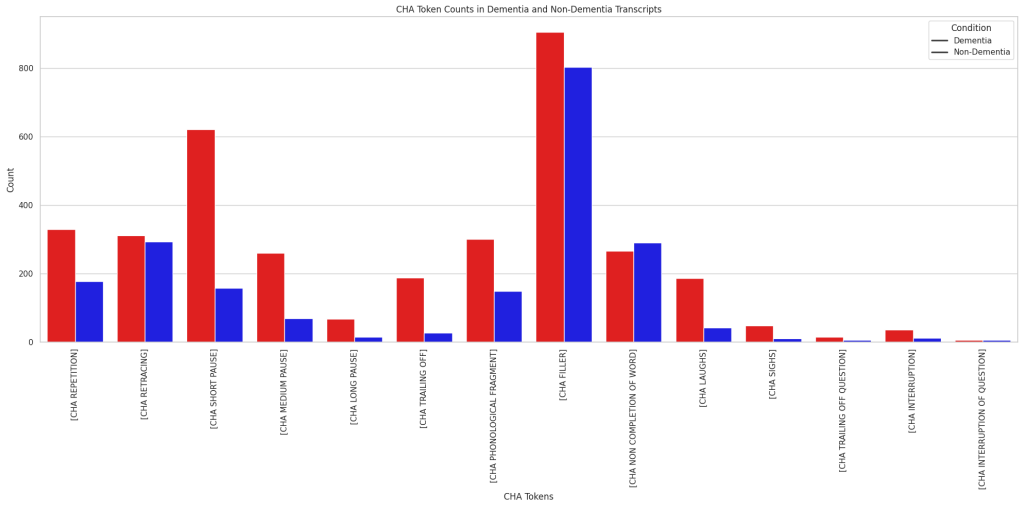
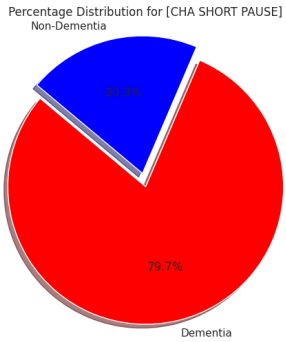


Figure 5.1: Frequencies of CHA tokens in Dementia and Non-Dementia transcripts

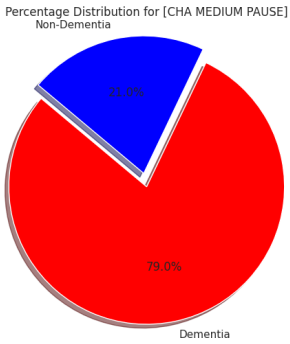
CHA Token Percentages

The next step of our analysis was to investigate the distribution of the CHA tokens in the segments further. This was done by calculating and visualising the percentage distribution of each token in the segments. First we compute the total number of occurrences of each token for both groups, and then calculate the percentage of dementia and non-dementia occurences for each token. For each token a pie chart is generated in order to visualise the distribution of the token in the segments. This visualization provides a clear, comparative view of how frequently each phonological feature appears in dementia versus non-dementia speech, highlighting potential diagnostic markers for dementia. Intuitively, the large the percentage of a token in the dementia group, the more likely it is for this token to be an indicator of dementia. The figure 5.3 includes examples of CHA token distributions that possibly indicate dementia. On the other hand, there were tokens that were similarly distributed in both groups, which may not be useful for diagnostic purposes. Interestingly enough, there was no distribution of CHA tokens that were more frequent in the non-dementia group. These findings suggest that CHA tokens could be a valuable tool for diagnosing dementia from spontaneous speech data. They also agree with the literature on the subject, which suggests that phonological features can be indicative of cognitive decline.

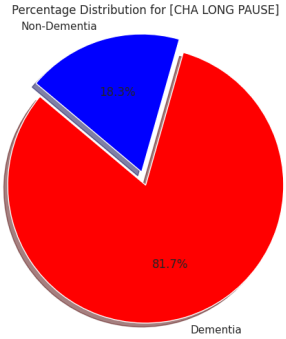
Below are some examples of CHA token distributions that possibly indicate dementia. These examples are mostly related to the patient pausing or trailing off during speech, which suggests cognitive decline, distortion in the patient's thought process, or difficulty in finding the right words. Other examples include the patient's inability to complete words, or the repetition of words, which could be indicative of memory loss or confusion. These tokens are more frequent in the dementia group, as indicated by the pie charts, and could be used as indicators of dementia in patients' speech.



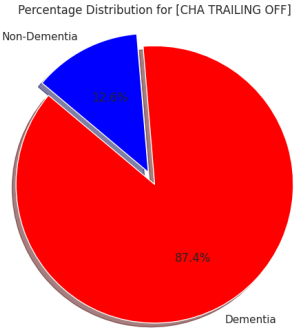
(a) [CHA SHORT PAUSE]



(b) [CHA MEDIUM PAUSE]



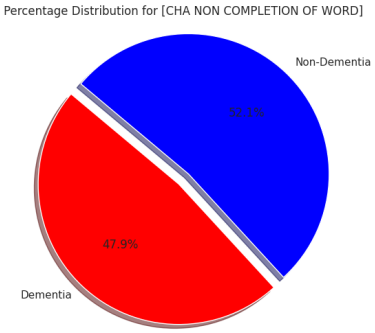
(c) [CHA LONG PAUSE]



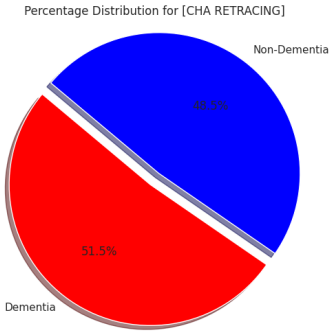
(d) [CHA TRAILING OFF]

Figure 5.2: Examples of CHA token Distributions that possibly indicate Dementia

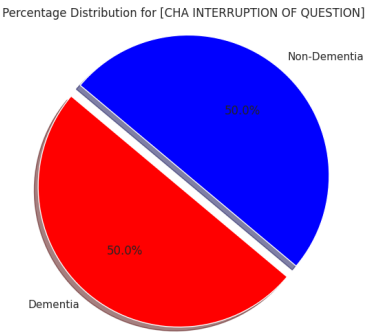
Below are some examples of CHA token distributions that are similarly distributed in both groups. These tokens provide little to no diagnostic value, but could be linked to dementia in other ways. Possibly in how they are used in speech, or how they relate to other tokens. Other ways could be the sequential use of tokens and speech patterns to indicate possible cognitive decline. This is a topic that requires further investigation and analysis, but will be more prevalent in the later stages of our research. Specifically, in the model training phase and the evaluation of the explainability of the models. Some of the tokens that are present below are in fact indicative of dementia, such as the non-completion of word token, but it is possible that they are used in a different context in the non-dementia group. Such context could be the patient's inability to complete a word due to a different reason, such as a speech impediment or a different cognitive issue, or even the patient's accent or dialect. These are all factors that need to be taken into account when analyzing the data, and could provide valuable insights into the diagnostic value of the CHA tokens.



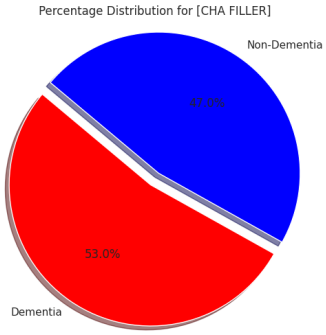
(a) [CHA NON COMPLETION OF WORD]



(b) [CHA RETRACING]



(c) [CHA INTERRUPTION OF QUESTION]



(d) [CHA FILLER]

Figure 5.3: Examples of CHA token Distributions that are similarly distributed in both groups

Co-occurences of CHA Tokens

The third step of our analysis was to investigate the coexistence of different CHA tokens in a single segment. By analysing the co-occurences of tokens we can associate different phonological features with each other, and possibly identify patterns that are indicative of dementia. To do so, we created a correlation matrix of the CHA tokens, which shows the pair wise correlation coefficients between different tokens across all trascripts. We then used the matrix to generate a heatmap, in order to visualise the co-occurences of the tokens in an intuitive way. The heatmap provides a clear view of the relationships between tokens. Positive correlations between two different tokens indicate that they are likely to co-occur in the same segment, while negative correlation indicates that they are unlikely to co-occur in the same segment. The larger the absolute value of the correlation coefficient, the stronger the relationship between a pair of tokens. This analysis provides a clear and comprehensive visualisation of the relationships between different phonological features, hightlighting potential patterns. By cross-referencing this information with the token distributions and frequencies we can provide valuable insights into the diagnostic value of a group of tokens appearing together in a segment.

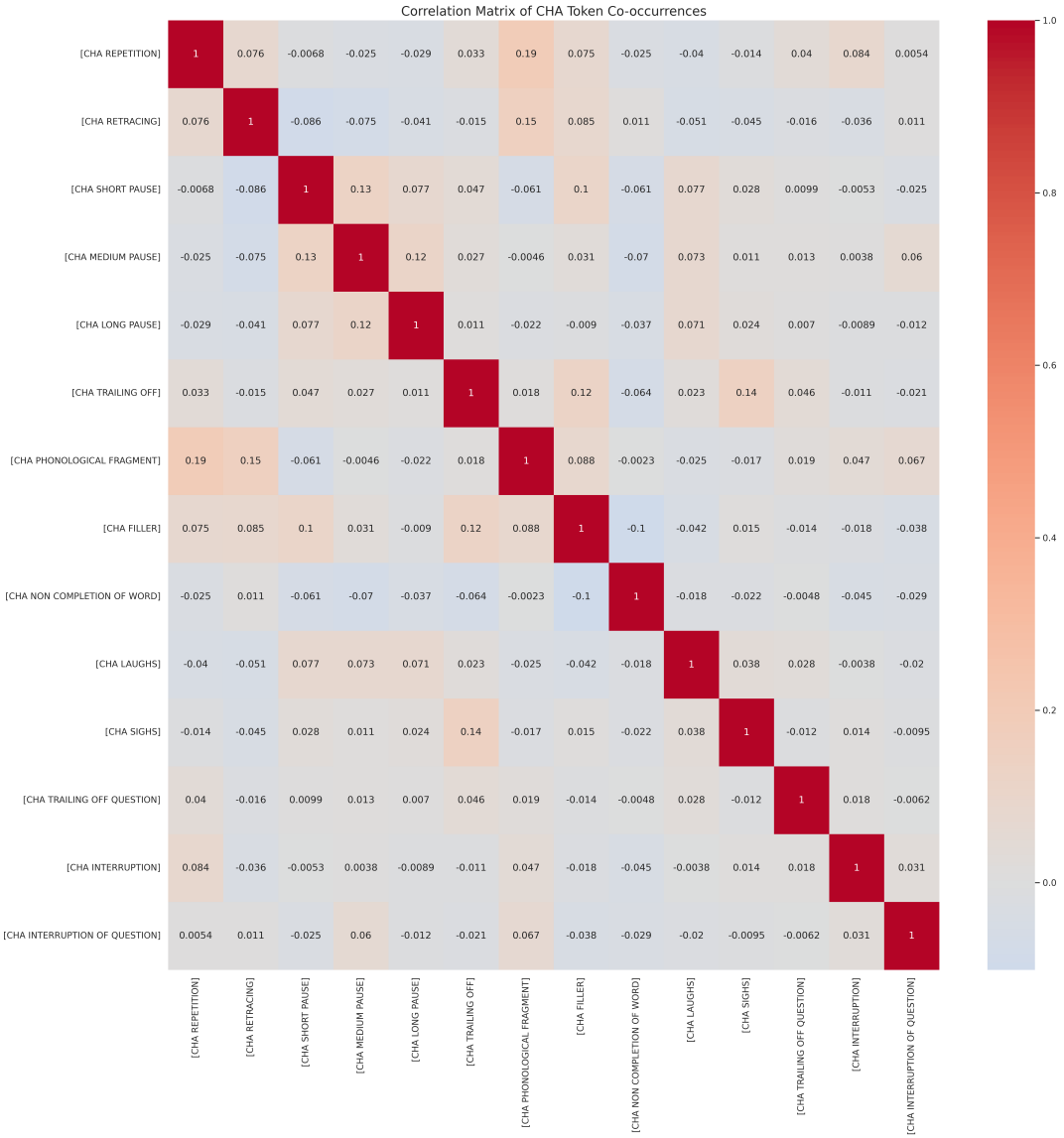


Figure 5.4: Co-occurences of CHA tokens in Dementia and Non-Dementia transcripts

In the figure 5.4 we can see the co-occurences of CHA tokens in the trascripts. The heatmap

provides strong evidence that certain tokens, much like the ones related to pausing and trailing off, are indicative of dementia, and also appear together in multiple transcripts as shown by their strong positive correlation. We can associate these tokens with cognitive decline, memory loss or confusion, and the fact that they appear together in the same segment provides further evidence of dementia. Other co-occurences of tokens can also be indicative of other cognitive issues or even speech impediments or mannerisms in speech. Another reason for specific co-occurences can be the patient's accent or dialect, ethnicity, or the manner in which the interview was conducted. Since the interviews were transcribed by manual labour, the clinician or professional conducting the interview and providing the trascripts could also insert their own biases or mannerisms into the transcripts, due to the way they trascribe and understand the patient's speech.

It is important to note that more analysis is required in order to determine the diagnostic value of the CHA tokens. Possibly, a multi-dimensional analysis, including features such as speech duration, pitch, and other phonological features could provide more insights into the diagnostic value of these tokens. Furthermore, different characteristics of the patient, such as age, ethnicity, educational level and other factors could also play a role in the diagnostic value of these tokens. These factors can provide a more comprehensive view of the patient's cognitive state and ability to communicate, thus making each token a multi-dimensional feature that can be used to further investigate the patient's cognitive state. That way, we could explicitly sway the token distribution in favor of dementia or non-dementia. This approach could be very useful in the research and development of AI models for diagnosing dementia from speech data, and could provide valuable insights into the patient's cognitive abilities, but it is an approach outside the scope of this thesis.

The data analysis phase of our research is crucial for several reasons. First, it provides evidence that the CHA tokens derived from the transcripts are valuable indicators of dementia in patients' speech and should be used in the development of AI models for diagnosing dementia. This approach should be investigated further in order to determine the diagnostic value of the tokens, and how they can be used to develop AI models for diagnosing dementia. Secondly, by providing these tokens into the training process of AI models, we can enhance the models' performance and accuracy, and provide more comprehensive explanations to the clinicians and professionals. This approach also provides a data-driven basis for further investigating the diagnostic value of phological features in patients' speech for exploring new diagnostic markers and improving the diagnosis and treatment of dementia. Lastly, this approach provides a more interpretable format for clinical applications and decision-making, which allows for the development of non-invasive, cost-effective, and efficient diagnostic tools for diagnosing dementia from spontaneous speech. Overall, the analysis of the CHA tokens is a pivotal step in this research, providing critical insights into the speech characteristics of individuals with dementia. It enhances the development of diagnostic models, supports clinical applications, and lays the groundwork for future research.

5.3 Model Development

5.3.1 Selection Criteria

In this section we will discuss the reasons for selecting Transformer models for our research, and how they can be used to develop explainable AI models for diagnosing dementia from spontaneous speech data.

Attention Mechanism

The selection of AI models for healthcare applications is a critical decision that requires careful consideration. We need to take into account the complexity and variability of the data at hand, the need for interpretability and explainability, the performance and accuracy of the models, and the ethical and legal implications of using AI in healthcare. In our research, we chose Transformer models for several reasons. The first reason, is the ability of Transformer models to handle sequential data and capture dependencies. This is done by using self-attention mechanisms, which allow a model to determine relative importance between different parts of the input sequence [158]. This will prove instrumental in generating explanations and evaluating how different words and tokens in the input can indicate dementia. Essentially, three different vectors are assigned to each word in the sequence, a query vector, a key vector, and a value vector. The attention score is then calculated, which is the dot product of the query and key vectors, divided by the square root of the dimension of the key vector. The attention score is then applied to the value vector, which is then used to calculate the output of the self-attention mechanism. Attention is used as a fundamental building block in Transformer models, and allows them to capture long-range dependencies in the data, which is crucial for understanding the context of the input sequence. A current limitation of the self-attention mechanism is that it can be computationally expensive, especially for longer sequences.

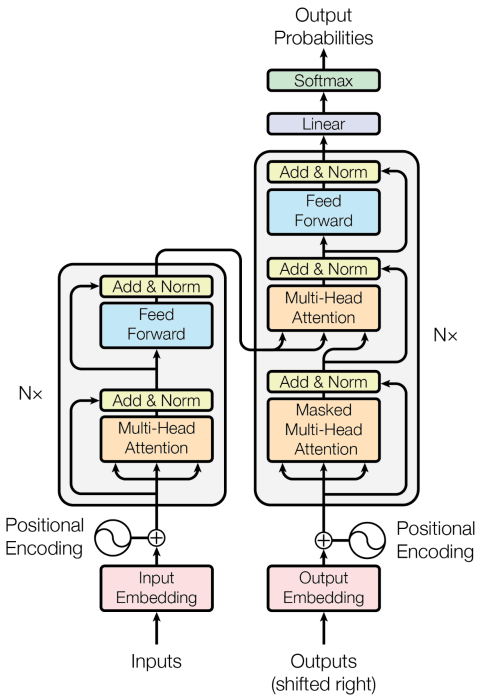


Figure 5.5: Transformer Architecture from "Attention is all you need"

Transfer Learning

The second reason is strongly related to the large number of pre-trained Transformer models available, which can be fine-tuned on specific tasks with relatively small amounts of data. This practice is called transfer learning, and is widely used in natural language processing tasks, where large pre-trained models are fine-tuned on specific tasks with smaller datasets. This is particularly useful in healthcare applications, especially in our use case, were data is scarce due to privacy concerns and the difficulty of the collection process. This issue of scarcity is common in healthcare applications, especially in fields were data collection requires patient consent and cooperation, such as dementia research. The collection process for these types of tasks can

prove to be challenging, time-consuming, and expensive, which is why transfer learning is a valuable tool for developing AI models for diagnosing dementia. This was a key factor in our decision to use Transformer models, as they can be fine-tuned on the data collected by the DementiaBank corpuses and used to develop our AI models for diagnosing dementia.

Interpretability and Explainability
The third reason is the ability to leverage the self-attention mechanism in transformer models in order to generate explanations. During the prediction process, we can inspect the self-attention scores of the encoder and decoder layers, in order to determine which parts of the input sequence are most important for the prediction. This allows us to generate explanations for the model's predictions, and provide insights into the decision-making process of the model. This is crucial for developing explainable AI models for diagnosing dementia, as it allows us to provide clinicians with evidence and reasoning for the model's predictions, and build trust and transparency in the model's decision-making process.

Support by the Literature
In addition to these reasons, the literature on the subject supports the use of Transformer models for diagnosing dementia and similar tasks as we have discussed in the literature review. Research has shown that these models can achieve state-of-the-art performance on natural language processing tasks, significantly outperforming traditional machine learning models and providing more accurate and reliable predictions. The literature also shows that the self-attention mechanism can assist in understanding the contextual relationships between different parts of the input sequence which is crucial in the task at hand. There is also strong evidence that Transformer models can be used to develop explainable AI models for diagnosing dementia, and provide clinicians with valuable insights into the decision-making process of the model. This evidence supports our decision to use Transformer models for our research, and provides a strong foundation for the development of AI models for diagnosing dementia.

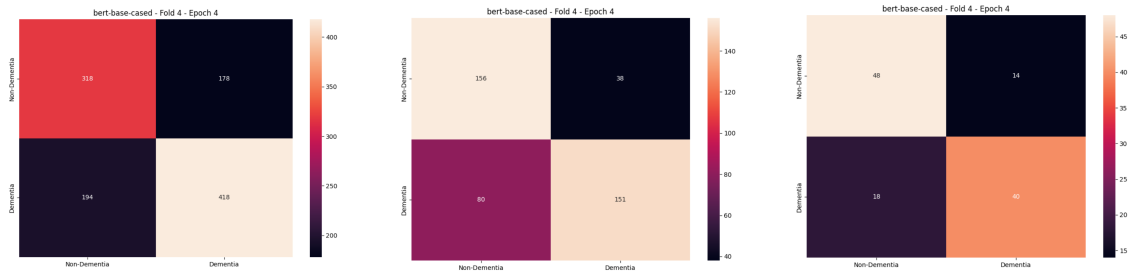
In summary, we concluded that Transformer models are the most suitable choice for our research, due to their ability to contextualise and capture dependencies in the trascripts, their support for transfer learning, their ability to be interpretable and explainable, and the strong evidence in the literature supporting their use in healthcare applications and in our specific use case. We will now discuss our proposed model architecture and the steps we took to develop and evaluate the model.

5.3.2 Model Architecture

In our research we propose an ensemble model, which combines different transformer models and classifiers in order to provide more accurate and reliable predictions. The ensemble model development is divided into two separate stages, the first stage being the fine-tuning of the transformer models on the data collected from DementiaBank, and the second stage being the training of multiple classifiers on the output of the transformer models. The final output of our ensemble model is a majority vote of the predictions made by the classifiers, which is then used to make the final prediction. We will now go into further detail about the architecture of the ensemble model and the steps taken to develop and evaluate it.

Fine-tuning Transformer Models
As we have already discussed, the first stage of our ensemble model development leverages the ability of transformer models to be fine-tuned on specific tasks with relatively small amounts

of data. In our research, we utilised the Hugging Face transformers library (huggingface.co) to select five different transformer models, namely, BERT, RoBERTa, DistilBERT, ClinicalBERT and BioBERT. After selection, we inserted the CHA tokens into each transformer model as special tokens, and the models were then fine-tuned on the data collected and processed from DementiaBank. The fine-tuning process involved training the different transformer models under 5-fold cross-validation and 5 epochs. The performance of the models was evaluated using different metrics, such as accuracy, precision, recall and F1-score. During the training process of the transformer models, all three segment sizes were used, 'short', 'medium', and 'large', in order to determine which size was best for training the models. We found that the 'medium' segment size was the most suitable for training the transformer models, as it provided the best performance and also allowed for a more comprehensive training of the attention mechanism. The 'short' size did not allow for the model to capture long-range dependencies in the data, which hindered the explainability of the model. The 'long' size did not provide adequate amounts of data for validation. Further research is in order on the subject of segment sizes and their impact on model performance but also on the explainability of the models. Larger segment sizes may be more suitable for capturing dependencies in the data, but that may add complexity and computational cost, which becomes a trade-off between performance and explainability. The figure 5.6 includes heatmaps of the different segment sizes for the BERT model, which show the performance of the model on each size. We also tested fine-tuning the models with different learning rates, batch sizes, and optimisers, in order to determine the best hyperparameters for training, without significantly affecting the performance of the models. Lastly we wanted to evaluate weather the models' performance was being hindered due to the addition of the CHA tokens as special tokens to each transformer which also showed no significant impact on the performance of the models, allowing us to proceed with our initial hypothesis and the induction of the CHA tokens in the overall architecture. We will further discuss the overall results of the fine-tuning process in the next chapter.



(a) Short segment (b) Medium segment (c) Long segment

Figure 5.6: Heatmaps of Short, Medium and Large segment sizes

Through the fine-tuning process we were able to obtain three different transformer models capable of making predictions on the segments with relatively high accuracy, precision, recall and F1-score. The models were not quite as accurate as the state-of-the-art models in the literature, but they provided a solid foundation for the development of the ensemble model and the explainability of the predictions. The next step was to train multiple classifiers on the output of the transformer models, in order to provide more accurate results.

Training Classifiers

The second stage of our ensemble model development involved training multiple classifiers and aggregating their predictions with a majority voting scheme which selected the most frequent prediction. The initial approach contained a single fold of the triad of transformer models,

which was then used to train a single classifier. We based our experiment on the paper of Julian Risch and Ralf Krestel [131] who used a similar approach to aggregate the predictions of multiple transformer models using bagging. We used a Bagging Regression classifier as the initial classifier, which was trained on on the concatenated output logits of each transformer model, and was then used to make predictions on a test set. The predictions showed very promising results, with a significant increase in all metrics, which was a strong indication of the potential of the ensemble model.

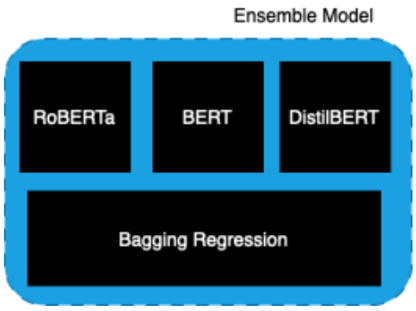


Figure 5.7: Single Ensemble Model Architecture

To extend our experiment, we trained multiple other classifiers such as Random Forest, Gradient Boosting, Support Vector Machines, Decision Tree Classifier, K-Nearest Neighbors, and Logistic Regression similarly to the bagging classifier. The results of the classifiers were then aggregated using a majority voting scheme, which selected the most frequent prediction. The final output of the ensemble model was the majority vote of the predictions made by the classifiers, and that final output represented the decision of our ensemble model. All classifiers were trained under grid search, in order to determine the best possible parameters for training. All classifiers showed state-of-the-art performance, with a significant increase in accuracy, precision, recall and F1-score compared to the transformer models.

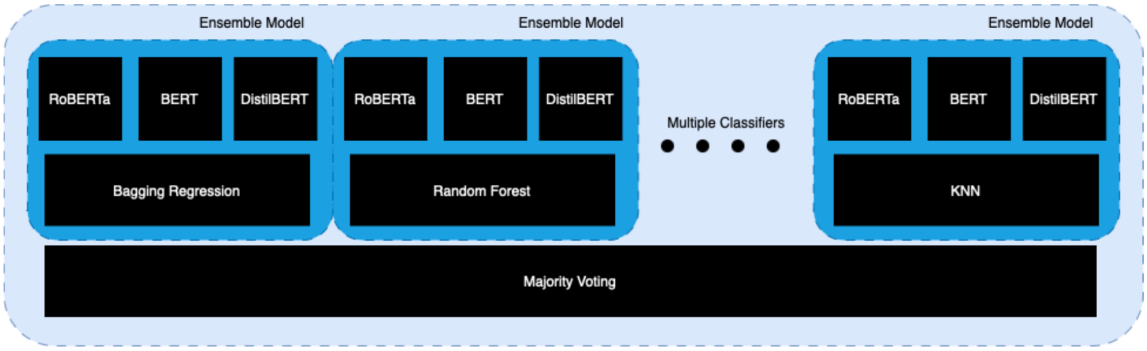


Figure 5.8: Ensemble Model Architecture

As we can see in the figure 5.8, the ensemble model architecture is a combination of multiple classifiers and transformer models, which are used to make predictions on the segments. The final output of the ensemble model is the majority vote of the predictions made by the classifiers, which is then used to make the final prediction. The ensemble model provides more accurate and reliable predictions compared to single transformer models or single classifiers.

5.4 Selection Criteria for explainable methods

In this section we will discuss the reasons for selecting the LIME explainable method for our research and subsequent user interface development. During the course of our research we evaluated multiple explainable methods capable of providing explanations for transformer models. These methods included LIME, Anchor and Transformers-Interpret. To evaluate the explainability of the methods we conducted a survey with clinicians and professionals experienced in treating dementia, who were asked to evaluate the visual explanations provided by each method and provide feedback on how informative and useful they were. Due to very low participation in the survey, we were unable to provide a comprehensive analysis of the results, but we were able to draw some conclusions from the feedback provided by the participants. The survey showed that LIME was the most preferable method out of all three.

5.4.1 Selection Criteria

LIME was selected for several reasons. We will shy away from stating its applicability and ease of use in incorporating it into transformer models, due to the fact that all three methods were equally capable of providing explanations for the transformer models and the code implementations were readily available and would not be a factor in the ease of use by professionals and clinicians. We will focus instead on the comparison of the visual explanations provided by each method, and how each method was able to highlight different parts of the input sequence and specifically the CHA tokens.

Visual Simplicity

The first reason for selecting LIME was the visual simplicity of the explanations provided. LIME uses simple and intuitive structures to provide insights into the decision-making process of the model, providing scores for each word or feature in the input sequence, and color coding them based on importance and relevance to dementia. Simplicity allows for non-technical users to understand the explanations and make informed decisions upon the explanation provided. Transformers-Interpret provided a similar interface, the feature importance scores were clear and the CHA tokens were highlighted in the explanations, but the interface, color-scheme and font selection were not as vibrant as LIME which made it less appealing to the participants. Anchor provided a more complicated interface and failed to tokenise the CHA tokens as well. Additionally, the textual representation of the explanations provided by Anchor contained UNK tokens, which made it difficult for the participants to understand the explanations. Lastly, a very important factor in the selection of LIME was the addition of a more informative scoring system, which provided the attribution scores used in the visualisation.

Feature Highlighting

The second reason is strongly dependent on the first. LIME, through its visualisation framework, provides the transcript back to the user, with the most important words and features highlighted in color. Dementia-related words are highlighted in red and orange shades, and non-dementia related words are highlighted in blue shades. This made the interpretation of the explanations much easier for the participants, and allowed for a more comprehensive understanding once the participants understood the color coding scheme and the CHA token representations. LIME was capable of tokenising the CHA tokens and highlighting them completely, which was a significant advantage over the other methods. Transformers-Interpret had a similar approach, though the highlighting was not nearly as clear as LIME, possibly due to the color selection or the way coloring was implemented through scoring. Anchor provided a complicated textual output that had no color coding or highlighting, using placeholder text and

cluttering its output, which made it difficult for the participants to understand the explanations.

Comparative Analysis
LIME, due to its local interpretability approach, was able to assess the importance of each word and CHA token in the input sequence relative to its context, meaning that for a single CHA token, LIME was able to provide insight into how the token swayed the model's prediction towards dementia or non-dementia. Depending on the input sequence, LIME would provide different explanations for the same token, which provided a more comprehensive view of the token's importance and relevance to the overall prediction. Transformers-Interpret provided insight on the CHA tokens, but was not as detailed as LIME, while Anchor, due to its example-based approach in providing explanations, requires a more focused and detailed analysis of its output.

Overall, LIME's flexibility, simplicity in visualisation and feature hightlighting capabilities made it the most preferable method out of the three and thus was selected for our ensemble explainer framework and user interface.

5.4.2 Providing Explainability

Our strategy for providing explainability consisted of creating an ensemble explainer framework, which combined the best performing explainable method from the survey. The ensemble explainer framework was developed in three stages, the first stage being, the generation of explanations from all three explainable methods, the second stage was to assess the quality of explanations through the afformentioned survey and lastly, the final stage was the aggregation of the explanations provided by LIME using a weighted average scheme. The final output of the ensemble explainer framework was a single aggregated explanation for the sequence which required an explanation. In this section we will discuss in detail, the steps taken to create the ensemble explainer framework.

Generation of Explanations
After the selection of transformer models and their fine-tuning, we proceeded to generate explanations for two different sequences, one for each group of dementia and non-dementia transcripts. The explainability methods utilised were LIME, Anchor and Transformers-Interpret, which were used to generate explanations out of the token representations provided by the three fine-tuned transformer models. For all the example explanations provided, a single sequence was used as input, which belonged to the dementia group. The sequence used is provided below.

youknow it [CHA FILLER]I [CHA REPETITION] I [CHA RETRACING] [CHA FILLER]excuse me but youknow I [CHA REPETITION] I

Figure 5.9: Example Sequence for Testing

We first generated explanations using LIME, which provided a comprehensive and intuitive visualisation of the explanations along with color coding and feature highlighting. These explanations also provided insights as to how much each word or token stimulated the model's prediction towards dementia or non-dementia. Various words and tokens, due to the token representations provided by the transformer models and the linear approximation of the model's decision function by LIME, were exponentially emphasized in the explanations. These phenomena were particularly evident in a variety of words which would not intuitively be associated with dementia, but were highlighted as important by the model. Such occurences could be

evidence of the model being biased towards certain words or tokens, due to their prevalence in the training data, or due to the way the model was making certain associations between words and dementia. The CHA tokens were also highlighted in the explanations, and were mostly in accordance with the token distributions and token analysis we conducted, although most of the CHA token scores were not as high as we expected possibly due to them being added to the tokenizers as special tokens and not being part of the vocabulary.

The LIME explanation for the BERT model is shown in the figure 5.10. The explanation shows that repetitons of words in this input sequence are indicative of dementia, and that retracing is not. The model also interprets the concatenated word "youknow", which should have probably been tokenised as a filler word but was not transcribed as such by the clinician, as strong evidence of dementia in this particular sequence. The figure 5.11 shows the LIME explanation for the RoBERTa model. The explanation here also highlights the importance of "youknow" in the sequence as a strong indicator of dementia, as well as repetitions of words, but defers from the BERT explanation in that filler words have a higher importance in showing non-dementia. Lastly, the explanation for the DistilBERT model, which is shown in the figure 5.12, highlights most of the tokens and words as strong indicators for dementia, especially the repetitions of words which is a significant difference from the other two explanations.

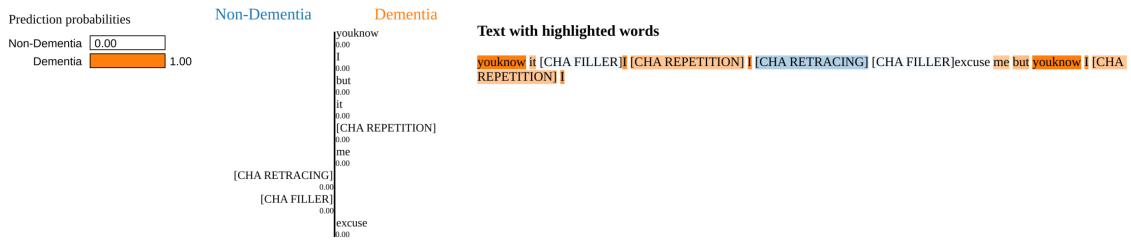


Figure 5.10: BERT LIME

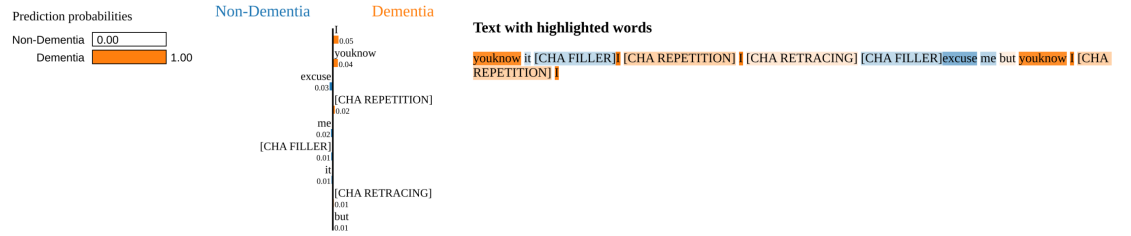


Figure 5.11: RoBERTa LIME

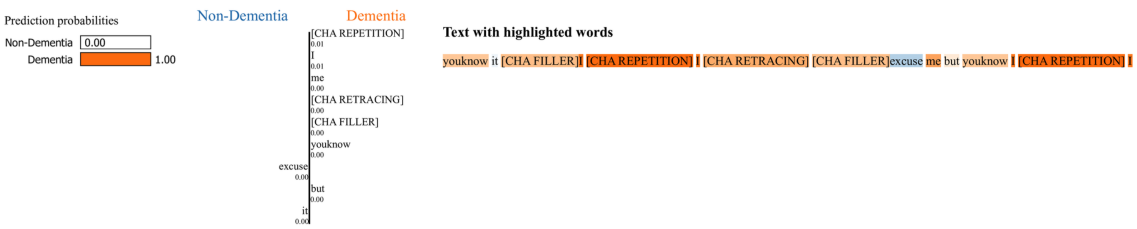


Figure 5.12: DistilBERT LIME

Quite interestingly, each model provided different reprsentations for the same tokens, which drove LIME to produce completely different explanations for a single sequence. Although the

explanations do provide insights into the model's decision-making process, they are not utterly reliable and should be further investigated.

The second candidate for generating explanations was Anchor, which is a rule-based explainable method that provides a non-linear approximation of the model's decision function, and is significantly more precise than LIME. Anchor's visualisation framework and examples provided for interpretation were not as intuitive as LIME's, and the explanations were not nearly as informative as the ones provided by LIME.

```
Model Predicts Dementia

Examples where anchor applies and model predicts Dementia:
youknow UNK [ CHA UNK UNK UNK REPETITION UNK I UNK UNK UNK UNK UNK UNK FILLER]excuse me but youknow I [ CHA REPETITION ] UNK
youknow UNK UNK UNK UNK [ CHA REPETITION ] I [ CHA UNK ] UNK CHA FILLER]excuse me but youknow I [ UNK UNK ] I
youknow it UNK UNK FILLER]I [ UNK REPETITION ] UNK UNK CHA RETRACING UNK [ CHA FILLER]excuse me but youknow UNK UNK CHA UNK ]
UNK it UNK CHA UNK [ UNK REPETITION ] I UNK CHA UNK ] [ CHA FILLER]excuse me but UNK I [ CHA UNK ] I
youknow it [ CHA UNK UNK UNK UNK UNK UNK CHA RETRACING UNK UNK CHA UNK me but youknow I [ UNK REPETITION UNK UNK
youknow UNK UNK CHA FILLER]I UNK CHA UNK ] UNK [ UNK UNK ] [ UNK UNK me UNK UNK I [ CHA UNK ] I
youknow UNK UNK CHA FILLER]I [ UNK UNK ] UNK UNK CHA RETRACING ] [ CHA UNK me UNK youknow I [ CHA REPETITION UNK UNK
UNK UNK UNK UNK FILLER]I UNK UNK UNK ] UNK UNK UNK UNK UNK UNK FILLER]excuse me but UNK I UNK UNK UNK ] UNK
UNK it [ CHA UNK [ CHA UNK ] I [ CHA UNK ] [ UNK UNK me UNK youknow I UNK CHA UNK UNK UNK
UNK it UNK CHA FILLER]I UNK UNK REPETITION ] I UNK UNK UNK ] UNK UNK UNK me but youknow UNK [ CHA REPETITION ] I

Examples where anchor applies and model predicts Non-Dementia:
```

Figure 5.13: BERT Anchor

```
Model Predicts Non-Dementia

Examples where anchor applies and model predicts Dementia:
youknow UNK UNK CHA FILLER]I [ UNK REPETITION UNK I [ CHA RETRACING ] [ UNK UNK me but youknow UNK UNK UNK REPETITION ] I
youknow UNK [ CHA UNK UNK CHA UNK ] UNK UNK CHA RETRACING ] UNK UNK FILLER]excuse UNK UNK youknow UNK [ UNK UNK ] UNK
youknow UNK UNK CHA FILLER]I [ CHA REPETITION ] UNK [ UNK UNK UNK UNK CHA FILLER]excuse me but youknow UNK UNK UNK REPETITION
     UNK UNK
youknow UNK [ UNK FILLER]I [ CHA REPETITION ] I UNK UNK UNK ] [ CHA UNK UNK but youknow I UNK CHA REPETITION UNK UNK
UNK UNK [ CHA UNK UNK CHA REPETITION UNK UNK [ CHA UNK ] [ UNK FILLER]excuse UNK UNK youknow I UNK CHA REPETITION UNK I
UNK it UNK UNK FILLER]I UNK CHA UNK ] I [ UNK RETRACING ] UNK UNK FILLER]excuse UNK UNK youknow I UNK CHA UNK ] I
youknow it UNK CHA UNK UNK CHA UNK ] UNK UNK CHA UNK ] [ UNK UNK me UNK youknow UNK [ CHA REPETITION ] UNK
youknow UNK [ UNK FILLER]I UNK CHA UNK UNK UNK UNK UNK UNK UNK [ CHA FILLER]excuse UNK but youknow UNK [ CHA REPETITION ] I
youknow UNK UNK CHA FILLER]I UNK CHA REPETITION ] I UNK CHA UNK UNK [ CHA FILLER]excuse me but youknow I [ CHA REPETITION ]
     UNK
UNK it [ UNK FILLER]I [ CHA REPETITION ] I [ CHA RETRACING UNK UNK CHA FILLER]excuse UNK UNK youknow UNK [ CHA UNK ] UNK

Examples where anchor applies and model predicts Non-Dementia:
```

Figure 5.14: RoBERTa Anchor

```
Model Predicts Dementia

Examples where anchor applies and model predicts Dementia:
UNK UNK [ CHA UNK [ UNK UNK UNK UNK [ UNK UNK ] UNK UNK UNK UNK but youknow I [ CHA UNK UNK I
UNK it [ CHA FILLER]I UNK UNK REPETITION UNK UNK UNK UNK UNK ] [ CHA UNK me but UNK UNK [ CHA UNK UNK UNK
youknow UNK UNK CHA UNK UNK CHA UNK ] UNK [ CHA UNK ] UNK CHA FILLER]excuse UNK UNK UNK I [ CHA REPETITION UNK UNK
youknow UNK [ UNK UNK UNK UNK REPETITION ] I [ CHA UNK ] UNK UNK UNK UNK UNK youknow I UNK CHA REPETITION UNK I
UNK it UNK UNK FILLER]I UNK CHA UNK ] UNK [ UNK RETRACING ] [ CHA UNK me but youknow UNK UNK CHA UNK UNK UNK
UNK UNK UNK UNK FILLER]I UNK UNK REPETITION UNK UNK UNK CHA UNK ] [ CHA FILLER]excuse UNK but UNK I UNK CHA REPETITION UNK I
youknow UNK UNK CHA FILLER]I UNK CHA UNK UNK UNK UNK UNK RETRACING ] [ UNK UNK UNK but youknow UNK UNK CHA REPETITION UNK UNK
youknow it UNK UNK FILLER]I [ CHA REPETITION UNK UNK [ CHA RETRACING ] UNK UNK FILLER]excuse UNK but youknow I UNK CHA UNK ] I
youknow UNK UNK CHA UNK UNK CHA UNK UNK I UNK CHA RETRACING UNK UNK CHA UNK me UNK youknow I UNK CHA REPETITION ] I
UNK UNK [ UNK UNK [ UNK UNK UNK UNK UNK CHA RETRACING ] UNK CHA FILLER]excuse me UNK youknow I UNK CHA REPETITION ] UNK

Examples where anchor applies and model predicts Non-Dementia:
```

Figure 5.15: DistilBERT Anchor

The Anchor explanation for the BERT model is shown in the figure 5.13. The explanation essentially provides multiple examples of subsequences that are indicative of the prediction made by the model. The subsequences contain different words and tokens that reside in the original input sequence and provide evidence of the model's prediction, which in this case is dementia. The Anchor explanation contains multiple examples, but it is apparent that the examples are visualised in a cluttered manner and require focusing on each example separately

word by word to fully comprehend the explanation. Anchor also uses UNK tokens as placeholders for words that are not part of a specific example in the explanation, which makes it difficult to understand the explanation. Another significant issue in the explanations provided by Anchor, is Anchor's inability to properly tokenise the CHA tokens, due to missing an internal tokeniser for special tokens that do not exist in the vocabulary. In multiple examples provided by Anchor, the CHA tokens were not tokenised properly, which concluded in them being split into multiple different tokens, and thus being misinterpreted as UNK tokens. In the figure 5.14 we can see the Anchor explanation for the RoBERa model, which at first glance seems to be similar to the BERT explanation, but upon closer inspection we can see disparities in the examples provided. The DistilBERT explanation, shown in the figure 5.15, suffers from the same issues as its two predecessors, and provides examples that are similar to the other two models. It is important to note that the Anchor explanations, despite their visual complexity, greatly favoured the CHA tokens, even though they were not tokenised properly, and rarely provided examples that did not contain them.

The third and final method we employed for generating explanations was Transformers-Interpret, a newly developed explainable open-source framework for creating explanations for transformer models. Transformers-Interpret calculates word and token attributions using the integrated gradients method, and provides visual explanations similar to the ones provided by LIME. Transformers-Interpret was able to provide clear and informative explanations, but the color-scheme and font selection were not as vibrant as LIME, which hindered the appeal of the explanations. The repetitions of words, for the BERT explanation in figure 5.16, were highlighted as indicative of non-dementia, while the RoBERTa explanation in figure 5.17 was indifferent. The DistilBERT explanation in figure 5.18 highlighted the repetitions as indicative of dementia, being the only explanation out of the three to focus on the CHA tokens. All three explanations showed larger inconsistencies in terms of evaluating importance of words and tokens compared to LIME. Transformers-Interpret failed, on some occasions, to properly tokenise words and tokens, which resulted in some placeholder text, without that being a notable issue in the explanations. Overall the explanations provided by Transformers-Interpret were informative, but the visuals and minor inconsistencies in feature importance made them less preferable than LIME for our intended purpose.

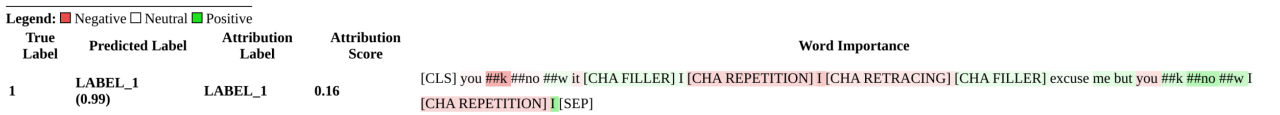


Figure 5.16: BERT Transformers-Interpret

Figure 5.17: RoBERTa Transformers-Interpret

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_1 (1.00)	LABEL_1	3.33	[CLS] you ##k ##no ##w it [CHA FILLER] I [CHA REPETITION] I [CHA RETRACING] [CHA FILLER] excuse me but you ##k ##no ##w I [CHA REPETITION] I [SEP]

Figure 5.18: DistilBERT Transformers-Interpret

Ensemble Explainer Framework

After generating explanations and assessing their quality and utility through the survey, we proceeded to aggregate the explanations provided by LIME using a weighted average scheme. The values of the weights were determined based on the best performing transformer model, in terms of token representations and overall performance. The final output of the ensemble was the aggregated explanation, using the interface provided by LIME.

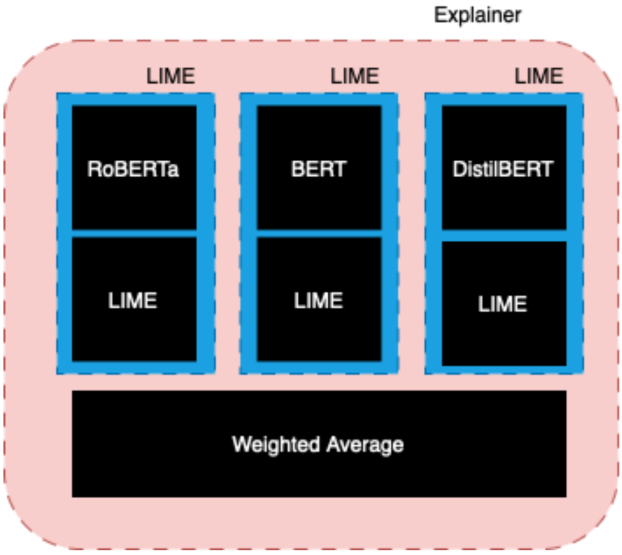


Figure 5.19: Ensemble Explainer Architecture

The ensemble explainer framework was able to smooth out the explanation by combining the individual explanations. The ensemble explainer framework explanation for the sequence is shown in the figure 5.20. A more detailed investigation of the best possible weights for the ensemble explainer framework is required, in order to determine the best possible weights, so that the CHA tokens attribution scores can be distributed effectively and provide a more comprehensive explanation in line with the literature concerning phonological features as evidence of dementia and the token analysis we conducted. Again, we see that even the ensemble explainer places a strong emphasis on the "youknow" token, which is due to the fact that all three individual explanations hightlighted it as a strong indicator of dementia which is an attribution made due to the self-attention mechanism in the training of the transformer models. A possibility is that this token is apparent in the group of dementia transcripts which inserts a bias in the model's decision-making process and the explanations provided. Again, we see that further research is required in order to determine the diagnostic value of each CHA token and how it can be used to develop interpretable explanations. An approach that could be used is for each CHA token, to produce a pair of new tokens, one for each group of dementia and non-dementia transcripts through the multi-dimensional analysis we discussed earlier. We will refer to these tokens as MVCHA tokens for the purpose of this discussion. That would allow for more reli-

able sequence generation during the data processing stage, where each CHAT trascript symbol would be replaced by either of the two new tokens, depending on the group it belongs to. That way, we could explicitly sway the token distribution in favor of dementia or non-dementia and provide more reliable explanations. Another important aspect of creating such an ensemble is diversity in the models used, an ensemble of larger variety of models could provide variance in the attributions and eliminate possible biases by single models. Lastly, an approach were all words are eliminated from the original transcripts, and the MVCHA tokens are significantly augmented in the sequence through speech and behavioural analysis, could prove promising in the development of comprehensive explanations.

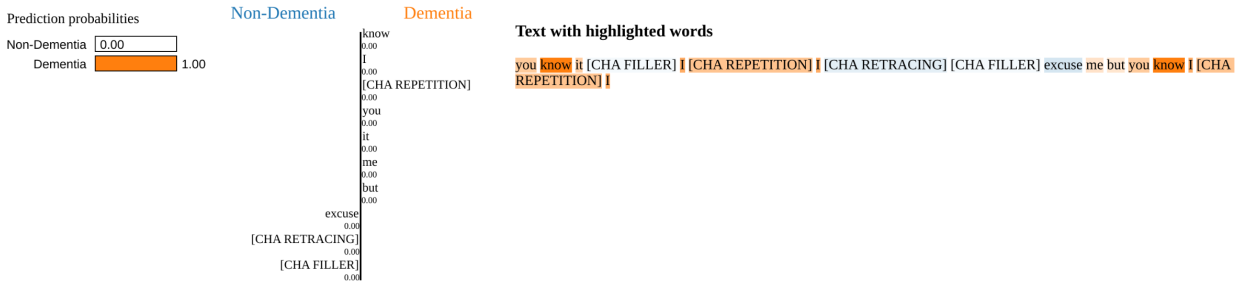


Figure 5.20: Ensemble Explanation

This approach can also be utilised with other explainable methods, such as Transformers-Interpret, where the approach of setting attributions and scores on words and tokens is quite similar to LIME. This way, we can provide more robust explanations for the transformer models. The ensemble explainer framework can possibly be extended to include multiple explainable methods, and provide a more comprehensive and reliable explanation for the model's predictions.

Evaluation metrics for explainability

We largely assessed the quality of the explanations using qualitative metrics related to the survey such as simplicity, and human evaluation. We also evaluated the explanations on account of their degree of proper usage of the CHA tokens and their visual appeal. In addition to the survey, we also evaluated the explanations based on fidelity and time of generation. Fidelity is a measure of how well the explanation approximates the model's decision function. We tested the fidelity of the explanations provided by LIME, Anchor and Transformers-Interpret by comparing the model's predictions to the predictions proposed by the explanations. All three methods showed similar fidelity scores. The results of the evaluation scores are presented in greater detail in the next chapter.

Overall, we have assessed that the explanations require further investigation and research, especially by incorporating the MVCHA tokens. The evaluation process has yielded promising results, and shows a comparative analysis of the explainability methods used in the research. The ensemble explainer framework also shows that an ensemble approach with greater diversity in the transformer models could in fact be a route to examine further to develop more reliable and comprehensive explanations.

Computational Resources

In order to expedite the computational processes involved in our experiments, we utilized GPU acceleration available on cloud computing services. The platform used provided access

to NVIDIA A100 Tensor Core GPUs, which significantly reduced the execution time for our experiments. The use of GPU acceleration was critical in handling the intensive computations required for fine-tuning the transformer models and also generating the explanations. We also utilized increased memory capacity which was essential in handling the explainer's computations of up to 83 GB of RAM. The experiments were conducted in a Python environment with relevant libraries such as TensorFlow and PyTorch optimized for GPU performance. We also tested the models on a local system with an ARM-based-system-on-chip M1 processor with 16 GB of RAM. The local system failed to generate the LIME explanations due to high memory consumption, which was a significant limitation in the development of the ensemble explainer framework. The cloud computing platform provided the necessary resources to handle the memory-intensive computations required for generating the explanations.

5.5 Ethical considerations

One of the foremost ethical considerations in our research is the privacy and confidentiality of the data used. The data provided by DementiaBank is highly sensitive and contains personal information about individuals. Handling of data such as this requires strict adherence to data protection laws and regulations. Researchers must ensure that such data is secured and anonymised to prevent any potential breaches of privacy. Another important ethical consideration, is the bias and fairness concern. The training of AI models on data that is biased or unfair can lead to biased and unfair outcomes for patients. We saw an example of bias in the LIME explanations provided, where certain words and tokens were highlighted as strong indications of dementia, even though they were not intuitively associated with the disease. Such biases can be introduced during the training process of the models, due to the way the data is collected and processed, or even during the transcribing process of the patients' speech. We know that speech patterns can vary drastically across different demographics, and the models should be evaluated to ensure they do not disproportionately affect any particular group. Transparency and accountability is another aspect to take into account when developing AI models for healthcare applications. Lack of transparency can lead to misuse or over-reliance on AI systems, potentially resulting in harmful consequences for patients. Strict guidelines and regulations also need to set in order to ensure accountability in the use of AI models for diagnostic purposes. Each participant in the decision making process should have a clearly defined role and responsibility so that accountability can be established. It is also essential that these tools are used in a supplementary manner, and not as a replacement for human judgement and expertise. Finally, the broader ethical use of AI in healthcare must be considered. AI tools should be used to enhance patient care and outcomes and should be implemented in a way that is consistent with ethical principles, including beneficence (doing good), non-maleficence (avoiding harm), autonomy (respecting patient choices), and justice (ensuring fairness and equality).

Chapter 6

The COMFORTAGE Project

This research was conducted as part of the COMFORTAGE project, formally known as "Prediction, Monitoring and Personalized Recommendations for Prevention and Relief of Dementia and Frailty", which aims to develop personalized and adaptive solutions for dementia and frailty prevention and management. This effort aims to establish a pan-European framework for prevention and intervention in dementia and frailty. It bolsters a multidisciplinary approach, combining expertise from the fields of medicine, social sciences, humanities and technology.

6.1 Dementia and Frailty

Dementia is defined as the loss of cognitive function to such an extent has it interferes with a person's daily life and activities. Dementia ranges in severity from the mildest stage, where the person is starting to notice mild cognitive decline, to the most severe, where the person must depend on others for basic activities of daily living, including eating, dressing, and bathing. Dementia affects memory, thinking, language, judgment, mood and behavior. It is a progressive disease that worsens over time and affects millions of people worldwide. It is not a normal part of the aging process of an individual. Signs of dementia become apparent when a person's healthy brain cells stop functioning, which results in different symptoms. Symptoms of dementia include:

- Memory loss, poor judgment and general confusion
- Difficulty in speaking, understanding, reading and writing
- Difficulty formulating thoughts, repeating oneself, inability to find the right words
- Difficulty in performing routine tasks, loss of interest in previously enjoyed activities, mood swings
- Hallucinations, delusions, paranoia, agitation, aggression, loss of empathy, apathy
- Loss of balance and modility issues

There are different types of dementia, with Alzheimer's disease being the most common. Other types of dementia include frontotemporal dementia, lewy body dementia and vascular dementia. There are several causes that would lead to dementia, such as damage or loss of brain cells, changes to the brain's structure and different types of proteins building up in the brain.

The process of diagnosing dementia involves ruling out other conditions that may cause congititive decline, such as vitamin deficiencies, thyroid problems, depression and hormonal

imbalances. The next step is to evaluate the patient's medical and family history and proceed with a physical and neurological examination. These examinations include non invasive evaluations such congitive and neuropsychological tests and phychiatric evaluations. The doctor may also order brain scans, blood tests, genetic tests and cerebrospinal fluid tests to assess the patient's condition. Early detection of dementia is important as it may, in some cases, be treated. However, there is no cure for dementia, but there are treatments that can help manage the symptoms and slow down the progression of the disease.

Frailty is a condition defined as a clinically recognizable state of increased vulnerability resulting from aging-associated decline in reserve and function across multiple physiologic systems such that the ability to cope with everyday or acute stressors is compromised. Frailty is a common condition in older adults and is associated with increased risk of adverse health outcomes, including disability, hospitalization, and mortality. Criteria for frailty include unintentinal weight loss, exhaustion, low grip strength and low general physical activity. Leading causes of frailty include aging, genetics, lifestyle, and environmental factors.

Treatment for frailty is highly individualized, depending on the patient's needs and life expectancy. Overall, the goal in treating frailty is improving the patient's quality of life and preventing further decline as much as possible. Treatment includes physical activity such as walking and light strength training, proper nutrition and hydration, and prioritising mental activity, mental health and social interaction.

6.2 Problem Objectives and Outcomes

Early detection and timely diagnosis of dementia and frailty are a crucial step in managing these conditions and mitigating their impact on the patient's quality of life. Different factors hinder the ability of healthcare institutions and professionals to provide personalised treatment and care for patients, and instead rely on a generalised approach, which may not consider the individual's needs and circumstances. The COMFORTAGE project aims to develop a personalized and adaptive solution for dementia and frailty prevention and management. In order to achieve this, the project will develop a pan-European framework for community-based and people-centric prevention, monitoring and progression managing solutions for dementia and frailty. COMFORTAGE's mission is to develop a Virtualised AI-Based Healthcare Platform (VHP), first of its kind, to centralise AI resources for risk assessment, early diagnosis, and personalized decision-making. COMFORTAGE brings together interdisciplinary experts from across Europe's leading research institutions, universities and companies. This collaboration aims to improve clinical outcomes, support clinicians in their decision-making processes, and ultimately enhance the quality of life for patients with dementia. The framework proposed by COMFORTAGE incorporates AI and big data innovations, along with domain specific expertise and IoT technologies to provide highly coordinated, personalized and proactive patient care through the VHP. This effort aims to improve the quality of life in dementia and frailty patients, reduce the burden on healthcare systems, and provide a more efficient and effective approach to managing these conditions.

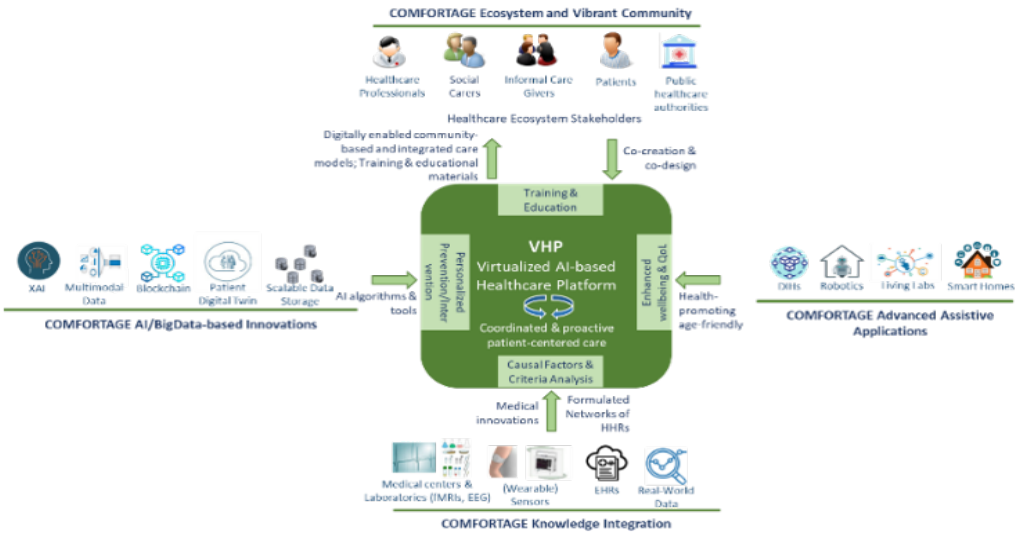


Figure 6.1: COMFORTAGE Framework

In our work, we aim to develop explainable AI systems that are capable of facilitating the efforts of the COMFORTAGE project, by providing an accurate and reliable means of predicting dementia through spontaneous speech. Our methodology is in line with the project's objectives and outcomes, as well as the overall mission. This research provides a framework for non-invasive, cost-effective and highly scalable solutions for dementia early detection and monitorin. We also aim to provide a clear path into future research directions that can be taken to further improve model performance and explainability.

6.3 DEMET: Our Contribution to the COMFORTAGE Project

Our contribution to the COMFORTAGE project is the development of DEMET, or Dementia Explainable Transformer, an explainable, AI-driven cognitive assessment agent for dementia detection through spontaneous speech. DEMET utilises state-of-the-art deep learning models and explainability techniques to provide an accurate prediction along with detailed insights into the model's decision-making process. DEMET's architecture is designed to be highly scalable and efficient, along with allowing for easy integration into the COMFORTAGE project. Our work is largely based upon the utilisation of phonogical features for explainability purposes and the implementation of ensemble learning techniques to improve model performance. We also propose an ensemble explainer framework that, with further investigation and development, can provide a more comprehensive understanding of the model's decision-making process and improve the overall explainability of the model. The ensemble explainer framework is designed to take into consideration the attributions of tokens provided by different trasnformer models in order to interpret how different phonological features in the patient's speech contrisolution the model's decision. DEMET provides both server and client-side implementations. The server-side implementation runs the model and explainer, along with an API that can easily be accessed by the client-side implementation. The client-side has a user-friendly and easy-to-use interface that allows for textual and audio input, and provides the user with the model's prediction and explanation. It can be used through the web application and the CLI tool application.

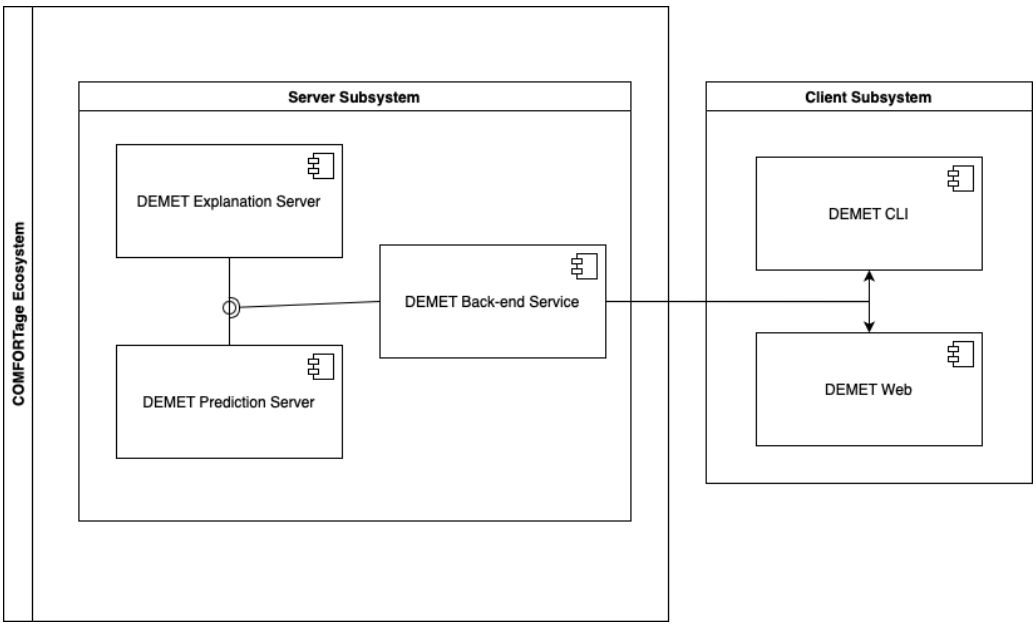


Figure 6.2: DEMET Application Architecture

DEMET allows for both textual and audio input. The Processor Object inside DEMET is responsible for processing the input into an appropriate format that can be used for prediction and explanation. The Processor Object is also responsible for extracting phonological features from the input, which are then used by the ensemble explainer to provide an explanation for the model's prediction. The DEMET model is trained on a processed version of the DementiaBank dataset, which contains trasncriptions of spontaneous speech from patients with dementia and healthy controls. The ensemble architecture of the predictor inside DEMET is based on three transformer models, namely BERT, RoBERTa and DistilBERT. These models were fine-tuned on the downstream task of dementia detection, and there outputs were combined into a concatenated feature vector. This vector was then fed into different classifiers. The classifiers used in the ensemble predictor are Random Forest, Gradient Boosting, Support Vector Machine, Logistic Regression, Decision Tree Classifier, K-Nearest Neighbors and Bagging Regressor. Ultimately, the final prediction of the ensemble predictor is the majority vote of the predictions provided by the classifiers. The explainer inside DEMET is based on LIME, which is a model-agnostic explainer that provides local explanations for the model's predictions. DEMET uses all three transformer models attributions to provide three different explanations for the model's prediction which are then aggregated into a single explanation. The aggregated explanation is the weighted average of the attributions of tokens provided by the individual explanations. The weights are determined by different factors, such as transformer model performance and qualitative metrics of the explanations provided by the different attributions of the transformer models. DEMET's performance is evaluated on a subset of DementiaBank manages an accuracy of 0.96%. The model's explainability is evaluated through the use of different qualitative and quantitative metrics, such as fidelity, simplicity, human evaluation and phonogical feature importance to name a few. The results of the prediction along with the explanation for the prediction made are provided to the user through the client-side implementation of DEMET in a user-friendly and easy-to-understand manner. Our explanation uses the scores provided by the ensemble explainer to color code the tokens in the input text, and highlight the most important phonological features and words that contributed to the model's prediction. The color coding is based on the same color scheme used in the LIME explainer, which is orange and red shades for dementia indicative tokens and blue shades for healthy indicative tokens. DEMET's next steps include further investigation into the ensemble predictor and ensemble explainer framework to improve diversity of transformer models and classifiers, and further

improvement into the phonological feature extraction scheme that was used in this research. We hope to also gamify the user experience of DEMET further, by providing the user with a more interactive and engaging experience and providing different cognitive assessment options and tests that can be used to further evaluate the patient's condition. The DEMET model architecture is provided in the figure below.

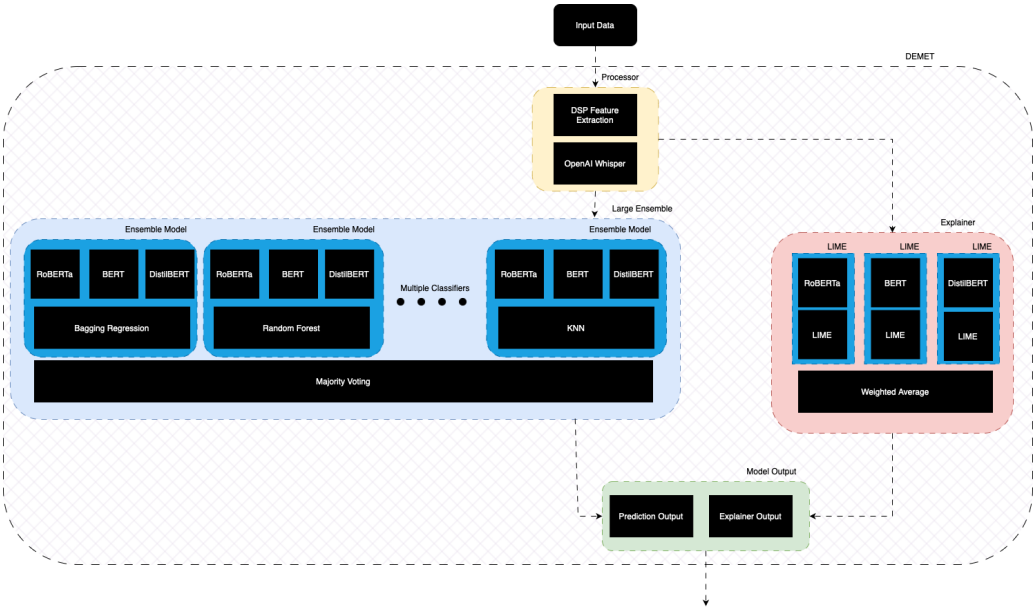


Figure 6.3: DEMET Model Architecture

The user can input text or audio along with the corresponding flags and metadata into the client-side application through the CLI tool or Web application to to receive the corresponding prediction and explanation as DEMET's output. The code implementations along with the documentation for DEMET can be found in the following link. These client-side applications are designed as a proof of concept that can be further developed and integrated into the COMFORTAGE project and are by no means the final product, or intended for clinical use as of the time of writing this thesis. Significantly more research and development is required to ensure that DEMET is a reliable and accurate tool for dementia detection through spontaneous speech, and the application of DEMET is subject to further validation and testing in order to ensure that it meets the necessary standards for clinical use and user satisfaction.

Through the development of DEMET, we hope to assist the COMFORTAGE project in achieving its objectives and outcomes, and provide a valuable contribution to the field of Explaianable AI applications for the improvement of quality of life in dementia patients. We also hope to provide a path for further research and development into the subject of dementia detection through spontaneous speech and urge the scientific community to further investigate the potential of AI-driven cognitive assessment agents for dementia detection and monitoring.

6.4 Algorithms and Models: Implementation and Results

In this section we present the results of our work on DEMET, an explainable, AI-driven cognitive assessment agent for dementia detection through spontaneous speech. We evaluated all

stages of the DEMET pipeline, including the ensemble predictor and ensemble explainer framework, and their subsequent components such as the transformer models and classifiers, as well as individual explainable methods used for the ensemble explainer development. All evaluations were conducted on a processed version of the DementiaBank dataset, were phonological features provided by the clinicians were transformed into a more interpretable format called CHA tokens. These tokens allowed for the generation of more interpretable explanations on the model's predictions.

6.4.1 Model Evaluation

Our metrics for evaluating model performace include accuracy, precision, recall and F1-score.

- **Accuracy** is the ratio of correctly predicted observations to the total observations. It is a measure of the model's ability to make correct predictions.

$$Accuracy = \frac{CorrectPredictions}{TotalPredictions}$$

- **Precision** is the ratio of correctly predicted positive observations to the total predicted positive observations. It is a measure of the model's ability to correctly predict positive observations.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall** is the ratio of correctly predicted positive observations to the all observations in actual class. It is a measure of the model's ability to find all the positive observations.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-score** is the weighted average of Precision and Recall. It is a measure of the model's accuracy on a particular dataset.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The initial stage of the DEMET pipeline was to find the best suitable format for the input sequences that would be fed into the transformer models. We tested three different sizes of the input sequences, 'short' for 5 words per sequence, 'medium' for 20 words per sequence, and 'long' for 50 words per sequence. We evaluated the transformer models on the different input sequence sizes to determine the best suitable format for the input sequences. The evaluation was done on a non balanced dataset under 5-fold cross validation, providing 20% of the data for testing. The evaluation results for all input sequence sizes are as follows:

Model	Accuracy	Precision	Recall	F1-score
BERT	0.66	0.70	0.68	0.69
RoBERTa	0.68	0.73	0.69	0.71
DistilBERT	0.64	0.64	0.84	0.72

Table 6.1: Evaluation Results for 'short' Input Sequence Size

Model	Accuracy	Precision	Recall	F1-score
BERT	0.74	0.79	0.76	0.77
RoBERTa	0.73	0.74	0.75	0.75
DistilBERT	0.69	0.73	0.80	0.76

Table 6.2: Evaluation Results for 'medium' Input Sequence Size

Model	Accuracy	Precision	Recall	F1-score
BERT	0.73	0.74	0.69	0.71
RoBERTa	0.78	0.80	0.77	0.79
DistilBERT	0.78	0.96	0.68	0.80

Table 6.3: Evaluation Results for 'long' Input Sequence Size

The results showed a clear improvement in model performance when using the 'medium' and 'long' input sequence sizes. We argue that the 'medium' input sequences provided an adequate amount of context for the attention-mechanism to work as well as enough samples for the model to be tested on. The 'short' sequences did not allow for the attention mechanism to work properly, and the 'long' sequences reduced the dataset size significantly, which after balancing left us with a very small amount of samples to train and test on. The 'medium' input sequence size was chosen as the best suitable format for the input sequences for our DEMET model. The dataset was then balanced and the transformer models were fine-tuned on the balanced dataset. The evaluation was done on a balanced dataset under 5-fold cross validation, providing 20% of the data for testing. The evaluation results for the transformer models are as follows:

Model	Accuracy	Precision	Recall	F1-score
BERT	0.85	0.87	0.86	0.86
RoBERTa	0.86	0.90	0.81	0.85
DistilBERT	0.83	0.80	0.86	0.83
ClinicalBERT	0.81	0.91	0.71	0.79
BioBERT	0.82	0.85	0.81	0.83

Table 6.4: Evaluation Results for Transformer Models

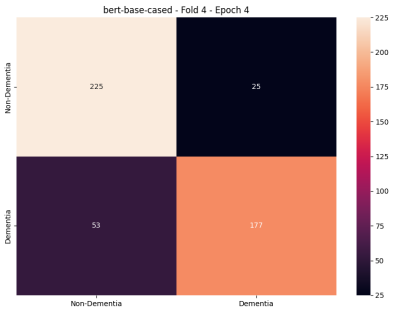


Figure 6.4: Heatmap for balanced BERT

We then proceeded to train and evaluate different classifiers on the transformer models' outputs. We refer to these models as single ensembles, as they combine the three concatenated

feature vectors provided by the transformer models as input to the classifiers. The classifiers used in this stage were Random Forest, Gradient Boosting, Support Vector Machine, Logistic Regression, Decision Tree Classifier, K-Nearest Neighbors and Bagging Regressor. The evaluation was done on a balanced dataset under grid search, providing 20% of randomly selected data for testing. The same was also done for the aggregation of the classifiers as a majority vote scheme. Aggregating the classifiers' output provided us with a singular prediction for the ensemble predictor. The evaluation results for the ensemble predictor are as follows:

Classifier	Accuracy	Precision	Recall	F1-score
Bagging Regressor	0.9616	0.9433	0.9609	0.9792
Random Forest	0.9666	0.9533	0.9662	0.9794
Gradient Boosting	0.9683	0.9533	0.9678	0.9828
Support Vector Machine	0.9616	0.9433	0.9609	0.9792
Logistic Regression	0.9600	0.9533	0.9597	0.9662
K-Nearest Neighbors	0.9633	0.9500	0.9628	0.9760
Decision Tree	0.9533	0.9366	0.9493	0.9623
Majority Voting	0.9666	0.9533	0.9662	0.9794

Table 6.5: Evaluation Results for Ensemble Predictor on 3 Transformer Models

The above results are the evaluation results for the ensemble predictor using three transformer models, BERT, RoBERTa and DistilBERT. They showed that the Gradient Boosting classifier provided the best performance for the ensemble predictor, with an accuracy of 0.9683, precision of 0.9533, recall of 0.9678 and F1-score of 0.9828. Through the evaluation process, we determined that the low variance in classifiers did not allow for the ensemble predictor to improve the model's performance significantly. We argue that the ensemble predictor's performance could be improved through the use of more diverse classifiers, and furhter investigation into the diversity of classifiers used in the ensemble predictor is required. In the figure 6.5 we provide a heatmap of the Gradient Boosting classifier's performance on the ensemble predictor. The heatmap shows the confusion matrix of the classifier.

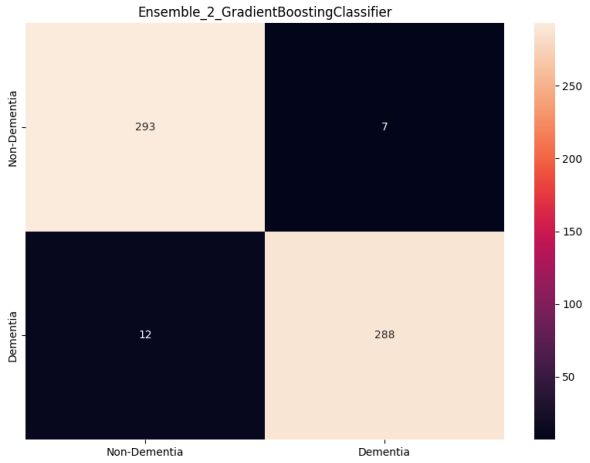


Figure 6.5: Heatmap for Gradient Boosting Classifier

We also evaluated the ensemble predictor using all five transformer models, BERT, RoBERTa, DistilBERT, ClinicalBERT and BioBERT with minor improvements in the ensemble predic-

tor's performance. The evaluation results for the ensemble predictor using all five transformer models are as follows:

Classifier	Accuracy	Precision	Recall	F1-score
Bagging Regressor	0.9683	0.9533	0.9678	0.9828
Random Forest	0.9766	0.9766	0.9766	0.9766
Gradient Boosting	0.9683	0.96	0.9681	0.9763
Support Vector Machine	0.9716	0.9633	0.9714	0.9796
K-Nearest Neighbors	0.9783	0.9733	0.9782	0.9831
Logistic Regression	0.9716	0.9633	0.9714	0.9796
Decision Tree Classifier	0.9633	0.9633	0.9633	0.9633
Majority Voting	0.9733	0.9666	0.9731	0.9797

Table 6.6: Evaluation Results for Ensemble Predictor on 5 Transformer Models

The results reinforced our initial hypothesis that the ensemble predictor's performance could be improved through the use of more diverse transformer models. Overall, DEMET's performance in predicting dementia through spontaneous speech yielded promising results and showed the potential of transformer models and ensemble learning techniques in improving model performance, which validated our initial hypothesis for this research. The single ensemble of the Gradient Boosting Classifier provided an accuracy of 0.9683, which is a significant improvement over the baseline transformer models' accuracies of 0.87, 0.85, 0.86, 0.81 and 0.82 for BERT, RoBERTa, DistilBERT, ClinicalBERT and BioBERT respectively, , along with improvements in precision, recall and F1-score. The ensemble predictor's performance could be further improved through diversifying both classifiers used, but also the transformer models used in each classifier respectively. The number of models used in the ensemble model should be investigated heavily in order to deduce the optimal number of models that would provide the best performance. Too few models and the ensemble predictor will not be diverse enough, too many and the ensemble predictor's judgements will be hindered by noise. For a majority voting scheme to be effective, the base models need to be diverse and provide different outlooks on the data. If all base models make similar errors and predictions, the majority voting scheme will not be able to improve the model's performance. Another approach to improving the ensemble predictor's performance is through the use of a weighted voting scheme, where the classifiers' predictions are weighted based on individual performance.

6.4.2 Explainability Evaluation

We evaluated three major explainable AI methods, namely LIME, Transformers-Interpret and Anchor, to determine the most suitable method for the ensemble explainer framework. Our metrics for evaluating explanations include both qualitative and quantitative metrics. We extracted the value of qualitative metrics through a survey that was conducted with clinicians and domain experts. The survey was designed to evaluate simplicity, and human evaluation, along with visual appeal and CHA token importance. We also evaluated the explainability of the methods through the quantitative metric of fidelity, the explanation's ability to approximate the model's decision and time of generation on GPU. Lastly, we included the qualitative metric of ease of use, due to the existence of a user-friendly interface in the client-side implementation of DEMET.

- **Simplicity** is the ease of understanding the explanation provided by the explainer. It is a measure of the explanation's ability to be easily understood by the user.

- **Human Evaluation** is the measure of the explanation's ability to be assessed as useful and informative by the user.
- **Visual Appeal** is the measure of the explanation's ability to be visually pleasing and engaging to the user.
- **CHA Token Importance** is the measure of the explanation's ability to highlight the most important phonological features
- **Fidelity** is the measure of the explanation's ability to approximate the model's decision.
- **Time of Generation** is the measure of time taken by the explainer to generate the explanation. and words that contributed to the model's prediction.
- **Ease of Use** is the measure of the explanation's ability to be easily generated by the user.

We scored the qualitative metrics on a scale of 1 to 5, with 1 being the lowest score and 5 being the highest. The resulting scoring for the three explainable AI methods, along with the quantitative metric we mentioned, are in the table below:

Type	Metric	LIME	T-I	Anchor	**DEMET**
Qlt	Simplicity	4	3	1	4
	Human Evaluation	4	3	1	4
	Visual Appeal	4	3	1	4
	CHA Token Importance	4	4	3	4
	Ease of Use	1	1	1	5
Qnt	Fidelity	0.73	1	0.9	1
	Time of Generation (GPU + 83 GB RAM)	55.47s	1.05s	0.11s	167.43s
	Time of Generation (CPU + 16 GB RAM)	**FAILED**	59.19s	9.43s	**FAILED**

Table 6.7: Evaluation Results for Explainable AI Methods

The results showed that the LIME explainer was the most suitable for the ensemble explainer framework, as it scored the highest in all qualitative metrics. The LIME explainer provided the most interpretable and understandable explanations, along with the highest visual appeal. All explainers managed to use the CHA tokens properly, and introduce their importance to the end user in a satisfactory manner. Anchor scored lower in the CHA token importance due to providing some examples in its explanations that were not tokenising the CHA tokens properly. As far as the quantitative metric of fidelity is concerned, Anchor scored the highest, due to its ability to provide a non linear approximation closer to the model's prediction function. The time of generation for Anchor took the least amount of time for both GPU and CPU, while LIME took the most amount, and failed in generating long sequence explanations on CPU due to memory constraints of the limited RAM of 16 GB. We tested the explanation generation on a subsequence of the original testing sequence **??** containing 5 words, and the generation took close to 5 minutes to complete. It is essential, due to computational constraints of using explainable AI methods, that services such as DEMET and the VHP of COMFORTAGE to run on high performance computing systems, possibly using microservices and cloud computing to provide the necessary computational resources for the model's prediction and explanation. After this assessment process, we determined that DEMET should use the LIME explainer for the ensemble explainer framework, as it provided the most interpretable and understandable explanations, even though it lacked in the quantitative metrics. Our key reasoning for choosing

LIME was that time of generation is not as critical of a factor as the qualitative metrics, since real time generation is not crucial for diagnosis and treating dementia. Also fidelity is being bypassed by the ensemble predictor and the explainer's approximation will not be a factor in DEMET's final output. If time of generation becomes a critical factor, we propose the use of Transformers-Interpret instead, as a faster alternative to LIME with similar performance in the qualitative metrics. DEMET scored the same as LIME in all metrics due to its architecture being a wrapper for multiple LIME explainers. DEMET was significantly easier to use that the other explainers, since it provided a user-friendly interface that allowed for easier access to the model's prediction and explanation without needing to write the code implementation for the explainer. It is important to note that the ensemble explainer framework is a proof of concept and is subject to further investigation and development in order to provide more comprehensive and accurate explanations.

6.5 Discussion

The results of our work on DEMET, an explainable AI-driven cognitive assessment agent for dementia detection through spontaneous speech, showed some promising results and provided valuable insights into the potential of AI-driven cognitive assessment agents. The ensemble predictor provided an accuracy of 0.96% on the DementiaBank dataset, which is a significant improvement over the baseline transformer models' accuracies. Our analysis of the CHA tokens showed that phonological features are important in generating explanations for clinicians and should not be omitted from the processed datasets used to train these models. A future direction of creating MVCHA tokens through multi-dimensional analysis of the CHA tokens is proposed, in order to sway the tokens towards the most appropriate class between dementia and non-dementia. The ensemble explainer framework provided valuable insights into the model's decision-making process and showed the potential of using ensemble learning techniques to improve model explainability. Through the increase of diversity in the ensemble explainer framework, we hope to provide more comprehensive and accurate explanations, with a focus on the importance of the CHA tokens in the input sequences. The computational constraints placed forth by the explainable methods showed that such systems require high performance computing systems to run efficiently as commercial applications and services. Our survey, despite its small sample size, showed that clinicians and domain experts found the research to be applicable and useful in the field of dementia detection and monitoring, not so much for clinical use, but for commercial applications and services, to provide tentative individuals with a non-invasive, cost-effective and efficient means of self-assessment and subsequent consultation with a healthcare professional. They also stressed the multi-faceted problem of dementia and frailty, and the need for constant monitoring of patients besides the use of automations and AI-driven assistance. The COMFORTAGE project aims to provide a solution to this problem, and we hope that DEMET can provide some assistance in achieving the project's goals.

Chapter 7

Ethical and Legal Considerations

Healthcare is in the midst of a profound transformation, being driven by the increasing amount and availability of data, the development of new and robust technologies and algorithms, and the increasing demand for personalised and precision medicine as the vision of Healthcare 5.0 comes to fruition. New discoveries and breakthroughs are being made in the field, and there seems to be no sign of stopping this progress. AI has proven to have the potential to become a valuable ally in this tranformation, providing evidence based insights and decision support to clinicians and healthcare professionals, enabling them to provide better healthcare services and import patient outcomes [152]. These improvements are prevalent in the areas of clinical diagnosis [18], drug discovery [154] [169], operational efficiency [78] and personalised healthcare [91].

It is apparent that such technologies can not safely and in good conscience be left unchecked. It is of paramount importance for a legislative framework to be put in place to combat what researchers and clinicians have stressed when discussing the impact of such technologies on healthcare and medicine. This framework should be designed to protect patients from unethical and miscalculated practices, and to ensure that the use of AI is in line with the ethical principles of healthcare and the betterment of patient outcomes that is, if AI is to be implemented and attached to clinical diagnosis.

AI healthcare is still in its infancy stage, but governments around the globe have shown great interest in the potential that it shows, investing heavily in the development of AI technologies [156] [171] [62]. That can prove problematic, granted that there is still a vast amount of uncertainty surrounding these ethical implications, namely biases, fairness and privacy. Researchers should now, more than ever before, focus their attention on the task of providing insight to governments about the implications of AI in healthcare, such the governmental bodies can legislate towards fair, ethical and safe use of AI.

7.1 Roadmap of ethical concerns

The map introduced by Mittelstadt et al. [109] provides a comprehensive overview of the ethical concerns that are associated with the use of AI in healthcare. The map is divided into three main pillars: the epistemic, the normative and the overarching ethical concerns.

1. **Epistemic concerns:** AI techniques have shown to be able to surpass human capabilities when it comes to evidence based decision making, and the ability to process and analyse vast amounts of data, where humans seem unable to do so. AI is scalable and can recognise patterns that are not visible to the human eye and intuition. This is a great

advantage, but it can be taken as fact, but rather as a support tool. Models can suffer from overfitting, miscalculations and lack of validation by external professional expertise. All these issues raise questions when it comes to the scientific backing of AI in healthcare, which is an obvious safety concern.
At the individual level, misdiagnosis and mistreatment of a patient can happen due to a wearable device with diagnostic capabilities, or a clinical decision support system that is taken at face value by a health professional. Research has shown that health professional can have a tendency of trusting the decisions of automated systems without critically assessing their output [32]. This can also lead to drop in skill level of health professionals [31], as they become more reliant on these systems and technologies.
These concerns can escalate to a societal level, where miscalulations in global health crises can have catastrophic consequences. A distortion of evidence of spreading diseases and viruses much like the influenza outbreaks [79] can lead to a lack of preparedness, mismanagement of resources and the distress of the people. These outcomes can have a significant impact on the global economy and the general trust in the healthcare system, government and academia.

2. **Normative concerns:** Arguments have been made and research is there to support that AI can produce unfair and biased outcomes for patients where an action or decision can have different effects depending on the group of people it applies to. This is a major concern of discrimination and fairness, where a model can learn to prioritise certain groups of people over others, who may have better predicted outcomes by the model [176] [29]. Another concern is the lack of patient oversight over the the collection of data being gathered passively by AI systems and IoT devices [147] which may or may not be used with ill intent. This predicament does not allow for the individual to have control over their own personal data, or have a say in the way it is being used while simultaneously being unable to assess and exert their own say as to how treatment is being administered.
Automations and AI systems in the field of healthcare can also strongly impact the relationship between healthcare professionals and patients. It is important to note that with excess automation of AI systems and diagnostic algorithms, healthcare can become dehumanised by the lack of human interaction and the feeling of not participating strongly in the decision making process for the treatment of patients. This can obviously reduce the capabilities of professionals, and also their rigour to provide the best care and service possible for their patients. Clinical practice, after all, is a process that is heavily based upon interactions between clinicians and patients, with constant evaluation, reassessment, trial and error to find the best possible treatment and provide favorable outcomes thus one should not base the entirety of the assessment process on the decisions of automated systems.
At a societal level, it is important to stress that algorithms and models are trained on data received by extensive research and sampling, which may be biased upon the population that is derived from. This produces models and algorithms that can help certain groups of people,able to generate enough data to train these models, and harm in a lot of cases, patients from other walks of life [156].
We thus arrive to the conclusion that there needs to be a clear and concise framework that delegates the amount of influence automations and AI systems can have in the decision making process of clinicians so that the risk of all these concerns can be reduced [117] [1].

3. **Overarching concerns** The concerns expressed in the previous excerpts are not isolated as mere technicalities in the ethics of healthcare, but rather paint a picture for a

very serious and complex issue that needs to be addressed. It is important to be able to address liabilities in these matters for when mistakes inevitably occur, and the ethics of responsibility and accountability play a significant role in this aspect [99]. The interaction between professionals, automated systems and AI support tools can make this an intricate matter, as it is difficult to determine who is responsible for the decisions that lead to mistreatment or misdiagnosis of a patient.
There needs to be a clear path to trace back to the source of the problem, and to be able to hold the responsible party accountable for their actions. This is a matter of transparency and trust, and it is important to be able to provide a clear and concise explanation to the patient as to why a certain decision was made, and who is responsible for it.
Danis and Solomon [97] make an argument for enabling patients to take action in their own healthcare, which is a matter of autonomy and informed consent on one's health and treatment. AI systems and automations can make this process very approachable by using IoT devices and wearables to gather data and provide insights to the patient. This can cause a multitude of issues. For one it can produce a conflict of interest between patient and automations where the patient may act upon ill decisions made by the system, thus causing harm to themselves. Lack of traceability and backtracking would make it difficult to hold the responsible part accountable since the authority assessing the situation would not be able to determine whether the patient did not adhere to the recommendations of the system, or if the system made a mistake.
The above aformentioned implications can scale to a larger issue of bias where certain groups of people can be assumed to care less about their general health and well-being and that they would be less likely to act upon decisions made by such systems. Koerber et al. [63] suggest that wrist-worn wearable devices have shown bias towards darker skin tones, when it comes to predicting arrythmias using LED reflections on the skin and thus creating the above assumption. This can lead to a general consensus that these groups of people are not to be trusted to act upon their own health, causing discrimination and unfair treatment by health institutions and insurance companies [104].
There needs to be regulation imposed by governing bodies to assess the responsibilities of all parties responsible for the creation, distribution and use of AI systems, from the developers and manufacturers, who need to ensure the quality of predictions of their products, to healthcare professionals who need to use these products in a responsible manner.

7.2 Patient privacy and consent

Privacy is defined as a human right by the United Nations [115] in the Universal Declaration of Human Rights, that no one is to be subjected to arbitrary interference with their privacy, family, home or correspondence. Laws and regulations have been put in place to protect these rights, like the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [35] but navigating the ethical implications of AI in healthcare even with these laws in place can be challenging. Firstly we should begin by explaining some core principles of privacy and consent as they relate to AI defined by HIPAA which is the regulatory entity that we will use as base to understand patient privacy and consent in more detail. Note that there are other entities with different regulatory systems in place, such as the GDPR in Europe where regulations may vary. The core principles are as follows:

7.2.1 Core principles of privacy and consent

1. **Protected Health Information (PHI)** is any information gathered by healthcare professionals during the course of providing health care services to a patient such as diagnosis, treatment, provisions and payments, that can be used to identify a person. PHI is used and distributed to researchers for necessary researchers and HIPAA regulations allow for such uses. PHI is clearly defined by HIPAA and is subject to the Privacy Rule. PHI is identifiable by 18 distinct identifiers and any data which does not contain any of these identifiers is not considered PHI and is not subject to the HIPAA guidelines for PHI. The 18 identifiers are as follows: names, dates, addresses/zip codes/geocodes, telephone numbers, fax numbers, email addresses, social security numbers, medical record numbers, health plan beneficiary numbers, account numbers, certificate/license numbers, vehicle identifiers and serial numbers, device identifiers and serial numbers, IP addresses, biometric identifiers, full face photos and any other unique identifying number, characteristic or code.

2. **Covered Entity** is a healthcare provider, health plan, or healthcare clearinghouse that uses and transmits PHI in electronic form. Covered entities are required to comply with the Privacy Rule and the Security Rule.

3. **Business Associate** is a person or entity that performs certain functions or activities on behalf of a covered entity, such as billing, claims processing, data analysis, or legal services. Business associates are required to comply with the Privacy Rule and the Security Rule and are required by law to have a Business Associate Agreement in place to mitigate the risk of using and disclosing PHI in a way that does not protect the patients privacy. This agreement covers the sharing of data between the covered entity and the business associate.

4. **The Privacy Rule** established a a set of national standards for the protection of PHI, and is enforced by the Department of Health and Human Services (HHS). The Privacy Rule applies to health plans, healthcare clearinghouses and healthcare providers who transmit health information in electronic form. The Privacy Rule requires that these entities implement safeguards to protect the privacy of PHI, and to provide patients with a notice of privacy practices that explains how their PHI will be used and disclosed. The Privacy Rule also gives patients the right to access their PHI, to request corrections to their PHI, and to request an accounting of disclosures of their PHI.

5. **The Security Rule** is a set of national standards for the protection of electronic PHI, and is enforced by the HHS. The Security Rule requires that covered entities implement administrative, physical, and technical safeguards to protect electronic PHI, and to conduct risk assessments to identify vulnerabilities in their systems. The Security Rule also requires that covered entities implement policies and procedures to prevent, detect, contain, and correct security violations.

6. **Consent** for usage and disclosure of PHI falls into three separate categories under HIPAA regulations. The first one is when no consent is required to use PHI, which is an exception when it comes to public health and safety concerns, a covered entity's usage to provide treatment, payment and healthcare operations, and to prevent or lessen imminent danger or pain to a person or the public, for example disclosing information to law enforcement in regards to a criminal activity. The second category of consent is when verbal consent or acquiescence is required to use PHI, for the disclosure to family members, friends or other persons involved in the patient's care or disclosures in facility directories where patient

data is stored to be displayed for healthcare services. For when a patient is incapacitated or in an emergency situation, covered entities have to ability to disclose information without the patients consent if they determine that such actions are for the betterment of their health and well-being by their own professional judgement. Lastly, the third category of consent is the requirement of written and explicit consent by the patient for the usage and disclosure of PHI for general requirements, with exceptions being limited to the ones mentioned in the first category, phychotherapy notes, marketing and sales of PHI, and research purposes. In order for a written consent to be considered valid, as defined and regulated by HIPAA, the form of consent must contain the information to be used, the persons and entities to use said information, the purpose of usage, a date of expiration of usage, and lastly a place where the patient can provide their name and signature.

7. **Breach** is the acquisition, access, use or disclosure of PHI in a manner not permitted by the Privacy Rule which compromises the security or privacy of the PHI. Covered entities are required to notify affected individuals, the HHS, and in some cases the media if a breach of unsecured PHI is detected. Financial and or criminal penalties for the knowing and willful neglect of the Privacy Rule can be imposed on the perpretrator.

7.2.2 Research and development under privacy regulations

In order for researchers to be able to use PHI in the development of AI systems, they can proceed with one of two pathways. The first one is to deidentify the data, or the Safe Harbor method, which means to remove any and all occurrences of the 18 identifiers that are used to identify whether data is PHI or not. Once a researcher has managed to completely deidentify the data, they are free to use it in any way they see fit, as it is no longer subject to the Privacy Rule of HIPAA. This allows for complete freedom to train, validate, test and distribute AI technologies without restrictions. There's three main downsides with this method, the first one being that one has to deidentify the data in the first place, which can be a rigorous task. The second one is to ensure the deidentification of data has been successful in its entirety, as any remaining identifiers can still be used to identify a person and thus make the data unusable. And the third one is that by striping parts of the data the model may not be as robust as it could be, as it is not trained on the entirety of the data. Note that researchers and AI developers are considered Business Associates and are required to obtain a Business Associate Agreement that states information about how the data is going to be used and for what purpose. This approach is not without serious risks. Namely, there is a risk of reidentification of data by malicious actors, or the merging of non PHI data with other datasets to produce PHI and assume the ability to identify a person. Often times it is not expected by covered intities to be able to safeguard against such techniques and thus the deidentification of data is not a foolproof method. Additional risks include the exploitation of loopholes and grey areas in the Privacy Rule, and what really constitutes research and development for the betterment of healthcare and patient outcomes by corporations and business entities.
A second pathway is to obtai written and explicit consent from the patient to use their PHI in the development of AI systems as we discussed in the previous section concerning consent and its categories. With this method challenges arise when it comes to collecting and obtaining consent from a large number of patients in order to have a diverse and representative dataset capable of training a model sufficiently. In addition to the challenge of obtaining the data, researchers have to be accepted by permitted disclosure under specific guidelines and regulations, often rigorous and time consuming.

7.2.3 Privacy preserving AI techniques

Covered entities are required by law to make resonable efforts to disclose the minimum amount of PHI necessary to accomplish an intended purpose such as AI development. This poses a significant challenge to researchers and developers alike, since it is in the very nature of robust and complicated models to require vast amounts of data to be able to make accurate predictions and produce reliable outcomes. This duality of requirements can be a significant barrier in the breaching of cutting edge tech and also the assurance of patient privacy and consent.
Privacy preserving AI techniques have been developed and are still a subject of research to help mitigate the issues of privacy in health related data. Pati et al. [135] propose the use of federated learning to train their model collaboratively without the need to share data between institutions. This allows for global scale training of models and diversity and complexity of data beyond what a single institution can provide. Yoon et al. [12] use a different approach by proposing the use of synthentic data. They introduce a framework for generating highly realistic and privacy-preserving synthetic EHR data called EHR-Safe, combining sequential encoder-decoder networks with adversarial training while preserving the aspects and properties of the original data. A similar approach is taken by Kumar et al. [8] who use generative adversarial to produce synthetic data which are in turn solely used as the benefactor for the training of a deep learning model for glaucoma detection. One would be remiss not to mention the work of Chambon et al. [33] who leverage the Stable Diffusion model to generate domain specific imagery. Their approach consists of exploring various sub-components of the Stable Diffusion pipeline to fine tune their proposed model to generate synthetic medical images.
Synthentic data introduces a new set of posibilities by allowing the training of models on PHI-esque data without the need to disclose any real PHI [61]. From a privacy perspective, there is an issue to be argued about how much the synthetic data is trully representative of the real data, and also how much of the data is trully synthetic.

Researchers	**AI Method**	**Description**
Pati et al.	Federated Learning	Desentralised training of models without sharing data
Yoon et al.	EHR-Safe	Generation of synthetic EHR data
Kumar et al.	Generative Adversarial Networks	Generation of synthetic data for glaucoma detection
Chambon et al.	Stable Diffusion	Generation of synthetic medical images

Table 7.1: Privacy Preserving AI Techniques

7.3 Bias and fairness in AI algorithms

Healthcare is constantly adopting more AI use cases for detecting disease, predicting patient outcomes, generating treatment plans and supporting clinicians and healthcare professionals to make better decisions in order to provide better care and services to patients. As we already discussed, AI has been a major factor for progress in the area of healthcare but is not without fault and ethical implications. Questions about fairness have been plaguing researchers, developers and governing bodies, as AI systems often show inconsistent performance in different groups of people or propagate pre-existing biases in the data they are trained on. Data is collected by humans, and humans are inherently biased, which is a reality that can not be ignored.

We will now delve deeper into the ethical implications of bias and fairness in AI algorithms.

7.3.1 Core concepts of bias and fairness

Let us first set the stage by defining some core concepts that are relevant to the discussion as it pertains to bias and fairness in AI algorithms.

1. **A biased algorithm** is an algorithm that demonstrates significant differences in performances across different groups of people based on:
   (a) their demographic characteristics, such as race, ethnicity, age, gender, sex etc.
   (b) their socioeconomic standing, such as income, education, insurance status
   (c) their geographical location, such as urban or rural areas

   A biased algorithm can be biased towards a single element of the above, or a combination of them.
2. **Equality** is the giving of equal resources and opportunities to all individuals but does not ensure equal outcomes for all.
3. **Equity** takes into consideration the differences of people in a socioeconomical context and provides resources and opportunities so that people can reach equal outcomes [122].
4. **Disparity** refers to the differences in outcomes in the context of fairness.
5. **Type I Error** is a false positive error where the algorithm predicts a positive outcome when the true outcome is negative.
6. **Type II Error** is a false negative error where the algorithm predicts a negative outcome when the true outcome is positive.

7.3.2 Sources of bias in AI algorithms

There are a lot of different ways that bias can infiltrate an AI algorithm, and it is important to be able to identify where in the process of creating and distributing AI algorithmic solutions bias can be introduced.
The first way of introducing bias when trying to solve a problem is essential asking the wrong question. Strategic objectives in the implementation of AI algorithms in health should reflect health equity and justice which means that researchers and developers alike should be rethinking their approaches to developing AI systems and automations. Their strategy should involve moving from aggregated results such as high scoring models and growing service line volumes to more granular and patient-centric outcomes such as models that perform equally well across different and diverse populations and solutions that are accessible to all communities.
The second way is how bias can be introduced in the data that is used to train models for healthcare [114]. A lot of times models make crucial mistakes in the deployment stage of their lifecycle where they are asked to make predictions and support decisions for clinicians in real world situations and environments. It is obvious that mistakes like these can be mitigated if the data used for the training of said models is up to the task at hand. A comprehensive scoping review done on the subject by Daneshjou et al. [124] found that the majority of studies reviewed for detecting skin disease lacked significant information concerning the data used for

training and validating AI models like ethnicity or skin tone. A significant minority of the datasets where publicly available and close to 40% of datasets used for developing cutting edge AI tech where below the standard for labeling and annotation. The researchers concluded that the datasets being used for development of AI technologies lack transparency, have low standards and are unable to assess patient diversity. Some very interesting initiatives have been taken by various researchers to address the issue of bias in datasets. Namely the Data Nutrition Project is a driving force in the development of frameworks for assessing the quality of datasets used for training AI systems [67] enabling researchers to select the best datasets before developing their models with the the Dataset Nutrition Label. Their focus is on understanding the data by proper feature selection and data preprocessing, taking into account privacy, data completeness and proper representation. The knowledge of bias in datasets should strongly impact the way that research and development of AI models and systems is conducted, driving approach, methodology and standarisation of data collection, curation and preprocessing to produce the best possible software.
Yan et al. [100] make a strong argument for the incomplete capture of patient outcomes in electronic health records, which is another way that bias can creep into AI model decision making. Essentially the issue lies with the fact that datasets included in the training of AI systems should only contain data causal to the proper identification of a disease or condition, but often times data contains differential observability factors, for example race, that cause algorithms to be biased. These factors are unobservable in the data and thus researchers can not account for them in the testing process of the model and so they move forward with the assumption that there are no such differential observability concerns. Another issue with EHR is their inability to capture certain intricacies with social determinants of health and aggregated health statuses of patients, which is combated by using proxy variables. These proxy variables are not always accurate and can introduce bias themselves. In its essence the issue lies in the actual way healthcare is provided to the patients and this leads to an incomplete capture of patient information. This can lead to algorithms that perpetuate these findings and thus continue this cycle of bias.

7.3.3 Mitigating bias in AI algorithms

By discussing the sources of bias in AI algorithms, we should now be able to address ways to mitigate bias and ensure fairness in AI algorithms in healthcare. Genevieve Smith and Ishite Rustagi [146] in their leadership playbook, inform on the importance of setting up a diverse and multi-disciplinary team to develop AI systems. Engaging individuals in the ethics and social sciences space as well as domain experts in healthcare and artificial intelligence that prioritise equity and justice. Much like the Data Nutrition Project, they also stress the importance of practicing responsible dataset curation and preprocessing, along with establishing policies for corporate governance, leadership and accountability. Kerstin et al. [161] propose a roadmap outlining how bias can be mitigated during the entire lifecycle of an AI system, from its inception to its deployment. The roadmap is divided into four different stages: the data collection and preparation stage, the model development, the model evaluation and lastly the deployment of the model. During each step of the way, the researchers propose specific actions to be taken to ensure that bias is not introduced into the final product. PROBAST [7] is a tool that can be used to assess the risk of bias in prediction model studies and can be used to evaluate the quality of the data used for training AI systems during the data collection and preparation stage. In the development stage Brian Hu Zhang et al. [170] propose the use of adversarial learning to reduce demographic bias, such as gender or zip code, in AI systems drawing from the work of Hardt et al. [60]. This is done by simultaneously training a predictor and an adversary for a given input X, to predict an output Y, for example an income bracket, and a protected variable

Z, for example gender. During the model evaluation phase, an obvious solution would be to use explainable and intrepretable solutions, much like the ones mentioned in our literature review 3.4. Lastly, after the model has been developed and evaluated, during the deployment stage of its lifecycle, researchers and developers should be monitoring the model for differences between patient cohort in clinical practice and patient cohort in the training data, a phenomenon called data shift, which we will address in detail later in 7.4.3. This way errors to the model's decision making can be detected and corrected. The researchers stress that these strategies should become best practices in the development of AI systems in healthcare to ensure that bias is mitigated and fairness is ensured.

7.3.4 Frameworks to reduce bias in healthcare

Researchers are well aware of the implications of bias in AI algorithms and are working towards developing ways to reduce it for better patient outcomes. Ben Green [56] makes a serious case for ethics in the tech industry lacking the backbone to address some hard hitting issues when it comes to the development of AI systems, as some corporations address ethics in an abstract manner, and their positions are lacking in explicit commitments to fairness and justice with proper calls to action. Large corporations and industry giants like Google AI have proposed various frameworks enabling developers to assess the fairness of their data and models to battle claims like the ones Green has made. One such framework is the Fairness Indicators [127]. This framework is a suite of tools that can generate metrics for transparency reporting and fairness analysis, which helps developers produce models more responsibly. The toolkit is able to compute confidence intervals which can detect disparities in the model's performance across different groups of people. The What-If Tool [128] is another tool developed by Google AI, which allows developers to evaluate their models in hypothetical scenarios by changing the input stream of data and observing model behaviour. These two tools can be intergrated together to provide a comprehensive solution for assessing fairness and bias in the development of AI systems. Saleiro et al. [137] propose a similar framework called Aequitas, which is easy to use and apply to one's machine learning workflow to assess bias and fairness in the data used for training models. The framework tests data for bias across different groups of people and provides interpretable explanations for the results of the test. IBM AI has also been a prevalent force in the space with the development of the AI Fairness 360 toolkit [25] which contains metrics for assessing bias and fairness in data and models, and algorithms for mitigating bias in AI systems as well as Watson OpenScale [71] which is a development and deployment platform that provides solutions for bias detection in model data and decision making, regulation compliance auditing and model explainability. IBM has also developed the AI Explainability 360 toolkit [20] which provides a comprehensive set of algorithms for model explainability and interpretability. Another open source tool is Fairlearn [163] which is a Python package that provides help in considering a system's societal context based on the work of Ben Green.

Developer	**Method**	**URL**
Google AI	Fairness Indicators	https://www.tensorflow.org/tfx/guide/fairness_indicators
Google AI	What-If Tool	https://pair-code.github.io/what-if-tool/
Saleiro et al.	Aequitas	https://github.com/dssg/aequitas
IBM AI	AI Fairness 360	https://github.com/Trusted-AI/AIF360
IBM AI	Watson OpenScale	https://client-docs.aiopenscale.cloud.ibm.com/html/index.html#
IBM AI	AI Explainability 360	https://github.com/Trusted-AI/AIX360
Microsoft	Fairlearn	https://fairlearn.org/

Table 7.2: Bias Reduction Frameworks

7.4 Compliance with healthcare regulations

It is apparent that compliance with healthcare regulations is of paramount importance when it comes to the development and deployment of AI systems in healthcare after discussing all the ethical implications in privacy and consent, bias and fairness in the above excerpts. Safety and efficacy of AI systems are at the core of the regulatory frameworks in place to ensure that patients are protected from harm. The Food and Drugs Administration in the US has been at the forefront of this initiative, proposing regulatory frameworks for modifications in AI systems [47] based on the risk categorisation principles of the International Medical Device Regulators Forum (IMDRF) [48] as well as released an action plan that outlines a multi-faceted approach for the entire lifecycle of AI products and systems [46] to aid in the development of systems that assist in better patient outcomes.

7.4.1 AI Act

Regulatory compliance is about ensuring that bussinesses and organisations adhere to the laws, regulations and specifications which are relevant to their business activities and operations. We have already discussed some of the regulations concerning patient privacy and consent placed forth by HIPAA in section **??**, which have been a great aid in ensuring safety and efficacy of AI systems in healthcare. The General Data Protection Regulation (GDPR) in Europe has also produced valuable efforts in the protection of patient data and privacy, driving the development of AI systems and promoting accountability and transparency in the space. Europe has also been the first to propose a comprehensive and unified regulatory framework in the Artificial Intelligence Act [34] which aims to define a global standard for the development and use of AI systems in all domains and sectors. Many other countries have followed suite to impose their own policies in relation to Artificial Intelligence. The Act assigns applications of AI into four separate and distinct categories: unacceptable risk, high risk, limited and minimal risk. AI Act explicitly states the banning of all systems deemed of unacceptable risk by the Commission, for their potential to seriously harm the safety, livelihoods and rights of individuals. This european legislative framework also states a comprehensive list of high risk applications that are to be subject to strict and rigorous requirements and obligations before they can be deployed for use. In addition to the above, the Act clearly defines the obligations that providers of AI systems have to adhere to, which state many of the prefered requirements that we have already discussed in the previous sections about privacy and bias. These matters include:

1. adequate risk assessment and mitigation systems in place
2. the usage of high quality data devoid of discriminatory biases
3. proper logging during development lifecycle activities for traceability and transparency
4. Clear and concise documentation stating the purpose, functionality and limitations of the system
5. Clear and concise documentation for the deployer and user of the system
6. Appropriate human oversight and control mechanisms in place for risk mitigation and assessment
7. High level of robustness, accuracy and security

The case of limited risk is mostly associated with lack of transparency. The Act introduces clauses concerning transparency obligations of developers and providers in general when it comes to informing the public about content generated by AI systems, like chatbot communications and deep fakes. It is the providers' responsibility to ensure that their AI systems are in fact identifiable as such. These steps are taken to promote trust between the public and corporations developing such systems. Lastly, on the low or no risk side of the spectrum, there are no specific regulations imposed by the Act. Applications in this category include AI systems used in entertainment and spam filtering, and compose the majority of AI systems in use today in the EU.

In addition to regulations and requirements imposed by the AI Act, they offer a pipeline for providers to follow to be in accordance with the legislation. Once a provider has developed an AI system, they are required to undergo a conformity assessment procedure by the Commission to ensure compliance with the Act. In cases where domain specific knowledge is required, a notified body is also to be involved in the assessment process. The system is then registered in the EU database for AI systems. Once the conformity assessment is completed, a declaration of conformity is issued upon the system which bears the CE marking [1] and the system is then ready for deployment. In the case of notable changes to the system, it is required by the provider to undergo a new conformity assessment procedure to ensure that the system is still in compliance with the Act. Once a system is deployed, it is up to all parties involved to ensure that the system is functioning as intended with systems in place to monitor and assess the system. Perperteators of the Act are subject to fines and penalties for non-compliance. The AI Act is expected to be fully implemented by Spring of 2024.

7.4.2 Corporate Adaptation

It is apparent that with new legislative bodies in place to regulate the development and deployment of AI systems and automations, corporations and businesses need to adapt in order to be able to comply with the ever changing regulatory landscape. Corporations are strongly encouraged to adopt a culture of ethical responsibility when it comes to AI, in order to showcase their commitment to the safety and well-being of the public. It is not the first time that corporations have been faced with a paradigm shift in regulatory frameworks, where they were tasked to adapt to new standards and practices. Once such occurrences take place, corporations are required to reevaluate their business models and strategies, and those who manage are viewed as socially responsible and ethical which can be a great promotional leverage. The companies that are delayed in their compliance can fall behind in the market and lose their competitive edge. So it is in everyone's best interest move forward in accordance with the new regulations.

There are specific steps that a corporation can take to ensure compliance with the latest regulations. In the executive level, corporations should start their process by aligning their business strategy with the goal of creating responsible AI applications along with resource allocation and proper budgeting for the development of said systems, such as hiring new and capable talent and investing in the latest tech. In addition to this, corporations should also establish an ethics committee for the reviewing of ongoing projects. Teams involved in the development process should be diverse and multi-faceted, including domain experts, lawyers and business professionals. Training and education should be provided to all members of the team to ensure adherence to established regulations and standards. Education should also aim to instill a culture of ethical responsibility and AI governance. At the operational level, teams

[1] CE marking is a certification mark that indicates conformity with health, safety, and environmental protection standards for products sold within the European Economic Area (EEA).

are advised to set up workflows and processes that ensure that all requirements by governing bodies are met. These processes entail collecting, processing and storing data, assessing model performance and safety, logging activities and producing reliable documentation and reports. These steps can greatly increase a corporation's ability to comply with regulations while also promoting a culture of ethical responsibility in building AI software.
Consulting firms and legal entities have also entered the space and have started offering services to corporations in order to aid them in their compliance with regulatory systems and standards offering software solutions and tools, automating a lot of the tasks that are required by corporations.

7.4.3 Regulatory Challenges

It is apparent that navigating through the legislative landscape can be challenging for corporations in many ways. It requires great planning and resources to manage to adhere to regulations every step of the way but it is also a great opportunity for corporations to move the needle forward when it comes to responsible Artificial Intelligence. Deriving from the above discussion, it is clear that data is a significant driving force in AI software development. The quality of data often dictates the quality of the model and the outcomes it produces. But what happens when data becomes outdated or irrelevant? Zech et al. [125] found that the performance of a deep learning model for predicting pneumonia disease from chest x-rays decreased significanlty when altering deployment settings and environments. Schulam et al. [141] trained a model on lab clinical EHR to predict risk of an adverse event [2] on data from 2011-2013, which tested well on data for the following next year, but had a significant drop in performance when testing on data from 2015. This effect is known as data shift and is a significant challenge for AI systems in industry, and as it so happens, it is one of the main reasons why developers have to reassess their models with the authorities after deployment.
The problem of data shift arises from differences in the environment and settings a model was trained on and the environment and settings it is deployed on. These changes can occur by changes in environment over time, i.e. outdated, or differences caused by situational factors in deployment, i.e. irrelevant. Models who are trained on data that are subject to data shift can begin over time, to generalise poorly and produce inaccurate predictions [148].
We will focus on the specifics of data shift in regards to healthcare. The main reasons as to why data shift happens in healthcare mostly has to do with three main categories of interest, namely changes in technology, changes in patient populations and setting and lastly change in behaviour [51]. Changes in technology usually occur when there is a shift in a piece of equipment used or an infrastracture changes. Changes in patient populations and setting refers to demographic changes in population or changes in disease prevalence and incidence. Behavioural changes are associated with how patients interact with the healthcare system, how the clinicians change their practices and how incentives are structured and altered.

There are two main pathways to combat data shift. The first one involves reacting to the changes in data by retraining the model on new data, specific to all target environments, shifted towards the proper distributions in accordance with the changes in setting and environment. This requires a clear understanding of the target domain, plethora of data and constant monitoring and maintenance of a model. Now this approach can be costly and time consuming, factors that can be a significant barrier for corporations to overcome when it comes to large scale software. Gretton, in his work [10], and following this approach, matches the distributions of training and testing data in a high dimensional space by using a kernel mean matching

[2] An undesired effect of a drug or other type of treatment, such as surgery

technique. Sugiyama et al. [98] propose a similar approach by using a covariate shift correction based on importance weighting. In critically important situations it is very important for models to be able to function properly without major changes after deployment. It is apparent that clinicians and healthcare professionals require a different approach. This approach is called proactive adaptation, the second pathway to combat data shift. In this case, the model does not require the amount and specificity of data that a reactive approach would require. Models trained on this paradigm should be able to anticipate and identify possible data shifts and their potential risk to the target outcome. Subbaswamy et al. [149] [150] have done significant work in this area by proposing a graph based approach where they use causal graphs to identify the relationships between variables that do not generalise across different environments. The graphs identify paths that cause instability in statistical influence and are subsequently removed, favouring relationships between variables that are stable across different environments.

Concluding this section, the ethics of AI in healthcare is a complicated matter and a lot of great minds have gathered to approach the subject from different angles, providing solutions and frameworks to combat the issues that arise. The importance of patient privacy is not be understated for it is a basic human right and researchers and developers should be wary of that. The necessity for unbiased, fair and responsible AI systems is prevalent as well, and researchers have shown great efforts in developing with these principles in mind. Governments and legislative bodies have also taken a stand, in what it seems to be a global phenomenon which will change the way humans receive healthcare, in legislating for responsible endeavours in AI development and deployment in accordance with the vision for Healthcare 5.0.

Chapter 8

Future Directions and Challenges

8.1 Emerging trends in explainable AI for healthcare

In recent years there has been growing interest in the literature for explainability and transparency in the use of AI systems. This is particularly important in fields where the decisions made are of critical importance to the well-being and safety of individuals, organisations and society in general. In healthcare, the need for explainability is ever more important as the decisions made by AI systems can have significant consequences on the health and well-being of patients. The responsible use of AI applications is becoming a necessary utility in healthcare, in order to address the ethical, legal and social implications of these systems. Explainability is a key component in ensuring that AI systems are used responsibly and ethically, and is a tool that is ever changing and evolving since its infancy in 2004. The following section will discuss the emerging trends in explainable AI for healthcare, and how these trends are shaping the future of AI in healthcare.

Interactive Explainability
Interactive explainability is a new trend in the field of explainable AI, and is the practice of allowing the user to interact with an explainable framework and tweak the parameters of the input data or the model to see how the explanation changes. This is usually done with the integration of an explainability method with a user interface. The interface provides different options to the user for altering parameters and seeing how the outcome of the explanation is affected. A prime example of such a system is the What-If Tool by Google AI, which allows users to interact with the input data and see how the model's behaviour is altered. These frameworks are particularly useful due to real-time feedback and the ability to explain by interpretable examples. This may allow for clinicians to adjust certain parameters in their patients' data and see how the AI behaves acoording to these adjustments, and also correlate their cases with similar cases in the past. Interactive explainability is a trend that has major significance in the field of healthcare, providing clinicians and healthcare professionals with enhanced understanding of a model's behaviour and the relationships between recommendations made by the system and the patient's data and profile. Subsequently, higher trust can be achieved and personalised insight about specific cases can be obtained, allowing for a more tailored and actionable approach. Such systems can also be utilised to detect potential biases and errors in the model's decisions when cross-referenced with clinical expertise in order to promote higher performance and explainability.

Integration with Electronic Health Records (EHR)
XAI systems are being intergrated with EHR systems to provide clinicians with real-time explanations of the AI's decisions based on the profile of the patients and his personal health records

and medical history. This is particularly important in providing clinicians with the necessary information and insight in order to provide personalised care and treatment to patients. This enhances the relationship between clinician and AI system interactions and provides holistic and comprehensive diagnostic tooling and recommendation systems for healthcare. When integrated with EHR systems, XAI explanations can be more interpretable and actionable, providing clinicians with the necessary information to make informed decisions and account for possible biases and errors, due to their own expertise and knowledge. These systems can be used with interactive dashboards much like the VHP system introduced by COMFORT-AGE, where patient data, model predictions and explanations are provided in an intuitive and user-friendly manner. A well-integrated system provides a seamless user experience where AI tools are embedded into the clinical workflow and provide real-time insights and explanations, even possibly streamlining routine operations and tasks and provide secure and reliable decision support and automations. Visualising EHR data may also provide clinicians with the ability to assess the model's behaviour and provide feedback on the model's performance and explanation quality for further improvement.

Regulatory Compliance

There is a growing focus on developing explainable AI systems that are not only reliable, but also ethical, trustworthy and compliant with regulatory standards. This is particularly important in healthcare, where decisions made are of critical importance. The development of explainable AI systems need to be designed with considerations for patient privacy, consent, and the overall impact on the healthcare system. Regulatory institutions such as the GDPR, HIPAA and FDA have set guidelines and standards for the development and deployment of AI systems, along with the proper handling of patient data. These regulations are put in place to ensure that AI systems are used responsibly and ethically, and organisational and corporate adherence to these standards is crucial in ensuring the safe and secure usage of AI systems in healthcare. Regulations include mitigating bias through different channels and techniques, providing data security and privacy by using methods suc as encryption, access controls and audit trails, and data sanitisation and anonymisation, by cleaning, de-identifying and imputating data.

IoT Healthcare

XAI is increasingly being intergated with IoT devices in healthcare in order to enhance the transparency, trust, and usability of AI-driven insights derived from IoT devices, which are extensively used for monitoring and managing patient health. IoT devices such as wearables, smart-home applications and monitoring devices collect real-time data on patients. XAI systems can provide interpretable insights and automations by detecting anomalies or trends in the patients health, providing clinicians with recommendations and insight into the patient's health status. This is especially applicable in elderly care and chronic disease management, where continuous monitoring and early detection of health risk is crucial. XAI can also be used in IoT servicing and maintenance, by providing predictions and explanations for possible failures or malfunctions to ensure timely intervention and maintenance reliability. The integration of XAI with IoT devices provides significant benefits in healthcare, by increasing reliability, trust and adoption of AI-driven insights and recommendations, increasing transparency and accountability and improving clinical decision-making and patient outcomes.

In conclusion, the integration of XAI in healthcare is driving significant advancements in patient care, clinical decision-making, and system reliability. As these technologies continue to evolve, the focus on transparency, ethical compliance, and user-centric design will be crucial in realizing the full potential of AI in healthcare, ultimately leading to improved patient outcomes

and a more efficient healthcare system.

8.2 Anticipated challenges and potential solutions

Integrating XAI in healthcare, although promising and certainly beneficial, is not without its challenges. These challenges need to be addressed to ensure effective and responsible use of AI systems in healthcare. In this section, we will discuss some of the anticipated challenges in the space and potential solutions to address them.

Explainability and Interpretability
Through this research, we have discussed the importance of explainability and interpretability in AI systems, and the difficulty of providing reliable and accurate explanations for complex models. Providing these explanations that are both accurate and understandable to clinicians and patients is difficult, especially with complex AI models like deep learning. Developing user-friendly interfaces and visualisation tools that can translate complicated computations of explainability methods into intuitive and actionable insights is crucial in ensuring that AI systems are used responsibly and ethically. It is important to develop these trools in collaboration with domain experts and end-users in order to ensure that the explanations provided are relevant and useful in clinical decision-making.

Bias and Fairness
XAI models can inheret biases from all stages of the model's development and deployment, including data collection, preprocessing, feature selection, model training and evaluation. Biases can also be introduced by outdated or incorrect data. This can lead to unfair and discriminatory outcomes, particularly in healthcare. To combat this issue, it is important for developers to implement different bias detection and mitigation techniques, much like the ones we discussed in previous sections. Additionally, it is important for developers to use diverse and representative datasets, hire and maintain diverse teams, and continuously monitor and evaluate the model's performance for biases and fairness. This can help ensure that the model is fair, unbiased and reliable, and that the decisions made by the model are ethical and just.

Interoperability
Integrating XAI systems with existing healthcare systems and infrastructure can be challenging due to differences in formatting, standards and protocols. This can lead to issues in integration, data sharing and overall communication between different components of a system or infrastructure. The way to address this challenge is to develop standardised data exchange protocols and APIs in order to facilitate the seamless integration between applications, systems and AI models. XAI systems should be designed into the clinical workflow and provide real-time, actionable insights and recommendations to clinicians and healthcare professionals. The tools should enhance the decision-making process and efficiency of the healthcare system in place.

Data Quality and Consistency
One of the major challenges in the development of XAI systems is the quality and consistency of the data used to train and validate machine learning models. In healthcare, data is often incosistent, noisy, incomplete and possibly biased. Data comes in many forms and by many sources, such as EHRs, medical imaging, wearables and clinical notes, and may be stored in different formats and structures. This can lead to challenges in data integration, processing and analysis, and may affect the overall performance and reliability of AI systems. To combat this issue, it is important to ensure robust data cleaning, standardisation, and integration processes. Ensuring data quality and consistency is crucial in developing reliable and accurate AI models,

and even more so in healthcare where the decisions made are of critical importance.

Data Privacy and Security
Another challenge in the development of XAI is data privacy and security concerns. In healthcare, patient data is higly sensitive and confidential, and must be handled in such a way that ensures patient privacy and compliance with regulatory standards such as HIPAA and GDPR. To ensure compliance with these standards and provide security in data handling, it is important to implement encryption schemes, access controls, anonymisation techniques and regular security audits and compliance checks in order to maintain the integrity and confidentiality of patient data.

Regulatory Compliance
Developing XAI systems that are compliant with regulatory standards is crucial in ensuring the responsible and ethical use of AI in healthcare. Regulatory standards such as the GDPR, HIPAA and FDA have set guidelines and standards for the development and deployment of AI systems, along with the proper handling of patient data. It is important for developers to adhere to these standards and stay updated with the latest regulatory guidelines. It is also important for organisations in the space to involve legal and compliance teams in the development and deployment of AI systems in order to ensure that the systems are compliant with these regulations.

Constant Adaptation
Healthcare is a dynamic field that is ever changing and evolving, especially with advancements in technology and medicine. AI systems need to be able to adapt to these changes and provide accurate and reliable recommendations and insights to clinicians and healthcare professionals. This requires regular monitoring and evaluation of the model's performance, and continuous re-training and updating of the models used in clinical practice on new data. There also needs to be a feedback loop between the AI systems in place and the clinicians using them, to ensure the models remain up-to-data and relevant to the clinical workflow.

Resource Allocation and Cost
Developing and deploying XAI systems in healthcare can be a costly and resource-intensive indeavour. It requires significant investment in infrastructure, data collection, model development and deployment, maintenance and personnel. It is important for organisations to demonstrate the long-term benefits and cost-effectiveness of these systems in order to justify possible investments. It is also important that these systems are developed through pilot projects and case studies to ensure proof of concept and feasibility before full-scale deployment. Funding and resources should be sought from government, private and public sectors and stakeholders, as well as partnerships with research institutions, policymakers and industry partners.

Integrating XAI in healthcare presents multiple challenges, from ensuring data quality and security to achieving regulatory compliance and clinician trust. Addressing these challenges requires a collaborative effort and a multifaceted approach among technologists, healthcare professionals, regulatory bodies, and patients to fully harness the potential of XAI for improved patient outcomes, enhanced clinical decision-making, and a more efficient healthcare system.

8.3 Recommendations for future research

Our research has highlighted the importance of explainable AI in healthcare, and the potential benefits and challenges of Integrating XAI systems in clinical practice. In this section, we will provide recommendations for future research in the field of XAI in healthcare.

Real-time Explainability
Developing real-time explainability methods that can provide clinicians with actionable insights and recommendations in real-time is crucial in enhancing the clinical decision-making process. Dynamic and interactive explainability tools are very important as patients' health status can change rapidly and require timely intervention. Real-time explainability can provide clinicians with the necessary information to make informed decisions and account for possible biases and errors in the model's predictions. The issue of real-time explainability is particularly important in certain clinical scenarios, such as emergency care and critical care, where timely intervention is crucial. Future research should focus on developing real-time explainability methods that can provide interpretable insights and recommendations to clinicians in real-time, and enhance the overall clinical workflow. In this work, we discussed how LIME, even though we used extremely powerful hardware, was still slow in providing explanations.

Human-Centered Design
The development of XAI systems should be done in a manner that emphasises human-centered design principles. Developers should conduct usability studies to understand how clinicians and healthcare professionals interact with XAI systems. This includes how different explanation styles such as visual, textual, and interactive explanations are perceived and understood by end-users, and how they impact decision-making, trust and user satisfaction in a clinical setting. Understanding clinician and patient perspectives on AI and ensuring that explanations are accessible to both experts and non-experts alike can help in creating systems that are more widely accepted and trusted by the public.

Ensemble Explainability
An aspect that we explored in this work was the use of multiple instances of a single explainability method to provide a more comprehensive and robust explanation. We propose that future research should focus on developing ensemble explainability methods that combine different explainability techniques, multiple instances of the same technique, or different methods to provide a more hollistic explanations that derives from different sources. This may provide more reliable and accurate feature importance scores and attributions, and enhance the overall interpretability and trustworthiness of the model's predictions.

Longitudinal Studies on XAI Impact
Conducting longitudinal studies to evaluate the impact of XAI systems on clinical practice and patient outcomes is crucial in understanding the long-term benefits and challenges of integrating XAI in healthcare. Such studies should measure how the use of XAI influences trust, accuracy and the adoption of AI-driven insights in clinical decision-making over time. They should also evaluate how XAI systems affect patient outcomes, healthcare costs, and the overall efficiency of the healthcare system.

Interdisciplinary Collaboration
Encouraging interdisciplinary collaboration in research between technologists, healthcare professionals, policymakers, and patients is crucial in developing XAI systems that are relevant, reliable and ethical. Future research should focus on fostering collaboration between different

stakeholders in the healthcare ecosystem to ensure that XAI systems are developed with considerations for patient safety, privacy, and regulatory compliance. Such collaboration is essential for addressing the complex challenges associated with XAI in healthcare and ensuring that research findings are practical, ethical, and aligned with the needs of the healthcare industry.

Improvement of Interpretability
There is a need for further research on improving the pre-existing explainability methods and developing new methods that can provide more accurate and reliable explanations for complex AI models such as deep learning. This needs to be done in a way that ensures that accuracy is not compromised in favour of interpretability. Investigating the trade-offs between how to interpret accuracy and interpretability in different explainability methods and developing methods that can provide a balance between the two is crucial in ensuring that AI systems are used responsibly and ethically in healthcare. This could involve developing new techniques for extracting meaningful explanations from black-box models or creating inherently interpretable models that are still powerful enough for healthcare applications.

Education and Training
Educating and training clinicians, healthcare professionals, and patients on the use of XAI systems is crucial in ensuring that these systems are used at their full potential. It is important that all stakeholders are aware of the benefits and limitations of XAI, as well as the ways of interpreting and acting upon the explanations provided by these systems.

Advancing XAI in healthcare requires a multifaceted research approach that addresses technical, ethical, and practical challenges. By focusing on domain-specific methods, improving interpretability, integrating systems into clinical workflows, and ensuring fairness and scalability, future research can help realize the full potential of XAI in improving patient outcomes and healthcare delivery. Collaboration across disciplines and active involvement of healthcare professionals and patients will be crucial in guiding this research toward meaningful and impactful solutions.

Chapter 9

Conclusion

9.1 Summary of key findings and contributions of the thesis

Our research and work has shed significant light upon the possible explainability solutions that researchers and developers can deploy to make AI systems more reliable as well as transparent and trustworthy. These methods allow for clinical use of AI systems in healthcare and other critical domains. In this section, we summarise the key findings and contributions of our thesis in the applications of XAI in healthcare.

9.1.1 CHA Tokens

Our initial approach of utilising phonological features inside the CHAT Transcript Protocol files as a means of explainability in the context of diagnosing dementia through spontenous speech was a significant step towards understanding how these features can be used to explain the model's decision. We proposed a significantly more intrepretable format for these features, which we call CHA tokens, which are a more human-readable format of the phonological features. Most of the research done in the field, tends to overlook and ignore these phonological features, ommitting them from the finalised, processed dataset that is used to train their models.

During our research, We found that certain phonological features provided significant evidence of possible dementia in the patient, which agrees with the existing literature. We also found that certain features were appearing as pairs in certain cases, and these pairings were at times evidence of dementia as well. We argue that these features can be further investigated to understand the underlying mechanisms of dementia and how they manifest in speech. During the process of providing explanations for the model's decision, we found that certain features were sometimes evidence of dementia, and sometimes not, which probes the question of how we may further refine each feature to provide more accurate explanations. A way to do this, is by using the CHA tokens and other features derived from the CHAT Transcript Protocol files, to provide clusters of features that are evidence of dementia, and those that are not through a unsupervised learning algorithm. After clustering we may inquire as to why certain features are evidence of dementia, and rename them to be more accurate in the explanation. We call these tokens, MVCHA tokens, from their multi-variate nature, which is an approach we propose for future research.

9.1.2 Ensemble Models

Our research into providing accurate predictions for the diagnosis of dementia also showed the importance and power of ensemble learning, by using multiple transformer models' outputs as inputs to a final classifier. We found that this approach provided a significant increase in all metrics, and by providing a larger variety of transformers, the metrics increased even further. We argue that through the use of ensemble learning, and diverse transformer and classifier selections, we can provide a more accurate and reliable model for NLP tasks, and specifically diagnosing dementia through spontaneous speech. This approach can be extended by using more features derived from the CHAT Transcript Protocol files, such as age, rate of speech, ethnicity and other features. These additional features can be concatenated with the transformer outputs, and used as inputs to the final classifier to provide a more hollistic approach to diagnosing dementia, though the amount of data required to train such a model would be significantly larger and more diverse, which was not available to us during our research.

In our work, we tried to provide an ensemble of classifiers trained on the afformentioned data, but found that, without extensive diversity in the transformers that were used on each singular classifier, the final ensemble did not provide any significant increase in the metrics, but mearly scored a similar result to the best classifier in the ensemble. We argue that this is due to the lack of different transformers entering each classifier, and that by providing a more diverse set of transformers, each different for each classifier, we may be able to provide a more accurate and reliable ensemble, even though our results were in fact state-of-the-art.

9.1.3 Ensemble Explainer

Our research into providing explanations sought to conduct a comparative analysis of three of the most popular explainability methods for transformer models, seeing as to how they performed in qualitative and quantitative metrics for the task of diagnosing dementia through spontaneous speech. We compared LIME, Anchors and Transformers-Interpret, and found that LIME scored the highest in all quantitative metrics through a survey we conducted with clinicians and healthcare professionals. We also found that Transformers-Interpret, should the clinical use case require a faster explanation, may also be used instead of LIME, as it scored better in the quantitative metrics for speed and fidelity. In our work, we propose a novel approach to providing explanations based on ensemble learning, which combines the outputs of the three LIME instances to provide a more hollistic and comprehensive explanation for the model's decision and the CHA tokens in the sequence of the patient's speech. This approach can be used with LIME and Transformers-Interpret, and we argue that by using a multitude of transformers, we can provide a more unbiased and reliable explanation on the attributions of the CHA tokens and how they contribute to the model's decision and sway the final prediction towards a diagnosis of dementia. A next step to provide further explanations, would be extending our explanations by also explaining the classifier's decision, and how each feature derived from the CHAT Transcript Protocol files, and the transformer outputs, contribute to the final prediction as a more hollistic approach to providing explanations. The final explanation would be a visual representation of all finding by the ensemble LIME or Transformers-Interpret instances, and the classifier's explanation of features.

9.1.4 User-Centric Design

In our research, we developed a user-centric design for using the ensemble model and explainer in a clinical setting through the use of a CLI and Web Application client, named DEMET, which is a framework for diagnosing dementia through spontaneous speech. The explainability methods we used in our research, were integrated into the DEMET framework, as we discussed

in the previous sections, to provide an easy-to-use and intuitive interface for clinicians and healthcare professionals to use. DEMET provides an API that can be intergrated into any existing system, and is our contribution to the COMFORTAGE project. A very important aspect of using these types of systems that provide explainability, is the fact that they are computationally expensive, and require a lot of resources in order to produce explanations in a timely manner without sacrificing the performance of the system. We urge future research to investigate how we may provide explanations in a more efficient manner. If these systems are to be used effectively in a clinical setting, they must be able to provide explanations efficiently, consistently and reliably, and we argue that this is a significant challenge that must be addressed in future research.

9.2 Implications for the healthcare industry

The integration of XAI in the healthcare industry has profound implications, affecting various aspects of healthcare delivery, clinical decision-making, patient outcomes and the broader healthcare ecosystem. In this section, we discuss the implications of our research findings for the healthcare industry.

9.2.1 Enhancing trust in AI technologies

XAI builds trust among healthcare professionals and patients by providing transparent and understandable explanations for AI-driven decisions. This increased trust can lead to broader acceptance and adoption of AI technologies in healthcare, enabling more widespread use of AI in clinical practice. By offering interpretable explanations for the model's decision, XAI supports clinicians and healthcare professionals in understanding the AI system's reasoning and making informed decisions based on the model's predictions. This approach can inhance diagnostic accuracy, improve patient outcomes and personalise treatment plans and patient care. Through the use of XAI, clinicians may better understand the underlying mechanisms of model predictions and decision making processes, leading to a more confident and evidence-based clinical decision. XAI can also help to identify possible biases and errors in model predictions, promoting fairness in healthcare delivery and reducing the risk of misdiagnosis or incorrect treatment. In this manner, AI decisions will be more accountable and will not disproportionately affect certain patient groups or demographics.

Through the use of XAI models, clinicians and healthcare professionals facilitate the integration of AI technologies into clinical practice by simultaneously complying with regulatory requirements and ethical standards that govern the use of AI in healthcare. The use of XAI can also help to address challenges in accountability, transparency and trustworthiness of AI-driven systems, potentially influencing legal frameworks around liability and accountability in healthcare. When using AI-driven solutions for critical healthcare tasks, certain liability and legal issues may arise, specifically in the case of misdiagnosis or incorrect treatment, which was influenced by the use AI. XAI can help to mitigate these risks by providing transparent and interpretable explanations.

9.2.2 Patient-centric approach

XAI has the potential to empower patients by providing them with more information and control over their healthcare decisions, by offering more intrepretable, understandable and accessible explanations for AI-driven insights and recommendations. Patients who can better understand the AI system's predictions and recommendations are more likely to trust the system and seek treatment and medical advice should they require it. This allows patients to be more actively

involved in their healthcare and treatment plans, which leads to significantly better patient satisfaction and surely better adherence to proposed treatment plans. XAI, incorporated into healthcare systems through the use of IoT devices, mobile applications, monitoring systems and EHR patient portals, can provide patients with real-time insights and explanations for their health data, enabling them to make informed decisions about their health and well-being and take proactive steps to improve their health outcomes.

As XAI becomes more and more integrated and widely adopted in healthcare, healthcare organisations and patients alike will be able to benefit from the cost efficiencies associated with AI-driven automation and decision support. By streamlining workflows, automating routine and tedious tasks, and providing real-time insights, recommendations and alerts on patient health conditions, XAI can contribute in a more cost-effective and efficient healthcare delivery system, ultimately leading to better financial outcomes for patients and driving costs down. This will also allow healthcare organisations to allocate resources more effectively, reduce administrative burden and improve overall operational efficiency.

9.2.3 Breaking new ground in Explainability

Our research has contributed to the field of XAI by proposing novel methods and approaches for enhancing the explainability of AI models in healthcare, specifically in the context of diagnosing dementia through spontaneous speech. By introducing CHA tokens and the ensemble explainer, we have demonstrated how XAI can be used to provide more interpretable, reliable and comprehensive explanations for transformer models' decisions without sacrificing performance. Our research findings have significant implications for the healthcare industry, enabling more accurate and transparent AI-driven diagnostics and decision-making processes. By integrating XAI into clinical practice, we can enhance trust in AI technologies, improve patient outcomes, empower patients and drive innovation in the healthcare industry. We believe that our work has opened new avenues for research and development in the field of XAI in healthcare, by providing an ensemble explanation through the use of multiple instances of different explainability methods, utilising the weighted averages of the attributions provided by each instance, thus providing a more comprehensive and hollistic explanation, eliminating single instance bias and providing reliability into how CHA tokens contribute to the model's decision.

9.2.4 Conclusion

We find ourselves at a critical juncture in the evolution of AI in healthcare, where the integration of XAI is transforming the way care is delivered. The implications of XAI in the healthcare industry are far-reaching, influencing trust, decision-making, equity, regulatory compliance, and patient empowerment. As XAI becomes more integrated into healthcare systems, it has the potential to transform how care is delivered, making AI-driven insights more accessible, transparent, and reliable. However, these advancements also bring challenges related to computational resources, accuracy of explanations, ethics, legality, and policy, which must be carefully managed to fully realize the benefits of XAI in healthcare. XAI has still a long way to go, and we are excited to see how the field evolves in the coming years.

9.3 Concluding remarks

In this thesis, we have explored the applications of XAI in healthcare, focusing on the diagnosis of dementia through spontaneous speech. We have proposed a novel approach to explainability, using CHA tokens to provide more interpretable explanations for transformer models. We have also demonstrated the power of ensemble learning in improving the accuracy of AI models, and proposed an ensemble explainer to provide more comprehensive and hollistic explanations, when it comes to phonological features found in patients' speech, for the model's decision. We have developed a user-centric design for using the ensemble model and explainer in a clinical setting, through the DEMET framework, which is our contribution to the COMFORTAGE project. Our research findings have significant implications for the healthcare industry, enhancing trust in AI technologies and enabling more widespread adoption of AI in clinical practice. We hope that our work will inspire further research in the field of XAI in healthcare, and contribute to the development of more reliable, transparent and trustworthy AI systems for clinical use. We are looking forward to further investigating the applications of XAI in healthcare, and exploring new methods and tools for enhancing the explainability of AI models. We believe that XAI has the potential to revolutionize healthcare delivery, improve patient outcomes, and empower patients to take control of their health and well-being. By integrating XAI into clinical practice, we can address key challenges in healthcare, such as diagnostic accuracy, patient empowerment, and regulatory compliance, while also unlocking new opportunities for innovation and collaboration in the healthcare industry. As we move forward, it is essential to continue exploring the field of XAI in healthcare, by conducting further research, developing new tools and methods, fostering collaboration between researchers, clinicians, and industry partners, and ensuring that these technologies are used ethically, responsibly, and equitably. By working together to advance the field of XAI in healthcare, we can create a more transparent, trustworthy, and patient-centric healthcare system that benefits all stakeholders and improves health outcomes for patients worldwide. It was a pleasure to work on this thesis, as part of the COMFORTAGE project, and we are excited to see how the future of XAI in healthcare unfolds. We will continue to work towards developing a more sophisticated and reliable AI system for diagnosing dementia through spontaneous speech, and we hope that our research will contribute to the broader field of XAI in healthcare.

Bibliography

[1] Rajkomar A et al. "Ensuring Fairness in Machine Learning to Advance Health Equity". In: (2018). URL: https://doi.org/10.7326/M18-1990.

[2] Dubey A Tiwari A. *Artificial intelligence and remote patient monitoring in US healthcare market: a literature review*. 2023.

[3] Alaa A. Abd-alrazaq et al. "An overview of the features of chatbots in mental health: A scoping review". In: *International Journal of Medical Informatics* 132 (2019), p. 103978. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2019.103978. URL: https://www.sciencedirect.com/science/article/pii/S1386505619307166.

[4] C.R. Aditya and M.B. Sanjay Pande. "Devising an interpretable calibrated scale to quantitatively assess the dementia stage of subjects with alzheimer's disease: A machine learning approach". In: *Informatics in Medicine Unlocked* 6 (2017), pp. 28–35. ISSN: 2352-9148. DOI: https://doi.org/10.1016/j.imu.2016.12.004. URL: https://www.sciencedirect.com/science/article/pii/S2352914816300491.

[5] Chirag Agarwal and Anh Nguyen. *Explaining image classifiers by removing input features using generative models*. 2020. arXiv: 1910.04256 [cs.LG].

[6] Simon Bin Akter, Sumya Akter, and Tanmoy Sarkar Pias. "Stroke Risk Prediction from Medical Survey Data: AI-Driven Risk Analysis with Insightful Feature Importance using Explainable AI (XAI)". In: *medRxiv* (2023). DOI: 10.1101/2023.11.17.23298646. eprint: https://www.medrxiv.org/content/early/2023/11/20/2023.11.17.23298646.full.pdf. URL: https://www.medrxiv.org/content/early/2023/11/20/2023.11.17.23298646.

[7] Robert F. Wolff et al. "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies". In: (2019). URL: https://doi.org/10.7326/M18-1376.

[8] Ashish Jith Sreejith Kumar et al. "Evaluation of Generative Adversarial Networks for High-Resolution Synthetic Image Generation of Circumpapillary Optical Coherence Tomography Images for Glaucoma". In: (2022). URL: https://doi.org/10.1001/jamaophthalmol.2022.3375.

[9] Das D et al. "An interpretable machine learning model for diagnosis of Alzheimer's disease". In: *PeerJ* (2019). URL: https://peerj.com/articles/6543/.

[10] Gretton Arthur et al. *Covariate Shift by Kernel Mean Matching*. 2009. URL: http://www.gatsby.ucl.ac.uk/~gretton/papers/covariateShiftChapter.pdf.

[11] Jahan S et al. "Explainable AI-based Alzheimer's prediction and management using multimodal data." In: *PLoS One* (2023). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10653516/.

[12] Jinsung Yoon et al. "EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records". In: (2023). URL: https://doi.org/10.1038/s41746-023-00888-7.

[13] Mahmud T et al. "An Explainable AI Paradigm for Alzheimer's Diagnosis Using Deep Transfer Learning." In: *Diagnostics (Basel)* (2024). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10855149/.

[14] Maximilian Alber et al. "iNNvestigate Neural Networks!" In: *Journal of Machine Learning Research* 20.93 (2019), pp. 1–8. URL: http://jmlr.org/papers/v20/18-540.html.

[15] Ahmed H. Alkenani et al. "Predicting Alzheimer's Disease from Spoken and Written Language Using Fusion-Based Stacked Generalization". In: *Journal of Biomedical Informatics* 118 (2021), p. 103803. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2021.103803. URL: https://www.sciencedirect.com/science/article/pii/S1532046421001325.

[16] Naomi S. Altman. "An introduction to kernel and nearest-neighbor nonparametric regression". In: (1992).

[17] Nicola Amoroso et al. "A Roadmap towards Breast Cancer Therapies Supported by Explainable Artificial Intelligence". In: *Applied Sciences* 11.11 (2021). ISSN: 2076-3417. DOI: 10.3390/app11114881. URL: https://www.mdpi.com/2076-3417/11/11/4881.

[18] Mugahed A. Al-Antari. "Artificial Intelligence for Medical Diagnostics-Existing and Future AI Technology!" In: (2023). URL: https://doi.org/10.3390/diagnostics13040688.

[19] Malik Arman, Farzan Muhammad, and Abbas Asad. *Explainable AI for Healthcare Decision Support Systems*. Nov. 2023.

[20] Vijay Arya et al. *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*. 2019. URL: https://arxiv.org/abs/1909.03012.

[21] Gupta Ashutosh and Singh Anand. "Healthcare 4.0: recent advancements and futuristic research directions". In: *Wireless Personal Communications* 129.2 (2023), pp. 933–952. DOI: 10.1007/s11277-022-10164-8. URL: https://doi.org/10.1007/s11277-022-10164-8.

[22] Sebastian Bach et al. *On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation*. 2015. URL: https://doi.org/10.1371/journal.pone.0130140.

[23] David Balduzzi et al. *The Shattered Gradients Problem: If resnets are the answer, then what is the question?* 2018. arXiv: 1702.08591 [cs.NE].

[24] Hubert Baniecki et al. "dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python". In: *Journal of Machine Learning Research* 22.214 (2021), pp. 1–7. URL: http://jmlr.org/papers/v22/20-1473.html.

[25] Rachel K. E. Bellamy et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oct. 2018. URL: https://arxiv.org/abs/1810.01943.

[26] Pronaya Bhattacharya et al. "BinDaaS: Blockchain-Based Deep-Learning as-a-Service in Healthcare 4.0 Applications". In: *IEEE Transactions on Network Science and Engineering* 8.2 (2021), pp. 1242–1255. DOI: 10.1109/TNSE.2019.2961932.

[27] Przemyslaw Biecek. "DALEX: Explainers for Complex Predictive Models in R". In: *Journal of Machine Learning Research* 19.84 (2018), pp. 1–5. URL: http://jmlr.org/papers/v19/18-416.html.

[28] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. ISBN: 9780367135591. URL: https://pbiecek.github.io/ema/.

[29] Garattini C et al. *Big Data Analytics, Infectious Diseases and Associated Ethical Impacts*. 2019. URL: https://pubmed.ncbi.nlm.nih.gov/31024785/.

[30] Zhou C.M. et al. *Constructing a Prediction Model for Difficult Intubation of Obese Patients Based on Machine Learning*. 2021.

[31] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. "Unintended Consequences of Machine Learning in Medicine". In: *JAMA* 318.6 (Aug. 2017), pp. 517–518. ISSN: 0098-7484. DOI: 10.1001/jama.2017.7797. eprint: https://jamanetwork.com/journals/jama/articlepdf/2645762/jama_cabitza_2017_vp_170094.pdf. URL: https://doi.org/10.1001/jama.2017.7797.

[32] Robert Challen et al. "Artificial intelligence, bias and clinical safety". In: *BMJ Quality & Safety* 28.3 (2019), pp. 231–237. ISSN: 2044-5415. DOI: 10.1136/bmjqs-2018-008370. eprint: https://qualitysafety.bmj.com/content/28/3/231.full.pdf. URL: https://qualitysafety.bmj.com/content/28/3/231.

[33] Pierre Chambon et al. *Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains*. 2022. arXiv: 2210.04133 [cs.CV].

[34] European Commission. "Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)". In: (2021). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[35] United States Congress. "Health Insurance Portability and Accountability Act of 1996". In: (1996). URL: https://www.hhs.gov/hipaa/index.html.

[36] Simon D'Alfonso. "AI in mental health". In: *Current Opinion in Psychology* 36 (2020). Cyberpsychology, pp. 112–117. ISSN: 2352-250X. DOI: https://doi.org/10.1016/j.copsyc.2020.04.005. URL: https://www.sciencedirect.com/science/article/pii/S2352250X2030049X.

[37] Yeboah Dacosta et al. *An Explainable and Statistically Validated Ensemble Clustering Model Applied to the Identification of Traumatic Brain Injury Subgroups*. 2020. DOI: 10.1109/ACCESS.2020.3027453.

[38] Luc Devroye, Laszlo Gyorfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996. URL: https://link.springer.com/book/10.1007/978-1-4612-0711-5.

[39] Natalia Díaz-Rodríguez et al. "EXplainable Neural-Symbolic Learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: The MonuMAI cultural heritage use case". In: *Information Fusion* 79 (Mar. 2022), pp. 58–83. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.09.022. URL: http://dx.doi.org/10.1016/j.inffus.2021.09.022.

[40] Carlo Dindorf et al. "Classification and Automated Interpretation of Spinal Posture Data Using a Pathology-Independent Classifier and Explainable Artificial Intelligence (XAI)". In: *Sensors* 21.18 (2021). ISSN: 1424-8220. DOI: 10.3390/s21186323. URL: https://www.mdpi.com/1424-8220/21/18/6323.

[41] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[42] Erik Edwards et al. "Multiscale System for Alzheimer's Dementia Recognition Through Spontaneous Speech". In: *Proc. Interspeech 2020*. 2020, pp. 2197–2201. DOI: 10.21437/Interspeech.2020-2781.

[43] Sean C. Ermer et al. *An Automated Algorithm Incorporating Poincaré Analysis Can Quantify the Severity of Opioid-Induced Ataxic Breathing*. 2020.

[44] Alireza Farrokhi et al. *Application of Internet of Things and artificial intelligence for smart fitness: A survey*. 2021. DOI: https://doi.org/10.1016/j.comnet.2021.107859. URL: https://www.sciencedirect.com/science/article/pii/S1389128621000360.

[45] Ruth C. Fong and Andrea Vedaldi. "Interpretable Explanations of Black Boxes by Meaningful Perturbation". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017. DOI: 10.1109/iccv.2017.371. URL: http://dx.doi.org/10.1109/ICCV.2017.371.

[46] Food and Drug Administration. "FDA Releases Artificial Intelligence/Machine Learning Action Plan". In: (2021). URL: https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan.

[47] Food and Drug Administration. "Proposed regulatory framework for modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD)". In: (2019). URL: https://apo.org.au/node/228371.

[48] International Medical Device Regulators Forum. ""Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations". In: (2014). URL: https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf.

[49] Jerome H. Friedman and Bogdan E. Popescu. "Predictive learning via rule ensembles". In: *The Annals of Applied Statistics* 2.3 (Sept. 2008). ISSN: 1932-6157. DOI: 10.1214/07-aoas148. URL: http://dx.doi.org/10.1214/07-AOAS148.

[50] Christopher Frye, Colin Rowat, and Ilya Feige. *Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability*. 2021. arXiv: 1910.06358 [stat.ML].

[51] Finlayson S. G. et al. "The Clinician and Dataset Shift in Artificial Intelligence". In: (2021). URL: https://doi.org/10.1056/NEJMc2104626.

[52] A.S.A. Garcez, K. Broda, and D.M. Gabbay. *Neural-Symbolic Learning Systems: Foundations and Applications*. Perspectives in Neural Computing. Springer London, 2002. ISBN: 9781852335120. URL: https://books.google.gr/books?id=6NZYVSOyD-UC.

[53] Alison K. Godbolt et al. "The Natural History of Alzheimer Disease: A Longitudinal Presymptomatic and Symptomatic Study of a Familial Cohort". In: *Archives of Neurology* 61.11 (Nov. 2004), pp. 1743–1748. ISSN: 0003-9942. DOI: 10.1001/archneur.61.11.1743. eprint: https://jamanetwork.com/journals/jamaneurology/articlepdf/787136/noc40077.pdf. URL: https://doi.org/10.1001/archneur.61.11.1743.

[54] Sabrina Goellner, Marina Tropmann-Frick, and Bostjan Brumen. *Responsible Artificial Intelligence: A Structured Literature Review*. 2024. arXiv: 2403.06910 [cs.AI].

[55] Alicja Gosiewska and Przemyslaw Biecek. *iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models*. Mar. 2019.

[56] Ben Green. "The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice". In: *Journal of Social Computing* 2.3 (Sept. 2021), pp. 209–225. ISSN: 2688-5255. DOI: 10.23919/jsc.2021.0018. URL: http://dx.doi.org/10.23919/JSC.2021.0018.

[57] David Gunning et al. "DARPA 's explainable AI (XAI) program: A retrospective". In: *Applied AI Letters* 2 (Dec. 2021). DOI: 10.1002/ail2.61.

[58] Tang H, Miri Rekavandi A, and Rooprai D et al. "Analysis and evaluation of explainable artificial intelligence on suicide risk assessment". In: (2023). DOI: https://doi.org/10.1038/s41598-024-53426-0.

[59] Esteva H. et al. "Neural Networks as a Prognostic Tool of Surgical Risk in Lung Resections". In: *Ann. Thorac. Surg.* (2002).

[60] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: (2016). arXiv: 1610.02413 [cs.LG].

[61] Zach Harned. *HIPAA, Medical AI, and Privacy*. 2023. URL: https://www.youtube.com/watch?v=P3KkCcFQDO8.

[62] Department of Health and Social Care. *Health Secretary announces £250 million investment in artificial intelligence*. 2019. URL: https://www.gov.%20uk/government/news/health-secretary-announces-250-million-investment-in-%20artificial-intelligence.

[63] Accuracy of Heart Rate Measurement with Wrist-Worn Wearable Devices in Various Skin Tones: a Systematic Review. "Koerber D. and Khan S. and Shamsheri T. and Kirubarajan A. and Mehta S." In: (2023). URL: https://doi.org/10.1007/s40615-022-01446-9.

[64] Lisa Anne Hendricks et al. *Generating Visual Explanations*. 2016. arXiv: 1603.08507 [cs.CV].

[65] Krebs HI and Volpe BT. *Rehabilitation robotics*. 2013.

[66] Robert R Hoffman et al. "Metrics for Explainable AI Challenges and Prospects". In: (2018).

[67] Sarah Holland et al. *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards*. 2018. arXiv: 1805.03677 [cs.DB].

[68] Robert C. Holte. *Very Simple Classification Rules Perform Well on Most Commonly Used Datasets*. 1993. URL: https://doi.org/10.1023/A:1022631118932.

[69] Andreas Holzinger et al. "Explainable AI Methods - A Brief Overview". In: *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Ed. by Andreas Holzinger et al. Cham: Springer International Publishing, 2022, pp. 13–38. ISBN: 978-3-031-04083-2. DOI: 10.1007/978-3-031-04083-2_2. URL: https://doi.org/10.1007/978-3-031-04083-2_2.

[70] Qiang Huang et al. *GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks*. 2020. arXiv: 2001.06216 [cs.LG].

[71] IBM. *IBM Watson OpenScale*. 2019. URL: https://www.ibm.com/downloads/cas/BK0OK0EA.

[72] ICO and The Alan Turing Institute. "Explaining Decisions Made with AI: Draft Guidance for Consultation—Part 1: The Basics of Explaining AI". In: (2019).

[73] Muhammad Idrees and Ayesha Sohail. *Explainable machine learning of the breast cancer staging for designing smart biomarker sensors*. 2022. DOI: https://doi.org/10.1016/j.sintl.2022.100202. URL: https://www.sciencedirect.com/science/article/pii/S266635112200047X.

[74] Loukas Ilias and Dimitris Askounis. "Explainable Identification of Dementia From Transcripts Using Transformer Networks". In: *IEEE Journal of Biomedical and Health Informatics* 26.8 (2022), pp. 4153–4164. DOI: 10.1109/JBHI.2022.3172479.

[75] Peng J et al. “An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients”. In: (2021).

[76] Anders Christopher J. et al. “Software for Dataset-wide XAI: From Local Explanations to Global Insights with Zennit, CoRelAy, and ViRelAy”. In: *CoRR* abs/2106.13200 (2021).

[77] Hwang JJ et al. *An overview of deep learning in the field of dentistry*. 2019.

[78] Sudhanshu Joshi et al. “Modeling Conceptual Framework for Implementing Barriers of AI in Public Healthcare for Improving Operational Excellence: Experiences from Developing Countries”. In: *Sustainability* 14.18 (2022). ISSN: 2071-1050. DOI: 10.3390/su141811698. URL: https://www.mdpi.com/2071-1050/14/18/11698.

[79] Ortiz JR et al. *Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends*. 2011. URL: https://doi.org/10.1371/journal.pone.0018687.

[80] Elma et al. Kerz. “Toward explainable AI (XAI) for mental health detection based on language behavior.” In: *Frontiers in psychiatry* (). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10748510/.

[81] Been Kim et al. *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)*. 2018. arXiv: 1711.11279 [stat.ML].

[82] Junghyun Koo et al. “Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer’s Dementia Recognition”. In: *Proc. Interspeech 2020*. 2020, pp. 2217–2221. DOI: 10.21437/Interspeech.2020-3153.

[83] Abhinav Kumar, Jyoti Kumari, and Jiesth Pradhan. “Explainable Deep Learning for Mental Health Detection from English and Arabic Social Media Posts”. In: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (Nov. 2023). Just Accepted. ISSN: 2375-4699. DOI: 10.1145/3632949. URL: https://doi.org/10.1145/3632949.

[84] Rodrigues L et al. *Drug Repurposing for COVID-19: A Review and a Novel Strategy to Identify New Targets and Potential Drug Candidates*. 2022.

[85] Wang L, Lin ZQ, and Wong. *A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images*. 2020.

[86] J. Laaksonen and E. Oja. “Classification with learning k-nearest neighbors”. In: *Proceedings of International Conference on Neural Networks (ICNN’96)*. Vol. 3. 1996, 1480–1483 vol.3. DOI: 10.1109/ICNN.1996.549118.

[87] Alyssa M. Lanzi et al. “DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses”. In: *American Journal of Speech-Language Pathology* 32.2 (2023), pp. 426–438. DOI: 10.1044/2022_AJSLP-22-00281. eprint: https://pubs.asha.org/doi/pdf/10.1044/2022_AJSLP-22-00281. URL: https://pubs.asha.org/doi/abs/10.1044/2022_AJSLP-22-00281.

[88] Michael van Lent, William Fisher, and Michael Mancuso. “An explainable artificial intelligence system for small-unit tactical behavior”. In: *Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence*. IAAI’04. San Jose, California: AAAI Press, 2004, pp. 900–907. ISBN: 0262511835.

[89] Benjamin Letham et al. “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model”. In: *The Annals of Applied Statistics* 9.3 (Sept. 2015). ISSN: 1932-6157. DOI: 10.1214/15-aoas848. URL: http://dx.doi.org/10.1214/15-AOAS848.

[90] Changye Li et al. "Useful blunders: Can automated speech recognition errors improve downstream dementia classification?" In: *Journal of Biomedical Informatics* 150 (2024), p. 104598. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2024.104598. URL: https://www.sciencedirect.com/science/article/pii/S1532046424000169.

[91] Wen-Ling Liao and Fuu-Jen Tsai. "Personalized medicine: A paradigm shift in healthcare". In: *BioMedicine* 3.2 (2013), pp. 66–72. ISSN: 2211-8020. DOI: https://doi.org/10.1016/j.biomed.2012.12.005. URL: https://www.sciencedirect.com/science/article/pii/S2211802012001027.

[92] Meng Liu et al. "DIG: A Turnkey Library for Diving into Graph Deep Learning Research". In: *Journal of Machine Learning Research* 22.240 (2021), pp. 1–9. URL: http://jmlr.org/papers/v22/21-0343.html.

[93] Xianjing Liu et al. "Predicting skin cancer risk from facial images with an explainable artificial intelligence (XAI) based approach: a proof-of-concept study". In: (2024). DOI: https://doi.org/10.1016/j.eclinm.2024.102550.

[94] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI].

[95] Scott M. Lundberg et al. *Explainable AI for Trees: From Local Explanations to Global Understanding*. 2019. arXiv: 1905.04610 [cs.LG].

[96] Schinle M et al. "Explainable Artificial Intelligence in Ambulatory Digital Dementia Screenings". In: *Stud Health Technol Inform* (2022). URL: https://pubmed.ncbi.nlm.nih.gov/35612031/.

[97] Danis M. and Solomon M. "Providers, payers, the community, and patients are all obliged to get patient Activation and engagement ethically right". In: (2013). URL: https://doi.org/10.1377/hlthaff.2012.1081.

[98] Sugiyama M., Suzuki T., and Kanamori T. *Covariate Shift Adaptation by Importance Weighted Cross Validation*. 2007. URL: https://doi.org/10.1109/TPAMI.2007.70726.

[99] Turilli M. and Floridi L. "The ethics of information transparency". In: (2008). URL: https://doi.org/10.1007/s10676-009-9187-9.

[100] Yan M. et al. "Observability and its impact on differential bias for clinical prediction models". In: (2022). URL: https://doi.org/10.1093/jamia/ocac019.

[101] Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. "An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery". In: *Computers in Biology and Medicine* 126 (2020), p. 104041. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2020.104041. URL: https://www.sciencedirect.com/science/article/pii/S0010482520303723.

[102] Anshu Malhotra and Rajni Jindal. "XAI Transformer based Approach for Interpreting Depressed and Suicidal User Behavior on Online Social Networks". In: *Cognitive Systems Research* 84 (2024), p. 101186. ISSN: 1389-0417. DOI: https://doi.org/10.1016/j.cogsys.2023.101186. URL: https://www.sciencedirect.com/science/article/pii/S1389041723001201.

[103] Jiayuan Mao et al. *The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision*. 2019. arXiv: 1904.12584 [cs.CV].

[104] Andrea Martani, David Shaw, and Bernice Simone Elger. "Stay fit or get bit – ethical issues in sharing health data with insurers' apps". In: *Swiss Medical Weekly* 149.2526 (2019), w20089. DOI: 10.4414/smw.2019.20089. URL: https://smw.ch/index.php/smw/article/view/2641.

[105] Dastan Maulud and Adnan Mohsin Abdulazeez. "A Review on Linear Regression Comprehensive in Machine Learning". In: *Journal of Applied Science and Technology Trends* 1 (Dec. 2020), pp. 140–147. DOI: 10.38094/jastt1457.

[106] Aghamohammadi Mehrdad et al. *Predicting Heart Attack Through Explainable Artificial Intelligence*. June 2019. DOI: 10.1007/978-3-030-22741-8_45.

[107] Anna Meldo et al. *The natural language explanation algorithms for the lung cancer computer-aided diagnosis system*. 2020. DOI: https://doi.org/10.1016/j.artmed.2020.101952. URL: https://www.sciencedirect.com/science/article/pii/S0933365720303900.

[108] Bahman Mirheidari et al. "Spoken language-based automatic cognitive assessment of stroke survivors". In: *Language and Health* (2024). ISSN: 2949-9038. DOI: https://doi.org/10.1016/j.laheal.2024.01.001. URL: https://www.sciencedirect.com/science/article/pii/S2949903824000010.

[109] Brent Mittelstadt et al. *The ethics of algorithms: Mapping the debate*. 2016. URL: https://doi.org/10.1177/2053951716679679.

[110] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022. URL: https://christophm.github.io/interpretable-ml-book.

[111] Grégoire Montavon et al. "Explaining nonlinear classification decisions with deep Taylor decomposition". In: *Pattern Recognition* 65 (May 2017), pp. 211–222. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2016.11.008. URL: http://dx.doi.org/10.1016/j.patcog.2016.11.008.

[112] Jessica Morley et al. "The ethics of AI in health care: A mapping review". In: *Social Science and Medicine* 260 (2020), p. 113172. ISSN: 0277-9536. DOI: https://doi.org/10.1016/j.socscimed.2020.113172. URL: https://www.sciencedirect.com/science/article/pii/S0277953620303919.

[113] Ahmad Wisnu Mulyadi et al. "Estimating explainable Alzheimer's disease likelihood map via clinically-guided prototype learning". In: *NeuroImage* 273 (2023), p. 120073. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2023.120073. URL: https://www.sciencedirect.com/science/article/pii/S1053811923002197.

[114] Norori N. et al. "Addressing bias in big data and AI for health care: A call for open science." In: (2021). URL: https://doi.org/10.1016/j.patter.2021.100347.

[115] United Nations. "Universal Declaration of Human Rights". In: (2022). URL: https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[116] An-phi Nguyen and María Rodríguez Martínez. "On quantitative aspects of model interpretability". In: (2020). arXiv: 2007.07584 [cs.LG].

[117] Ezio Di Nucci. "Should we be afraid of medical AI?" In: *Journal of Medical Ethics* 45.8 (2019), pp. 556–558. ISSN: 0306-6800. DOI: 10.1136/medethics-2018-105281. eprint: https://jme.bmj.com/content/45/8/556.full.pdf. URL: https://jme.bmj.com/content/45/8/556.

[118] David Ortiz-Perez et al. "A Deep Learning-Based Multimodal Architecture to predict Signs of Dementia". In: *Neurocomputing* 548 (2023), p. 126413. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.126413. URL: https://www.sciencedirect.com/science/article/pii/S0925231223005362.

[119] Amisha Malik P, Pathania M, and Rathaur VK. *Overview of artificial intelligence in medicine*. 2019.

[120] Joanne Peng, Kuk Lee, and Gary Ingersoll. "An Introduction to Logistic Regression Analysis and Reporting". In: *Journal of Educational Research - J EDUC RES* 96 (Sept. 2002), pp. 3–14. DOI: 10.1080/00220670209598786.

[121] Agrawal Prashansa. *Artificial Intelligence in Drug Discovery and Development*. Jan. 2018. DOI: 10.4172/2329-6887.1000e173.

[122] Milken Institute School of Public Health. *Equity vs Equality: What's the Difference?* URL: https://onlinepublichealth.gwu.edu/resources/equity-vs-equality/.

[123] Jennifer L. Quon et al. *Artificial intelligence for automatic cerebral ventricle segmentation and volume calculation: a clinical tool for the evaluation of pediatric hydrocephalus*. 2021. DOI: 10.3171/2020.6.PEDS20251. URL: https://thejns.org/pediatrics/view/journals/j-neurosurg-pediatr/27/2/article-p131.xml.

[124] Daneshjou R. et al. "Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms A Scoping Review". In: (2021). URL: www.doi.org/10.1001/jamadermatol.2021.3129.

[125] Zech J. R. et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study". In: (2018). URL: https://doi.org/10.1371/journal.pmed.1002683.

[126] Nickalus Redell. *Shapley Decomposition of R-Squared in Machine Learning Models*. 2019. arXiv: 1908.09718 [stat.ME].

[127] Google Research. "Fairness Indicators: Scalable Infrastructure for Fair ML Systems". In: (2019). URL: https://blog.research.google/2019/12/fairness-indicators-scalable.html.

[128] Google Research. "What-If Tool: Codeless Probing of Machine Learning Models". In: (2019). URL: https://pair-code.github.io/what-if-tool/.

[129] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. arXiv: 1602.04938 [cs.LG].

[130] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: high-precision model-agnostic explanations". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018. ISBN: 978-1-57735-800-8.

[131] Julian Risch and Ralf Krestel. "Bagging BERT Models for Robust Aggression Identification". English. In: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. Ed. by Ritesh Kumar et al. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 55–61. ISBN: 979-10-95546-56-6. URL: https://aclanthology.org/2020.trac-1.9.

[132] Marko Robnik-Šikonja and Igor Kononenko. "Explaining Classifications For Individual Instances". In: *IEEE Transactions on Knowledge and Data Engineering* 20.5 (2008), pp. 589–600. DOI: 10.1109/TKDE.2007.190734.

[133] Lior Rokach and Oded Maimon. "Decision Trees". In: vol. 6. Jan. 2005, pp. 165–192. DOI: 10.1007/0-387-25465-X_9.

[134] Matteo Rucco, Giovanna Viticchi, and Lorenzo Falsetti. "Towards Personalized Diagnosis of Glioblastoma in Fluid-Attenuated Inversion Recovery (FLAIR) by Topological Interpretable Machine Learning". In: *Mathematics* 8.5 (2020). ISSN: 2227-7390. URL: https://www.mdpi.com/2227-7390/8/5/770.

[135] Pati S., Baid U., and Edwards B. et al. "Federated learning enables big data for rare cancer boundary detection". In: (2022). URL: https://doi.org/10.1038/s41467-022-33407-5.

[136] Stivaros S.M., Radon M.R., and Mileva R. et al. "Quantification of structural changes in the corpus callosumin children with profound hypoxic–ischaemic brain injury". In: (2016).

[137] Pedro Saleiro et al. "Aequitas: A Bias and Fairness Audit Toolkit". In: (2019). arXiv: 1811.05577 [cs.LG].

[138] Wojciech Samek et al. "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications". In: *Proceedings of the IEEE* 109.3 (2021), pp. 247–278. DOI: 10.1109/JPROC.2021.3060483.

[139] Deepti Saraswat et al. "Explainable AI for Healthcare 5.0: Opportunities and Challenges". In: *IEEE Access* 10 (2022), pp. 84486–84517. DOI: 10.1109/ACCESS.2022.3197671.

[140] Salih Sarp et al. "The Enlightening Role of Explainable Artificial Intelligence in Chronic Wound Classification". In: *Electronics* 10.12 (2021). ISSN: 2079-9292. DOI: 10.3390/electronics10121406. URL: https://www.mdpi.com/2079-9292/10/12/1406.

[141] Peter Schulam and Suchi Saria. *Reliable Decision Support using Counterfactual Models*. 2018. arXiv: 1703.10651 [stat.ML].

[142] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[143] Lee SH et al. *Exercise with a wearable hip-assist robot improved physical function and walking efficiency in older adults*. 2023.

[144] Makubhai Shahin, Pathak Ganesh, and Chandre Pankaj. "Predicting lung cancer risk using explainable artificial intelligence". In: *Bulletin of Electrical Engineering and Informatics* 13 (Apr. 2024), pp. 1276–1285. DOI: 10.11591/eei.v13i2.6280.

[145] Lloyd S. Shapley. *A value for n-person games*. 1953. arXiv: https://www.rand.org/pubs/papers/P0295.html [cs.GT].

[146] Genevieve Smith and Ishita Rustagi. "Mitigating Bias in Artificial Intelligence An Equity Fluent Leadership Playbook". In: (2020). URL: https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf.

[147] Sigrid Sterckx et al. ""You hoped we would sleep walk into accepting the collection of our data": controversies surrounding the UK care.data scheme and their wider relevance for biomedical research". In: (2015). URL: https://doi.org/10.1007/s11019-015-9661-6.

[148] Adarsh Subbaswamy. "From development to deployment: dataset shift, causality, and shift-stable models in health AI". In: (2020). URL: https://doi.org/10.1093/biostatistics/kxz041.

[149] Adarsh Subbaswamy and Suchi Saria. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms". In: (2018). arXiv: 1808.03253 [stat.ML].

[150] Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. "Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport". In: (2019). arXiv: 1812.04597 [stat.ML].

[151] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG].

[152] Mariarosaria Taddeo and Luciano Floridi. "How AI can be a force for good". In: *Science* 361.6404 (2018), pp. 751–752. DOI: 10.1126/science.aat5991. eprint: https://www.science.org/doi/pdf/10.1126/science.aat5991. URL: https://www.science.org/doi/abs/10.1126/science.aat5991.

[153] Weimin Tan et al. "The use of explainable artificial intelligence to explore types of fenestral otosclerosis misdiagnosed when using temporal bone high-resolution computed tomography". In: *Annals of Translational Medicine* 9.12 (2021). ISSN: 2305-5847. URL: https://atm.amegroups.org/article/view/72594.

[154] Bowen Tang, John Ewalt, and Ho-Leung Ng. "Generative AI Models for Drug Discovery". In: *Biophysical and Computational Tools in Drug Discovery*. Ed. by Anil Kumar Saxena. Cham: Springer International Publishing, 2021, pp. 221–243. ISBN: 978-3-030-85281-8. DOI: 10.1007/7355_2021_124. URL: https://doi.org/10.1007/7355_2021_124.

[155] Yeldar Toleubay et al. *Utterance Classification with Logical Neural Network: Explainable AI for Mental Disorder Diagnosis*. 2023. arXiv: 2306.03902 [cs.CL].

[156] Eric J Topol. "High-performance medicine: the convergence of human and artificial intelligence". In: *Nature medicine* 25.1 (2019), pp. 44–56. URL: https://www.nature.com/articles/s41591-018-0300-7.

[157] Bellini V., Guzzon M., and Bigliardi B. "Artificial Intelligence: A New Tool in Operating Room Management. Role of Machine Learning Models in Operating Room Optimization." In: *J Med Syst* (2020).

[158] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[159] Vikramkumar, Vijaykumar B, and Trilochan. *Bayes and Naive Bayes Classifier*. 2014. arXiv: 1404.0933 [cs.LG].

[160] Giulia Vilone and Luca Longo. "Explainable Artificial Intelligence: a Systematic Review". In: *CoRR* abs/2006.00093 (2020). arXiv: 2006.00093. URL: https://arxiv.org/abs/2006.00093.

[161] Kerstin N Vokinger, Stefan Feuerriegel, and Aaron S Kesselheim. "Mitigating bias in machine learning for medicine". In: (2021). URL: https://doi.org/10.1038/s43856-021-00028-w.

[162] Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. *Shapley Flow: A Graph-based Approach to Interpreting Model Predictions*. 2021. arXiv: 2010.14592 [cs.LG].

[163] Hilde Weerts et al. *Fairlearn: Assessing and Improving Fairness of AI Systems*. 2023. arXiv: 2303.16626 [cs.LG].

[164] Ronald J. Williams. *Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning*. 1992. eprint: https://doi.org/10.1007/BF00992696.

[165] Yao X et al. "Artificial Intelligence–Enabled Electrocardiograms for Identification of Patients with Low Ejection Fraction: A Pragmatic, Randomized Clinical Trial". In: *Nat. Med.* 27 (2021).

[166] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. *Scalable Bayesian Rule Lists*. 2017. arXiv: 1602.08610 [cs.AI].

[167] Hao Yuan et al. “Explainability in Graph Neural Networks: A Taxonomic Survey”. In: *arXiv preprint arXiv:2012.15445* (2020).

[168] Hao Yuan et al. “XGNN: Towards Model-Level Explanations of Graph Neural Networks”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*. KDD ’20. ACM, Aug. 2020. DOI: 10.1145/3394486.3403085. URL: http://dx.doi.org/10.1145/3394486.3403085.

[169] Xiangxiang Zeng et al. “Deep generative molecular design reshapes drug discovery”. In: *Cell Reports Medicine* 3.12 (2022).

[170] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. “Mitigating Unwanted Biases with Adversarial Learning”. In: (2018). arXiv: 1801.07593 [cs.LG].

[171] Luxia Zhang et al. “Big data and medical research in China”. In: *BMJ* 360 (2018). ISSN: 0959-8138. DOI: 10.1136/bmj.j5910. eprint: https://www.bmj.com/content/360/bmj.j5910.full.pdf. URL: https://www.bmj.com/content/360/bmj.j5910.

[172] Jianlong Zhou and Fang Chen et al. “Physiological Indicators for User Trust in Machine Learning with Influence Enhanced Fact-Checking”. In: (2019).

[173] Jianlong Zhou et al. *Correlation for user confidence in predictive decision making*. Launceston, Tasmania, Australia, 2016. DOI: 10.1145/3010915.3011004. URL: https://doi.org/10.1145/3010915.3011004.

[174] Jianlong Zhou et al. “Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics”. In: *Electronics* 10.5 (2021). ISSN: 2079-9292. DOI: 10.3390/electronics10050593. URL: https://www.mdpi.com/2079-9292/10/5/593.

[175] Attia ZI et al. “Screening for Cardiac Contractile Dysfunction Using an Artificial Intelligence–Enabled Electrocardiogram”. In: *Nat. Med.* 25.1 (2019).

[176] Obermeyer Ziad et al. *Dissecting racial bias in an algorithm used to manage the health of populations*. 2019. URL: https://www.science.org/doi/full/10.1126/science.aax2342.

[177] Luisa M Zintgraf et al. *Visualizing Deep Neural Network Decisions: Prediction Difference Analysis*. 2017. arXiv: 1702.04595 [cs.CV].

[178] Maryam Zolnoori, Ali Zolnour, and Maxim Topaz. “ADscreen: A speech processing-based screening system for automatic identification of patients with Alzheimer’s disease and related dementia”. In: *Artificial Intelligence in Medicine* 143 (2023), p. 102624. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2023.102624. URL: https://www.sciencedirect.com/science/article/pii/S0933365723001380.

Appendix A: Questionnaire

The following pages contain the questionnaire that was used during the research and evaluation of the explainability methods used in this thesis. The questionnaire was designed to gather feedback from medical professionals and clinicians with experience in the field of dementia and neurodegenerative diseases. We contacted a total of 20 professionals, including neurologists and psychiatrists and a total of 10 large medical institutions in Greece, England, the United States, Canada and Australia, and asked them to participate in the evaluation of the methods used in this thesis. The questionnaire was distributed and collected via email and was filled out by 3 participants, all of whom have experience in the field of dementia. Due to very low participation in the survey, we were unable to provide a comprehensive analysis of the results, but we were able to draw some conclusions from the feedback provided by the participants. The questionnaire is split into three sections, each containing questions about the validity of the CHA tokens, the explainability of the methods used in this thesis, and the overall usefulness of the methods in a clinical setting. The first section includes questions about the validity of the CHA tokens and their ability to provide evidence of dementia in a patient's speech. For each token used in the training of our models, the questionnaire contains a marker that asks the participant to rate the token's validity as evidence of dementia on a scale of 1 to 5, 1 meaning that the token provides no evidence for such a claim and 5 meaning that the token provides strong evidence of dementia. The second section provides two distinct examples, one of a patient with dementia and one of a patient without dementia, and asks the participant to rate the explainability of the model's visualizations for each example in regards to how informative they are in explaining the model's decision. The third section contains questions on the trustworthiness of the model's predictions and whether these explainable AI methods could be used in clinical practice. In regards to the use of the CHA tokens, the overall sentiment from the participants was that some tokens could possibly provide evidence of dementia, namely those that are related to repetition, non-completion of words or sentences, short to long pauses and trailing off, but highlighted that the validity of these tokens is highly dependent on the context in which they are used, which means that further investigation is in order. In terms of the explainability of the model's visualizations, the participants found LIME to be the most informative and easy to understand method, followed by Transformers-Interpret and Anchor. In terms of the trustworthiness of the model's predictions, the participants were generally skeptical, with most participants stating that the diagnosis of dementia is a multi-faceted and complicated process that requires constant monitoring and evaluation by medical professionals and could not be solely based on spontaneous speech. The participants do see potential in the use of these models in a supplementary setting, used by citizens who are worried about their cognitive health but are tentative to visit a clinician.

Explainable AI Applications for the improvement of quality of life in Dementia patients

This survey aims to assist in the understanding of how AI applications can be used to improve medical practices and introduce automations in the healthcare sector, specifically in the treatment of dementia using phonological features of spontaneous speech as tokens for training large AI models. These tokens can be used to introduce strong explainability, the ability that a model has to explain a certain decision that it made, in these models, and thus allow for trust and transparency. The survey is designed to extract information from medical professionals who have experience in the treatment of dementia patients. Three different AI models where used in this study, BERT, RoBERTa, and DistilBERT. Three different explainability techniques were used to explain the decisions of these models, LIME, Transformer-Interpret, and ANCHOR.
Please answer the following questions to the best of your ability. Your responses and personal information will be kept confidential and will be used for research purposes only. Thank you for your time.

About you

Please provide the following information:

1. **Your name**: _____
2. **Title**: _____
3. **Years of Experience**: _____
4. **Contact Information(e-mail)**: _____

Phonological Features of Dementia

Please evaluate the following phonological features as to how they provide strong evidence for the presence of dementia in a patient. Assume that the patient was interviewed and the following features were observed in their speech.

5a. **[CHA REPETITION] The patient is repeating a word**
Non Dementia □—□—□—□—□ Dementia

5b. **[CHA RETRACING] The patient starts to say something, stops, repeats the basic phrase, changes the syntax but maintains the same idea**
Non Dementia □—□—□—□—□ Dementia

5c. **[CHA SHORT PAUSE] The patient pauses for a short duration**
Non Dementia □—□—□—□—□ Dementia

5d. **[CHA MEDIUM PAUSE] The patient pauses for a medium duration**
Non Dementia □—□—□—□—□ Dementia

5e. **[CHA LONG PAUSE] The patient pauses for a long duration**
Non Dementia □—□—□—□—□ Dementia

5f. **[CHA TRAILING OFF] The patient's attention drifts off and the sentence is left incomplete** Non Dementia □—□—□—□—□ Dementia

5g. **[CHA PHONOLOGICAL FRAGMENT] The patient's phonological material differs between words** Non Dementia □—□—□—□—□ Dementia

5h. **[CHA FILLER] The patient uses fillers such as 'uh', 'um', 'you know' etc.**
Non Dementia □—□—□—□—□ Dementia

5i. **[CHA NON COMPLETION OF WORD] The patient does not complete a word**
Non Dementia □—□—□—□—□ Dementia

5j. **[CHA LAUGHS] The patient laughs during the conversation**
Non Dementia □—□—□—□—□ Dementia

5k. **[CHA SIGHS] The patient sighs during the conversation**
Non Dementia □—□—□—□—□ Dementia

5l. **[CHA TRAILING OFF QUESTION] The patient asks a question but does not complete it** Non Dementia □—□—□—□—□ Dementia

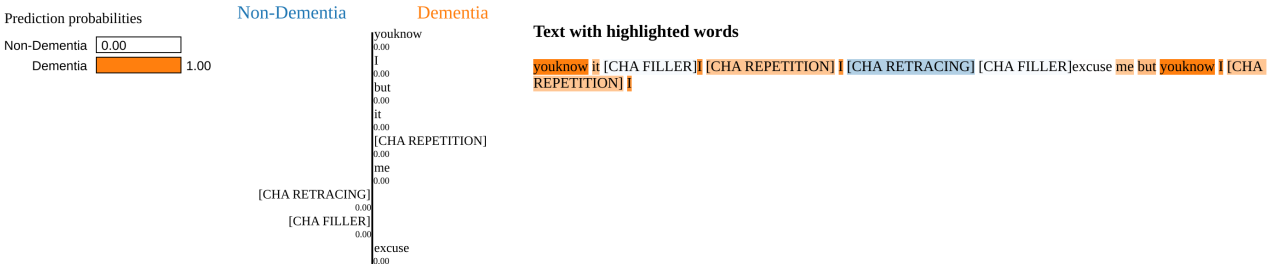
5m. **[CHA INTERRUPTION] The patient interrupts the interviewer**
Non Dementia □—□—□—□—□ Dementia

5n. **[CHA INTERRUPTION OF QUESTION] The patient interrupts the interviewer while they are asking a question** Non Dementia □—□—□—□—□ Dementia

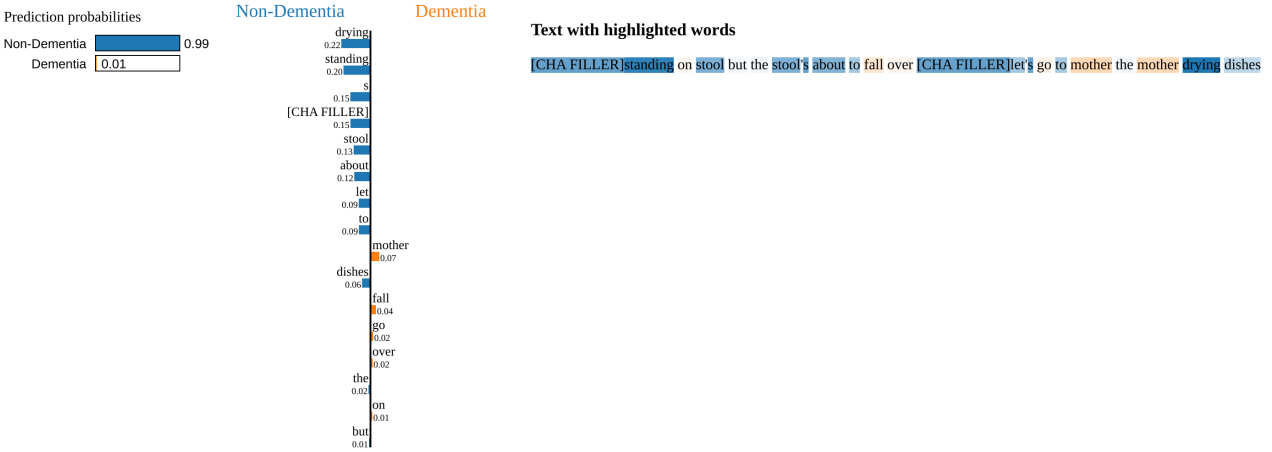
BERT Explanations

Please evaluate the following model explanations as to how informative they are in explaining the model's decision.

6a. Using LIME and predicting Dementia Not Informative ☐—☐—☐—☐—☐ Informative



6b. Using LIME and predicting Non Dementia Not Informative ☐—☐—☐—☐—☐ Informative



7a. Using T-I and predicting Dementia Not Informative ☐—☐—☐—☐—☐ Informative

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_1 (0.99)	LABEL_1	0.16	[CLS] you ##k ##no ##w it [CHA FILLER] I [CHA REPETITION] I [CHA RETRACING] [CHA FILLER] excuse me but you ##k ##no ##w I [CHA REPETITION] I [SEP]

7b. Using T-I and predicting Non Dementia Not Informative ☐—☐—☐—☐—☐ Informative

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	LABEL_0 (0.98)	LABEL_0	3.16	[CLS] [CHA FILLER] standing on stool but the stool ' s about to fall over [CHA FILLER] let ' s go to mother the mother drying dishes [SEP]

8a. Using ANCHOR and predicting Dementia Not Informative □—□—□—□—□ Informative

```
Model Predicts Dementia

Examples where anchor applies and model predicts Dementia:
youknow UNK [ CHA UNK UNK UNK REPETITION UNK I UNK UNK UNK UNK UNK UNK FILLER]excuse me but youknow I [ CHA REPETITION ] UNK
youknow UNK UNK UNK UNK [ CHA REPETITION ] I [ CHA UNK ] UNK CHA FILLER]excuse me but youknow I [ UNK UNK ] I
youknow it UNK UNK FILLER]I [ UNK REPETITION ] UNK UNK CHA RETRACING UNK [ CHA FILLER]excuse me but youknow UNK UNK CHA UNK ] I
UNK it UNK CHA UNK [ UNK REPETITION ] I UNK CHA UNK ] [ CHA FILLER]excuse me but UNK I [ CHA UNK ] I
youknow it [ CHA UNK UNK UNK UNK UNK UNK UNK CHA RETRACING UNK UNK CHA UNK me but youknow I [ UNK REPETITION UNK UNK
youknow UNK UNK CHA FILLER]I UNK CHA UNK ] UNK [ UNK UNK ] [ UNK UNK me UNK UNK I [ CHA UNK ] I
youknow UNK UNK CHA FILLER]I [ UNK UNK ] UNK UNK CHA RETRACING ] [ CHA UNK me UNK youknow I [ CHA REPETITION UNK UNK
UNK UNK UNK UNK FILLER]I UNK UNK UNK ] UNK UNK UNK UNK UNK UNK UNK FILLER]excuse me but UNK I UNK UNK UNK ] UNK
UNK it [ CHA UNK [ CHA UNK ] I [ CHA UNK ] [ UNK UNK me UNK youknow I UNK CHA UNK UNK UNK
UNK it UNK CHA FILLER]I UNK UNK REPETITION ] I UNK UNK UNK ] UNK UNK UNK me but youknow UNK [ CHA REPETITION ] I

Examples where anchor applies and model predicts Non-Dementia:
```

8b. Using ANCHOR and predicting Non Dementia Not Informative □—□—□—□—□ Informative

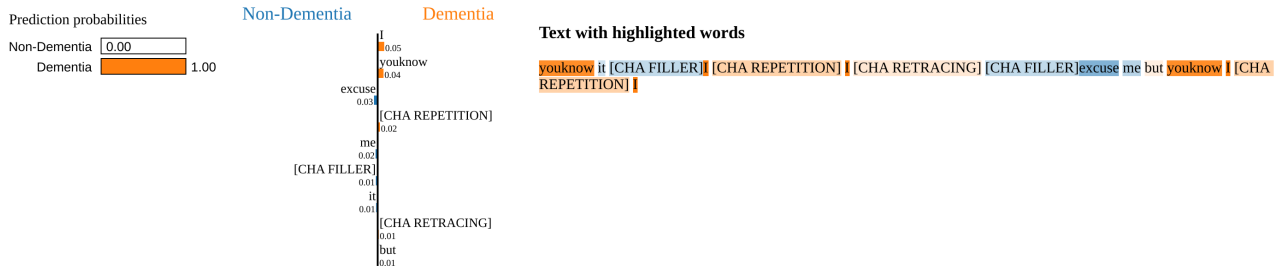
```
Model Predicts Non-Dementia

Examples where anchor applies and model predicts Non-Dementia:
UNK CHA FILLER]standing UNK UNK UNK UNK stool 's UNK to UNK UNK [ UNK FILLER]let UNK go to mother UNK mother UNK dish
[ CHA FILLER]standing on UNK UNK UNK stool 's UNK to fall over [ UNK UNK 's go to UNK the mother UNK UNK
UNK CHA FILLER]standing on stool UNK the stool UNK UNK to fall UNK [ CHA FILLER]let UNK UNK to mother UNK UNK drying
[ CHA FILLER]standing on stool but UNK stool 's UNK UNK fall over [ CHA UNK UNK UNK to mother the mother UNK dishes
[ UNK FILLER]standing on stool UNK the UNK UNK UNK UNK UNK over UNK CHA FILLER]let 's UNK to UNK the mother UNK dishe
[ CHA FILLER]standing on stool UNK UNK stool 's UNK UNK fall over UNK CHA UNK 's UNK to UNK UNK UNK drying UNK
[ CHA FILLER]standing UNK stool UNK the stool 's UNK to UNK over [ CHA UNK 's UNK UNK mother the UNK drying dishes
[ UNK FILLER]standing on UNK but UNK UNK 's about UNK UNK over [ CHA UNK 's go to UNK UNK mother UNK UNK
UNK UNK FILLER]standing UNK stool but the UNK 's about UNK fall UNK [ CHA UNK 's UNK UNK UNK UNK mother UNK UNK
[ CHA FILLER]standing on stool but UNK UNK UNK about UNK fall UNK [ UNK FILLER]let 's UNK UNK UNK the UNK drying dish

Examples where anchor applies and model predicts Dementia:
```

RoBERTa Explanations

Please evaluate the following model explanations as to how informative they are in explaining the model's decision.

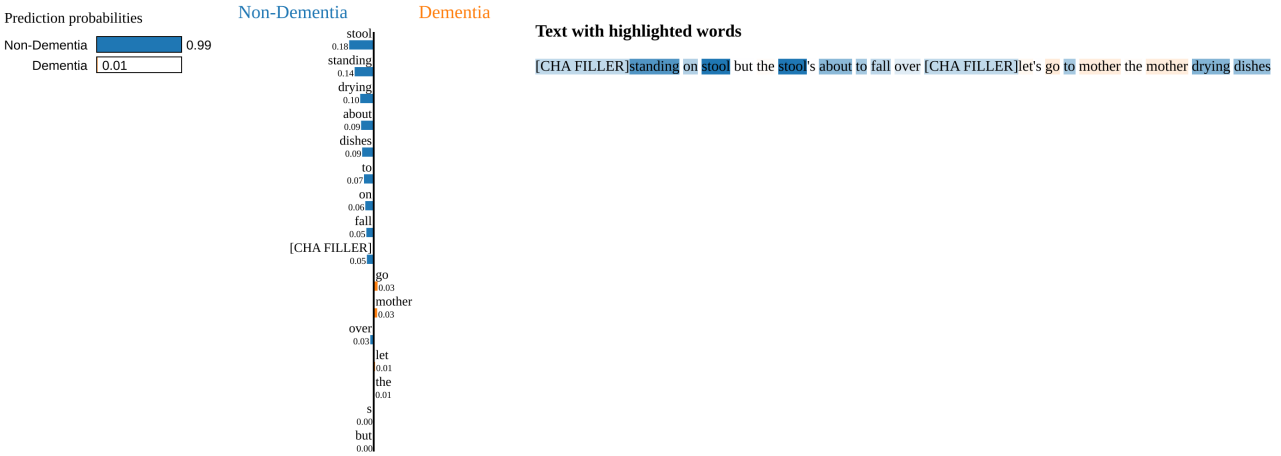
9a. Using LIME and predicting Dementia

Not Informative □—□—□—□—□ Informative



9b. Using LIME and predicting Non Dementia

Not Informative □—□—□—□—□ Informative



10a. Using T-I and predicting Dementia

Not Informative □—□—□—□—□ Informative

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_1 (1.00)	LABEL_1	0.65	#s you know it [CHA FILLER] I [CHA REPETITION] I [CHA RETRACING] [CHA FILLER] exc use me but you know I [CHA REPETITION] I #/s

10b. Using T-I and predicting Non Dementia

Not Informative □—□—□—□—□ Informative

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	LABEL_0 (0.99)	LABEL_0	2.16	#s [CHA FILLER] standing on stool but the stool 's about to fall over [CHA FILLER] let 's go to mother the mother drying dishes #/s

11a. Using ANCHOR and predicting Dementia Not Informative □—□—□—□—□ Informative

```
Model Predicts Dementia

Examples where anchor applies and model predicts Dementia:
youknow UNK UNK CHA FILLER]I [ UNK REPETITION UNK I [ CHA RETRACING ] [ UNK UNK me but youknow UNK UNK UNK REPETITION ] I
youknow UNK [ CHA UNK UNK CHA UNK ] UNK UNK CHA RETRACING ] UNK UNK FILLER]excuse UNK UNK youknow UNK [ UNK UNK ] UNK
youknow UNK UNK CHA FILLER]I [ CHA REPETITION ] UNK [ UNK UNK UNK UNK CHA FILLER]excuse me but youknow UNK UNK UNK REPETITION UNK UNK
youknow UNK [ UNK FILLER]I [ CHA REPETITION ] I UNK UNK UNK ] [ CHA UNK UNK but youknow I UNK CHA REPETITION UNK UNK
UNK UNK [ CHA UNK UNK CHA REPETITION UNK UNK [ CHA UNK ] [ UNK FILLER]excuse UNK UNK youknow I UNK CHA REPETITION UNK I
UNK it UNK UNK FILLER]I UNK CHA UNK ] I [ UNK RETRACING ] UNK UNK FILLER]excuse UNK UNK youknow I UNK CHA UNK ] I
youknow it UNK CHA UNK UNK CHA UNK ] UNK UNK CHA UNK ] [ UNK UNK me UNK youknow UNK [ CHA REPETITION ] UNK
youknow UNK [ UNK FILLER]I UNK CHA UNK UNK UNK UNK UNK UNK UNK [ CHA FILLER]excuse UNK but youknow UNK [ CHA REPETITION ] I
youknow UNK UNK CHA FILLER]I UNK CHA REPETITION ] I UNK CHA UNK UNK [ CHA FILLER]excuse me but youknow I [ CHA REPETITION ] UNK
UNK it [ UNK FILLER]I [ CHA REPETITION ] I [ CHA RETRACING UNK UNK CHA FILLER]excuse UNK UNK youknow UNK [ CHA UNK ] UNK

Examples where anchor applies and model predicts Non-Dementia:
```

11b. Using ANCHOR and predicting Non Dementia Not Informative □—□—□—□—□ Informative

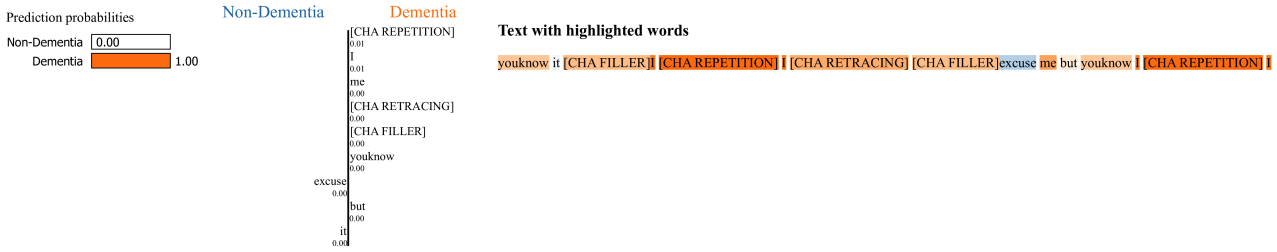
```
Model Predicts Non-Dementia

Examples where anchor applies and model predicts Non-Dementia:
[ UNK UNK on stool but UNK stool UNK UNK to UNK over [ CHA FILLER]let 's go UNK mother the mother drying dishes
[ UNK UNK on UNK UNK UNK stool 's UNK UNK fall UNK [ UNK FILLER]let 's go UNK UNK the mother drying dishes
[ CHA FILLER]standing on stool but UNK stool UNK about to fall over UNK CHA UNK UNK UNK UNK mother UNK mother dry
UNK UNK FILLER]standing on stool but UNK UNK 's about to fall UNK [ UNK UNK UNK go UNK UNK the UNK drying dishes
UNK CHA UNK UNK stool but the UNK 's UNK to fall UNK [ UNK UNK 's UNK to mother the mother drying UNK
UNK UNK UNK on stool but UNK stool 's about to UNK UNK UNK UNK FILLER]let 's go UNK mother UNK UNK drying dishes
[ UNK UNK UNK stool but UNK stool 's about UNK UNK UNK UNK CHA UNK UNK go to UNK the UNK drying UNK
[ UNK UNK UNK stool UNK UNK stool 's UNK UNK fall over [ CHA FILLER]let 's UNK UNK UNK the UNK drying dishes
[ CHA FILLER]standing on stool but the UNK 's about UNK UNK UNK [ UNK UNK 's UNK to mother the mother drying dish
UNK UNK FILLER]standing UNK stool but UNK stool 's about UNK fall over UNK UNK UNK 's go UNK mother UNK UNK dryin

Examples where anchor applies and model predicts Dementia:
```

Distil Explanations

Please evaluate the following model explanations as to how informative they are in explaining the model's decision.

12a. Using LIME and predicting Dementia

Not Informative □—□—□—□—□ Informative



12b. Using LIME and predicting Non Dementia

Not Informative □—□—□—□—□ Informative

13a. Using T-I and predicting Dementia

Not Informative □—□—□—□—□ Informative

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_1 (1.00)	LABEL_1	3.33	[CLS] you ##k ##no ##w it [CHA FILLER] I [CHA REPETITION] I [CHA RETRACING] [CHA FILLER] excuse me but you ##k ##no ##w I [CHA REPETITION] I [SEP]

13b. Using T-I and predicting Non Dementia

Not Informative □—□—□—□—□ Informative

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	LABEL_0 (0.96)	LABEL_0	2.81	[CLS] [CHA FILLER] standing on stool but the stool ' s about to fall over [CHA FILLER] let ' s go to mother the mother drying dishes [SEP]

14a. Using ANCHOR and predicting Dementia Not Informative □—□—□—□—□ Informative

```
Model Predicts Dementia

Examples where anchor applies and model predicts Dementia:
UNK UNK [ CHA UNK [ UNK UNK UNK UNK [ UNK UNK ] UNK UNK UNK UNK but youknow I [ CHA UNK UNK I
UNK it [ CHA FILLER]I UNK UNK REPETITION UNK UNK UNK UNK UNK ] [ CHA UNK me but UNK UNK [ CHA
    UNK UNK UNK
youknow UNK UNK CHA UNK UNK CHA UNK ] UNK [ CHA UNK ] UNK CHA FILLER]excuse UNK UNK UNK I [
    CHA REPETITION UNK UNK
youknow UNK [ UNK UNK UNK UNK REPETITION ] I [ CHA UNK ] UNK UNK UNK UNK UNK youknow I UNK CHA
     REPETITION UNK I
UNK it UNK UNK FILLER]I UNK CHA UNK ] UNK [ UNK RETRACING ] [ CHA UNK me but youknow UNK UNK
    CHA UNK UNK UNK
UNK UNK UNK UNK FILLER]I UNK UNK REPETITION UNK UNK UNK CHA UNK ] [ CHA FILLER]excuse UNK but
    UNK I UNK CHA REPETITION UNK I
youknow UNK UNK CHA FILLER]I UNK CHA UNK UNK UNK UNK UNK RETRACING ] [ UNK UNK UNK but youknow
     UNK UNK CHA REPETITION UNK UNK
youknow it UNK UNK FILLER]I [ CHA REPETITION UNK UNK [ CHA RETRACING ] UNK UNK FILLER]excuse
    UNK but youknow I UNK CHA UNK ] I
youknow UNK UNK CHA UNK UNK CHA UNK UNK I UNK CHA RETRACING UNK UNK CHA UNK me UNK youknow I
    UNK CHA REPETITION ] I
UNK UNK [ UNK UNK [ UNK UNK UNK UNK UNK CHA RETRACING ] UNK CHA FILLER]excuse me UNK youknow I
     UNK CHA REPETITION ] UNK

Examples where anchor applies and model predicts Non-Dementia:
```

14b. Using ANCHOR and predicting Non Dementia Not Informative □—□—□—□—□ Informative

```
Model Predicts Non-Dementia

Examples where anchor applies and model predicts Non-Dementia:
UNK CHA UNK on stool but the UNK UNK UNK UNK UNK UNK [ CHA FILLER]let 's go to UNK UNK mother
    drying UNK
[ CHA UNK on UNK but the stool UNK about UNK fall over [ CHA FILLER]let UNK UNK to mother the
    mother drying UNK
[ CHA UNK UNK stool UNK UNK UNK 's UNK UNK UNK UNK UNK CHA UNK 's go UNK UNK the UNK drying
    UNK
UNK UNK FILLER]standing UNK stool UNK UNK stool UNK about to fall UNK UNK CHA FILLER]let 's
    UNK to UNK UNK mother dryi
UNK UNK FILLER]standing UNK UNK UNK the stool 's UNK to fall UNK [ CHA FILLER]let 's go to UNK
     the UNK drying dishes
UNK UNK UNK on UNK but the UNK 's UNK to UNK over [ UNK UNK 's go UNK mother the UNK drying
    dishes
UNK UNK FILLER]standing on UNK but UNK UNK UNK about to UNK UNK [ CHA UNK UNK go UNK UNK the
    mother drying dishes
[ UNK FILLER]standing on UNK UNK the UNK 's about UNK UNK UNK UNK CHA FILLER]let UNK UNK UNK
    UNK UNK UNK drying dishes
[ CHA UNK on UNK UNK UNK UNK 's UNK to fall UNK UNK UNK FILLER]let UNK go to mother the UNK
    drying dishes
UNK UNK FILLER]standing on stool UNK UNK UNK 's UNK to fall over UNK CHA UNK UNK UNK to UNK
    the UNK drying UNK

Examples where anchor applies and model predicts Dementia:
```

General Questions

Please answer the following questions assuming that a model predicts correctly 95% of the time:

15. Which model do you think performed the best in terms of explainability?

- □ BERT
- □ RoBERTa
- □ DistilBERT

16. Which explainability technique do you think was the most informative?

- □ LIME
- □ Transformer-Interpret
- □ ANCHOR

17. Knowing that LIME takes 5 minutes to generate an explanation, Transformer-Interpret takes 1 minute, and ANCHOR takes 10 second, which technique would you prefer to use in practice?

- □ LIME
- □ Transformer-Interpret
- □ ANCHOR

18. Do you think that the use of AI models in the treatment of Dementia patients is beneficial?
Not at all □—□—□—□—□ Completely

19. Do you think that the use of AI models in the treatment of Dementia patients is ethical?
Not at all □—□—□—□—□ Completely

20. Would you be willing to use AI models in your practice?
Not at all □—□—□—□—□ Completely

21. Would you be willing to use AI models in your practice more if they were explainable?
Not at all □—□—□—□—□ Completely

22. Would an automated system that uses AI models to diagnose Dementia be beneficial?
Not at all □—□—□—□—□ Completely

23. Would an automated system that uses AI models and provides explanations for its decisions in diagnosing Dementia be beneficial?
Not at all □—□—□—□—□ Completely

24. How much would you trust an AI model in diagnosing Dementia?
Not at all □—□—□—□—□ Completely

25. Considering that data for this type of research is in low supply and high demand, would you agree that a global database should be created to facilitate research in this area?
Not at all □—□—□—□—□ Completely

26. Would you be willing to contribute to such a database?
Not at all □—□—□—□—□ Completely

27. Do you think your patients would be willing to contribute to such a database?
Not at all □—□—□—□—□ Completely

28. Do you think that such a database, provided that it is secure and anonymized, would be ethical? Not at all □—□—□—□—□ Completely

Give us your thoughts

29. Please provide any additional comments or thoughts you have about the use of AI models in the treatment of Dementia patients.

Appendix B: DEMET Interface

The DEMET model and interface are a proof of concept designed to be easily intergrated to the COMFORTAGE architecture framework in order to provide accurate predictions and interpretable explanations for the diagnosis and treatment of dementia through spontaneous speech. DEMET is our proposed model and research contribution to the COMFORTAGE project, and is designed to provide clinicians with the necessary insight into the model's decision-making process. DEMET utilises phonological features extracted from spontaneous speech transcripts provided by DementiaBank, which were transformed into a more interpretable and understandable format for clinicians and patients alike, called CHA tokens. DEMET provides a user friendly client-side interface that allows clinicians to interact with the model by providing textual and audio representation of the patient's speech. The output of the interface is a textual color coded explanation of the transcript, along with a prediction of the patient's diagnosis. The interface is designed to be highly intuitive and user-friendly. It can be used through the CLI tool as well as the Web Application interface. We aim to further investigate how CHA tokens can be used for explainability purposes, enhance the qualitative and quantitative metrics of explainability in our explainer, and provide a greater visualisation and user interface for DEMET for a greater user experience overall. We also aim to incorporate camera and computer vision features to DEMET for the analysis of facial expressions and gestures during the interview process. This will allow for the addition of more CHA tokens into the transcripts to be used for the model's predictions and explanations, and provide a more comprehensive and holistic approach to the diagnosis and treatment of dementia. The code implementation for DEMET are provided here. Contributions to the project are welcome and encouraged, and we aim to further develop the model and interface for the COMFORTAGE project in the future.

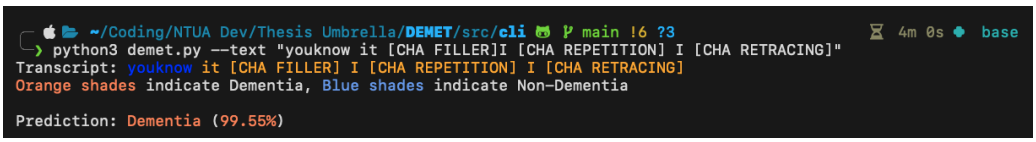


Figure 9.1: DEMET CLI Dementia Prediction

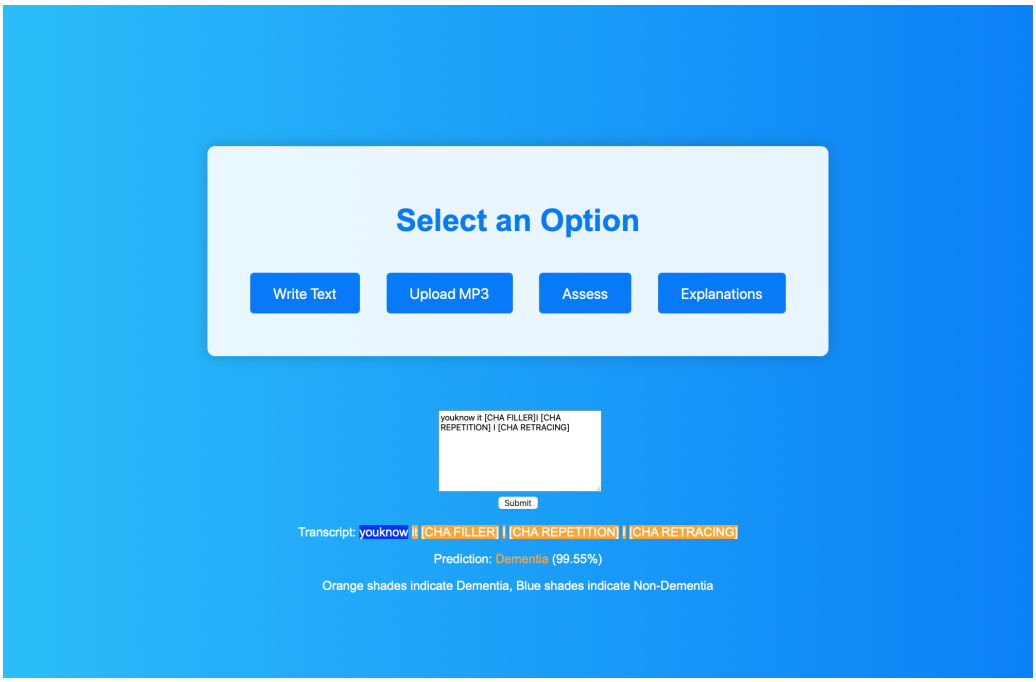


Figure 9.2: DEMET Web

List of Tables

1.1 Μετατροπή CHAT σε CHA tokens 6
1.2 Αποτελέσματα μοντέλων 11
1.3 Αποτελέσματα classifiers με τρεις transformers 11
1.4 Αποτελέσματα classifiers με πέντε transformers 11
1.5 Αποτελέσματα μεθόδων εξήγησης 12

3.1 Model Agnostic XAI Methods 29
3.2 Related Work 39
3.3 Related Work (Continued) 40
3.4 Datasets used by Researchers 41

5.1 Transformation of CHAT symbols into CHA tokens 49

6.1 Evaluation Results for 'short' Input Sequence Size 72
6.2 Evaluation Results for 'medium' Input Sequence Size 73
6.3 Evaluation Results for 'long' Input Sequence Size 73
6.4 Evaluation Results for Transformer Models 73
6.5 Evaluation Results for Ensemble Predictor on 3 Transformer Models 74
6.6 Evaluation Results for Ensemble Predictor on 5 Transformer Models 75
6.7 Evaluation Results for Explainable AI Methods 76

7.1 Privacy Preserving AI Techniques 83
7.2 Bias Reduction Frameworks 87

List of Figures

1.1 Κατανομή των CHA tokens στο σύνολο των δεδομένων 7
1.2 Κατανομή των CHA tokens ως προς τη συχνότητα εμφάνισης ζευγαριών στο ίδιο κείμενο 7
1.3 Συνολική αρχιτεκτονική μοντέλων 8
1.4 BERT LIME 9
1.5 RoBERTa LIME 9
1.6 DistilBERT LIME 9
1.7 Συνολική εξήγηση 9
1.8 DEMET Αρχιτεκτονική 10

5.1 Frequencies of CHA tokens in Dementia and Non-Dementia transcripts 50
5.2 Examples of CHA token Distributions that possibly indicate Dementia 51
5.3 Examples of CHA token Distributions that are similarly distributed in both groups 52
5.4 Co-occurences of CHA tokens in Dementia and Non-Dementia transcripts 53
5.5 Transformer Architecture from "Attention is all you need" 55
5.6 Heatmaps of Short, Medium and Large segment sizes 57
5.7 Single Ensemble Model Architecture 58
5.8 Ensemble Model Architecture 58
5.9 Example Sequence for Testing 60
5.10 BERT LIME 61
5.11 RoBERTa LIME 61
5.12 DistilBERT LIME 61
5.13 BERT Anchor 62
5.14 RoBERTa Anchor 62
5.15 DistilBERT Anchor 62
5.16 BERT Transformers-Interpret 63
5.17 RoBERTa Transformers-Interpret 63
5.18 DistilBERT Transformers-Interpret 64
5.19 Ensemble Explainer Architecture 64
5.20 Ensemble Explanation 65

6.1 COMFORTAGE Framework 69
6.2 DEMET Application Architecture 70
6.3 DEMET Model Architecture 71
6.4 Heatmap for balanced BERT 73
6.5 Heatmap for Gradient Boosting Classifier 74

9.1 DEMET CLI Dementia Prediction 127
9.2 DEMET Web 127