
Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
τομεας σηματων, ελεγχου και ρομποτικης


# Analysis of Audio Signals and Biometric Markers for Supporting Patients with Mental Health Disorders

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της


**Άρτεμις Ανδρώνη**

**Επιβλέπων:** Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.



ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ
Αθήνα, Μάρτιος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων

# Analysis of Audio Signals and Biometric Markers for Supporting Patients with Mental Health Disorders

## ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

**Άρτεμις Ανδρώνη**

**Επιβλέπων:** Πέτρος Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 7η Μαρτίου, 2025.

| | | |
|---|---|---|
| ..... | ..... | ..... |
| Πέτρος Μαραγκός | Αθανάσιος Ροντογιάννης | Ιωάννης Κορδώνης |
| Καθηγητής Ε.Μ.Π. | Αναπληρωτής Καθηγητής Ε.Μ.Π. | Επίκουρος Καθηγητής Ε.Μ.Π. |

Αθήνα, Μάρτιος 2025

.....
**Αρτεμις Ανδρωνη**
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

# Περίληψη

Η πρόβλεψη υποτροπής σε σοβαρές ψυχικές διαταραχές όπως η διπολική διαταραχή και οι διαταραχές του φάσματος της σχιζοφρένειας παραμένει μια σημαντική πρόκληση, συχνά απαιτώντας δαπανηρές νοσηλείες και επιφέροντας δυσκολίες στην καθημερινή ζωή των ασθενών. Οι πρόσφατες εξελίξεις στην ψηφιακή φαινοτυποποίηση (digital phenotyping) παρέχουν τη δυνατότητα συνεχούς παρακολούθησης συμπεριφορικών και φυσιολογικών δεικτών, επιτρέποντας έγκαιρες παρεμβάσεις στην περίπτωση υποτροπής. Η παρούσα έρευνα βασίζεται στο έργο e-Prevention–ένα ολοκληρωμένο σύστημα που υποστηρίζει ασθενείς με ψυχικές διαταραχές μέσω μεθόδων μηχανικής μάθησης για την ανίχνευση υποτροπών–επεκτείνοντας την ηχητική βάση δεδομένων του και αναπτύσσοντας νέα μοντέλα που συνδυάζουν ηχητικά και βιομετρικά σήματα.

Αρχικά, επεκτείναμε την ηχητική βάση δεδομένων του e-Prevention ώστε να συμπεριλάβουμε περισσότερους ασθενείς και περιστατικά υποτροπής. Στη συνέχεια, επαναξιολογήσαμε τα μοντέλα Συνελικτικών Αυτοκωδικοποιητών (Convolutional Autoencoders - CAEs) και Συνελικτικών Μεταβολικών Αυτοκωδικοποιητών (Convolutional Variational Autoencoder - CVAEs) που αναπτύχθηκαν στο πλαίσιο του e-Prevention, χρησιμοποιώντας το επεκταμένο σύνολο δεδομένων. Τα αποτελέσματα επιβεβαίωσαν ότι η αύξηση των δεδομένων βελτιώνει την ικανότητα ανίχνευσης ανωμαλιών στην ομιλία. Στη συνέχεια, αναπτύξαμε autoencoders βασισμένους σε Δίκτυα Μακράς Βραχείας Μνήμης (Long Short-Term Memory Networks - LSTMs), οι οποίοι συλλαμβάνουν τις χρονικές εξαρτήσεις στα φωνητικά σήματα. Διαπιστώσαμε ότι το LSTMAE μοντέλο ενίσχυσε περαιτέρω την απόδοση στην πρόβλεψη υποτροπών, ενώ το CVAE μοντέλο παραμένει η καλύτερη variational προσέγγιση.

Για να διερευνήσουμε τα οφέλη του συνδυασμού πολλαπλών πηγών δεδομένων, αντιστοιχίσαμε τις ηχογραφήσεις συνεντεύξεων με βιομετρικά δεδομένα (σήματα καρδιακού ρυθμού, επιταχυνσιομέτρου, και γυροσκοπίου) που συλλέχθηκαν μέσω ρολογιών (smartwatches). Αναπτύξαμε συνδυαστικές (joint) αρχιτεκτονικές autoencoders, οι οποίες αποτελούνται από ηχητικούς και βιομετρικούς κλάδους, και οι συνδυάζουν τις ηχητικές και βιομετρικές αναπαραστάσεις που προκύπτουν από τους κωδικοποιητές σε έναν κοινό λανθάνοντα χώρο, επιτυγχάνοντας βελτιωμένη ανίχνευση υποτροπών συγκριτικά με τις προσεγγίσεις που χρησιμοποιούν μόνο έναν τύπο δεδομένων. Τα πειραματικά αποτελέσματα υποδεικνύουν ότι τα εξατομικευμένα μοντέλα (personalized) υπερέχουν έναντι των καθολικών (global), υπογραμμίζοντας τη σημασία της προσαρμογής των μοντέλων στις ιδιαιτερότητες του κάθε ασθενούς. Επιπλέον, μέσω πειραμάτων απενεργοποίησης του κάθε κλάδου, επιβεβαιώνουμε την συνεισφορά του κάθε τύπου δεδομένων, αποδεικνύοντας ότι τα συνδυαστικά μοντέλα αξιοποιούν αποτελεσματικά τόσο τα ηχητικά όσο και τα βιομετρικά δεδομένα για βελτιωμένη ανίχνευση υποτροπής.

Συνοψίζοντας, η παρούσα εργασία υποστηρίζει ότι η ενσωμάτωση ηχητικών και βιομετρικών δεδομένων μέσω αρχιτεκτονικών autoencoders μπορεί να βελτιώσει την εγκαιρη ανίχνευση υποτροπών σε ασθενείς με διπολική διαταραχή και σχιζοφρένεια, συμβάλλοντας στις προσπάθειες για έγκαιρες κλινικές παρεμβάσεις και εξατομικευμένη φροντίδα σε ασθενείς με ψυχικές διαταραχές.

**Λέξεις Κλειδιά** — Ψυχικές Διαταραχές, Ψηφιακή Φαινοτυποποίηση, Μηχανική Μάθηση, Ανίχνευση Ανωμαλιών, Αρχιτεκτονικές Αυτοκωδικοποιητών, Αυθόρμητη Ομιλία, Βιομετρικοί Δείκτες, Συνδυασμός Δεδομένων

# Abstract

Relapse prediction in severe mental health conditions such as bipolar disorder and schizophrenia spectrum disorders (SSD) remains a pressing challenge, often necessitating costly hospitalizations and causing significant disruptions in patients' lives. Recent advancements in digital phenotyping offer the potential to monitor behavioral and physiological patterns continuously, thus enabling earlier intervention. This thesis builds upon the e-Prevention project–an integrated system designed to support patients with mental health disorders through machine learning-based relapse detection–by expanding its audio database and developing new models that fuse speech and biometric signals.

First, we expand the original audio database to include additional patients and relapse cases. We then re-evaluate the Convolutional Autoencoder (CAE) and Convolutional Variational Autoencoder (CVAE) models developed during the e-Prevention project on this expanded dataset, confirming that the larger dataset improves anomaly detection in speech. Subsequently, we introduce LSTM-based autoencoders (LSTMAE, LSTMVAE) to capture temporal dependencies in speech signals, finding that the LSTMAE further enhances predictive performance, whereas the CVAE remains the strongest variational approach.

To examine the benefits of multimodal fusion, we align audio recordings from clinical interviews with biometric data (heart rate variability, accelerometer and gyroscope signals) collected from smartwatches. We design joint autoencoder frameworks, that include biometric and audio branches, and combine the learned representations of each modality's encoder into a unified latent space, which yields improved relapse detection compared to unimodal approaches. Experimental results indicate that personalized (patient-specific) models tend to outperform global models, highlighting the importance of tailoring these models to individual patients. Furthermore, ablation experiments through branch disabling validate the contribution of each modality, demonstrating that the joint models effectively leverage both audio and biometric data for improved relapse detection.

In summary, this work demonstrates how integrating audio and biometric data through advanced autoencoder architectures can enhance the early detection of relapse in bipolar disorder and SSD, contributing to the efforts aimed at more timely clinical interventions and personalized care for patients with mental health conditions.

**Keywords** — Mental Health Disorders, Digital Phenotyping, Machine Learning, Anomaly Detection, Autoencoder Architectures, Spontaneous Speech, Biometric Markers, Multimodal Fusion

# Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Πέτρο Μαραγκό για την ευκαιρία που μου έδωσε να εκπονήσω την παρούσα διπλωματική εργασία στο εργαστήριο του. Μέσω αυτής της εμπειρίας, απέκτησα πολύτιμες γνώσεις και ανακάλυψα το ενδιαφέρον μου για τον τομέα της Μηχανικής Μάθησης και της Τεχνητής Νοημοσύνης.

Θα ήθελα επίσης να ευχαριστήσω θερμά τη Νάνσυ Ζλατίντση και τον Χρήστο Γαρούφη, οι οποίοι είναι οι συνεπιβλέποντες της διπλωματικής. Η βοήθεια, η καθοδήγηση και η υποστήριξή τους ήταν πολύ σημαντικές για εμένα και για την ολοκλήρωση της εργασίας μου.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, και όλους τους φίλους μου που ήταν δίπλα μου καθ' όλη τη διάρκεια των σπουδών μου. Ιδιαίτερα, ευχαριστώ τον Γιώργο, την Αφροδίτη και την Ρόζα για όλη την υποστηρίξη και την αγάπη τους.

Άρτεμις Ανδρώνη
Μάρτιος 2025

# Περιεχόμενα

# Λίστα Σχημάτων

# Κατάλογος Πινάκων

# Εκτεταμένη Περίληψη στα Ελληνικά

## Η Σημασία του Προβήματος: Πρόβλεψη Υποτροπής σε Ασθενείς με Ψυχικές Διαταραχές

Οι ψυχικές διαταραχές, όπως η διπολική διαταραχή και οι διαταραχές του φάσματος της σχιζοφρένειας, αποτελούν σοβαρές ψυχικές παθήσεις που επηρεάζουν εκατομμύρια ανθρώπους, με τη σχιζοφρένεια να πλήττει 24 εκατομμύρια ανθρώπους [Wor22] παγκοσμίως και τη διπολική διαταραχή να επηρεάζει το 2,4% του πληθυσμού [Zho+24]. Αμφότερες οι διαταραχές χαρακτηρίζονται από υψηλά ποσοστά υποτροπών, οι οποίες συχνά απαιτούν νοσηλεία και επιδεινώνουν την εξέλιξη της διαταραχής. Έως και το 52% των ασθενών με σχιζοφρένεια οι οποίοι έχουν νοσηλευτεί παρουσιάζουν υποτροπή μέσα σε έναν χρόνο από το εξιτήριο [BMV12], ενώ το 25% των ατόμων με διπολική διαταραχή εμφανίζουν σοβαρές υποτροπές, με το 40% να υποτροπιάζει πολλαπλές φορές σε διάστημα πέντε ετών [Het+23].

Οι διαταραχές αυτές επηρεάζουν σημαντικά την λειτουργικότητα και τη καθημερινή ζωή των ασθενών, με τη διπολική διαταραχή να χαρακτηρίζεται από εναλλαγές μεταξύ καταθλιπτικών και μανιακών επεισοδίων [Nie+23], ενώ η σχιζοφρένεια εκδηλώνεται με ψυχωτικά συμπτώματα και γνωσιακή δυσλειτουργία [VK09]. Οι υποτροπές σχετίζονται με αλλαγές στην ομιλία [LBG20; Fau+16], τη σωματική δραστηριότητα [WM17] και τη λειτουργία του αυτόνομου νευρικού συστήματος (ΑΝΣ), όπως η μεταβλητότητα του καρδιακού ρυθμού [Hen+10]. Οι παραδοσιακές μέθοδοι διάγνωσης βασίζονται σε κλινικές συνεντεύξεις και αναφορές των ίδιων των ασθενών, καθιστώντας δύσκολη την έγκαιρη παρέμβαση. Επομένως, δεδομένων αυτών των προκλήσεων και της περίπλοκης και απρόβλεπτης φύσης των διαταραχών, η συνεχής παρακολούθηση ασθενών με τη χρήση φωνητικών και βιομετρικών δεικτών θα μπορούσε να αποτελέσει κρίσιμο εργαλείο για την έγκαιρη ανίχνευση και πρόληψη των υποτροπών.

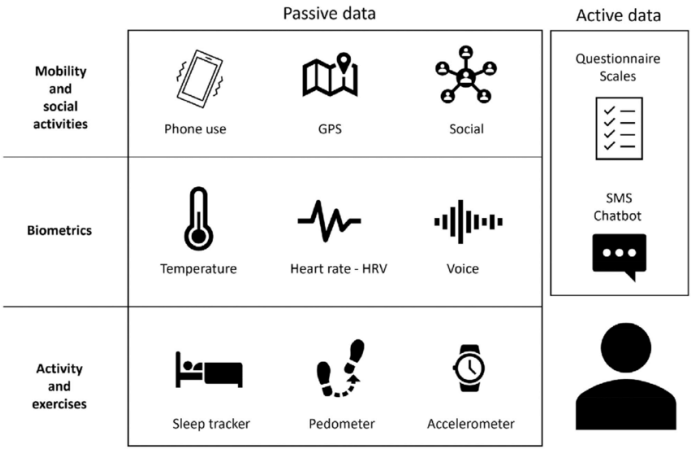



Figure 0.0.1: Παρουσίαση της "ψηφιακής φαινοτυποποίησης" [Mou+21].

Η "ψηφιακή φαινοτυποποίηση" (digital phenotyping), αποτελεί μια καινοτόμο προσέγγιση που αξιοποιεί δεδομένα από ψηφιακές συσκευές, όπως έξυπνα ρολόγια (smartwatches) και κινητά τηλέφωνα, για την ανάλυση συμπεριφορικών και φυσιολογικών μοτίβων [OR16; MZS17], με εφαρμογές σε διάφορους τομείς της υγείας (Σχήμα 0.0.1). Στον τομέα της ψυχικής υγείας, η συνεχής και μη παρεμβατική παρακολούθηση που προσφέρει

επιτρέπει την έγκαιρη ανίχνευση λεπτών μεταβολών που προηγούνται μιας υποτροπής–μεταβολών που συχνά δεν μπορούν να ανιχνευτούν μέσω των παραδοσιακών κλινικών μεθόδων αξιολόγησης–βελτιώνοντας έτσι την αποτελεσματικότητα της πρόληψης και της έγκαιρης παρέμβασης [Pan+18; Bar+18].

### Σκοπός της Έρευνας

Η παρούσα διπλωματική εργασία βασίζεται στο έργο e-Prevention [Zla+22], το οποίο ανέπτυξε ένα ολοκληρωμένο σύστημα για τη παρακολούθηση ασθενών με ψυχικές διαταραχές, ενσωματώνοντας τεχνικές μηχανικής μάθησης για την ανίχνευση υποτροπών. Το έργο επικεντρώθηκε κυρίως στην ξεχωριστή ανάλυση βιομετρικών και ακουστικών δεδομένων, χωρίς να εξετάζει πλήρως τη συνδυαστική τους χρήση για την ενίσχυση της πρόβλεψης της υποτροπής. Συνεπώς, η έρευνα αυτή στοχεύει στην επέκταση και τη βελτίωση των μοντέλων που αναπτύχθηκαν κατά την έρευνα του e-Prevention, εστιάζοντας στη συνδυαστική ανάλυση ακουστικών και βιομετρικών δεδομένων για τη βελτίωση της ανίχνευσης υποτροπών σε ασθενείς με ψυχικές διαταραχές.

Συγκεκριμένα, η έρευνα συνεισφέρει στον τομέα της ψηφιακής φαινοτυποποίησης στην ψυχική υγεία μέσω των εξής:

- Επέκταση της ηχητικής βάσης δεδομένων e-Prevention, δημιουργώντας ένα μεγαλύτερο σύνολο δεδομένων για βελτιωμένη αξιολόγηση και γενίκευση των μοντέλων.
- Εκ νέου αξιολόγηση των συνελικτικών μοντέλων αυτοκωδικοποιητών (autoencoders) που αναπτύχθηκαν κατά την έρευνα του e-Prevention, ώστε να εξεταστεί η απόδοσή τους στην ανίχνευση υποτροπών σε περισσότερους ασθενείς.
- Ανάπτυξη και αξιολόγηση autoencoder αρχιτεκτονικών βασισμένες σε Δίκτυα Μακράς Βραχείας Μνήμης (Long Short-Term Memory - LSTM) για την ανίχνευση υποτροπών από δεδομένα ομιλίας, συγκρίνοντας την αποτελεσματικότητά τους με τα συνελικτικά μοντέλα.
- Συνδυαστική ανάλυση των ακουστικών και βιομετρικών δεδομένων μέσω της ανάπτυξης συνεδυαστικών (joint) autoencoder μοντέλων για την ενίσχυση της ακρίβειας πρόβλεψης υποτροπών και την ανάδειξη των πλεονεκτημάτων των συνδυαστικών προσεγγίσεων.

## Θεωρητικό Υπόβαθρο

### Αναπαραστάσεις Ηχητικών Σημάτων

Τα ηχητικά σήματα αποτελούν σημαντική πηγή πληροφοριών για την ανάλυση προτύπων ομιλίας και την ανίχνευση μεταβολών σε αυτή. Για την επεξεργασία τους μέσω αλγορίθμων μηχανικής μάθησης, πρέπει να μετατραπούν σε κατάλληλες αναπαραστάσεις έτσι ώστε να εξαχθούν τα χρήσιμα χαρακτηριστικά τους.

Στο πεδίο του χρόνου τα ηχητικά σήματα απεικονίζονται ως κυματομορφές, όπου το πλάτος μεταβάλλεται με τον χρόνο. Ωστόσο, αυτή η αναπαράσταση δεν παρέχει άμεσα πληροφορίες για το φασματικό περιεχόμενο του σήματος. Επομένως, για την εξαγωγή του φασματικού περιεχομένου, χρησιμοποιούνται ο μετασχηματισμός Fourier (FT), για τα σήματα συνεχούς χρόνου, και ο Διακριτός Μετασχηματισμός Fourier (DFT), για τα σήματα διακριτού χρόνου, αντίστοιχα.

Για την ανάλυση μη στατικών σημάτων, χρησιμοποιείται ο Μετασχηματισμός Fourier Βραχέος Χρόνου (STFT), ο οποίος αναλύσει το σήμα εφαρμόζοντας ένα ολισθαίνον παράθυρο σε τμήματά του, με σκοπό να αποτυπώσει μεταβολές του φάσματος στον χρόνο. Μία από τις σημαντικότερες αναπαραστάσεις των ηχητικών σημάτων είναι το φασματογράφημα (spectrogram), το οποίο απεικονίζει την μεταβολή της ενέργειας του σήματος σε συνάρτηση με το χρόνο και τη συχνότητα. Υπολογίζεται ως εξής:

$$\text{spectrogram}(m, \omega) = |STFT\{s[n]\}(m, \omega)|^2 \tag{0.0.1}$$

Όπου:

- $|STFT\{s[n]\}(m, \omega)|^2$ είναι η πυκνότητα φάσματος ισχύος (PSD) του σήματος.

Ωστόσο, το φασματογράφημα απεικονίζει τις συχνότητες σε γραμμική κλίμακα, αντίθετα με την ανρθώπινη ακοή, η οποία αντιλαμβάνεται την τονικότητα με λογαριθμική κλίμακα. Επομένως, για την προσαρμογή στην ανθρώπινη

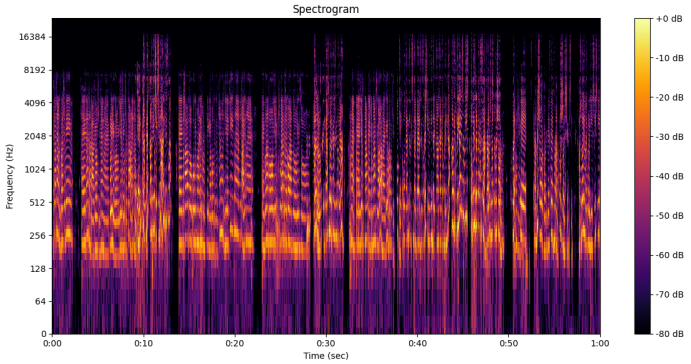



Figure 0.0.2: Απεικόνιση Mel-spectrogram ηχητικού σήματος.

ακουστική αντίληψη, χρησιμοποιείται το Mel-φασματογράφημα (Mel-spectrogram), το οποίο χρησιμοποιεί την κλίμακα Mel, που αντιστοιχεί γραμμικές συχνότητες $f$ σε Mel-συχνότητες $m(f)$.

## Αναπαραστάσεις Βιομετρικών Σημάτων

Τα βιομετρικά σήματα, που προέρχονται από αισθητήρες, μπορούν να προσφέρουν πολύτιμες πληροφορίες για τη φυσική κατάσταση και τις μεταβολές της συμπεριφοράς ενός ατόμου. Από τα σήματα αυτά συνήθως εξάγονται χρονικά, φασματικά ή μη-γραμμικά χαρακτηριστικά για την ανάλυση τους με αλγορίθμους μηχανικής μάθησης.

**Χρονικά Χαρακτηριστικά**: Χαρακτηριστικά όπως η ενέργεια βραχέος χρόνου (Short-Time Energy - STE), η μέση τιμή και η μεταβλητότητα, είναι χρήσιμα για την περιγραφή των διακυμάνσεων της έντασης και των στατιστικών ιδιοτήτων του σήματος.

**Φασματικά Χαρακτηριστικά**: Τα φασματικά χαρακτηριστικά αναλύουν την κατανομή της ισχύος του σήματος σε διαφορετικές συχνοτικές ζώνες, με μεθόδους όπως το περιοδογράφημα Lomb-Scargle [Sca82], που χρησιμοποιείται ευρέως σε βιομτερικά σήματα όπως η μεταβλητότητα καρδιακού ρυθμού (HRV). Η HRV μελετάται περαιτέρω μέσω συχνοτικών ζωνών (frequency bands) για την αξιολόγηση της δραστηριότητας του αυτόνομου νευρικού συστήματος [SG17a].

**Μη-Γραμμικά Χαρακτηριστικά**: Η ανάλυση του γραφήματος Poincaré αποτελεί μια δημοφιλή μέθοδο για την αναπαράσταση της μεταβλητότητας των βιομετρικών σημάτων, όπως της HRV, χρησιμοποιώντας τους περιγραφητές SD1 και SD2, οι οποίοι αντιστοιχούν στα μήκη των αξόνων της έλλειψης που περικλείει σημεία του γραφήματος, τα οποία ανιπροσωπεύουν ζεύγη διαδοχικών NN διαστημάτων του καρδιακού ρυθμού [BPK01; PG07].

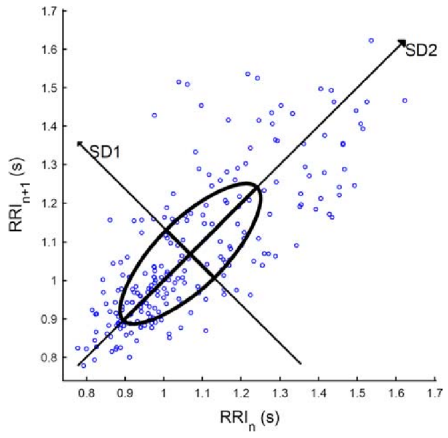



Figure 0.0.3: Παράδειγμα γραφήματος Poincaré ενός σήματος HRV [Cho+09].

Επιπλέον μη-γραμμικά χαρακτηριστικά, αποτελούν η εντροπία του σήματος, η οποία παρέχει πληροφορίες για το πόσο "ακανόνιστο" είναι το σήμα, και οι διαστάσεις fractal (Fractal Dimension - FD) [Man82], όπως η διάσταση Higuchi (HFD) [Hig88] και η πολυκλιμακωτή διάσταση fractal (Multiscale Fractal Dimension - MFD) [Mar94], που παρέχουν πληροφορίες για την πολυπλοκότητα των βιομετρικών σημάτων.

## Αρχιτεκτονικές Νευρωνικών Δικτύων

Τα νευρωνικά δίκτυα είναι υπολογιστικά μοντέλα εμπνευσμένα από τη δομή των βιολογικών νευρώνων. Η βασική δομή τους περιλαμβάνει συνδεδεμένα επίπεδα νευρώνων, τα οποία επεξεργάζονται δεδομένα εισόδου μέσω βαρών και συναρτήσεων ενεργοποίησης με σκοπό την αντιστοίχησή τους σε μία επιθυμητή έξοδο.

Συνήθεις συναρτήσεις ενεργοποίησης αποτελούν οι σιγμοειδής (Sigmoid), η υπερβολική εφαπτομένη (Tanh) και η ReLU (Rectified Linear Unit), οι οποίες εισάγουν μη-γραμμικότητα και επιτρέπουν στο δίκτυο να μάθει πολύπλοκες σχέσεις μεταξύ των δεδομένων.

Η εκπαίδευση του δικτύου γίνεται μέσω τεχνικών βελτιστοποίησης (backpropagation, gradient descent) και τεχνικών κανονικοποίησης (regularization) όπως η L1/L2 και η κανονικοποίηση παρτίδων (batch normalization), που βελτιώνουν τη γενίκευση και αποτρέπουν την υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης.

### Συνελικτικά Νευρωνικά Δίκτυα (CNNs)

Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks - CNNs), αποτελούν μία από τις πιο σημαντικές προσεγγίσεις στον τομέα της επεξεργασίας εικόνας και ήχου λόγω της ικανότητάς τους να εξάγουν ιεραρχικά χαρακτηριστικά από τα δεδομένα με τη χρήση συνελικτικών επιπέδων και την εφαρμογή φίλτρων [MIB20; HGD17]. Τα φίλτρα αυτά επικεντρώνονται σε μικρές, τοπικές περιοχές των δεδομένων εισόδου, γεγονός που τα καθιστά κατάλληλα για την ανάλυση δεδομένων πολλών διαστάσεων.

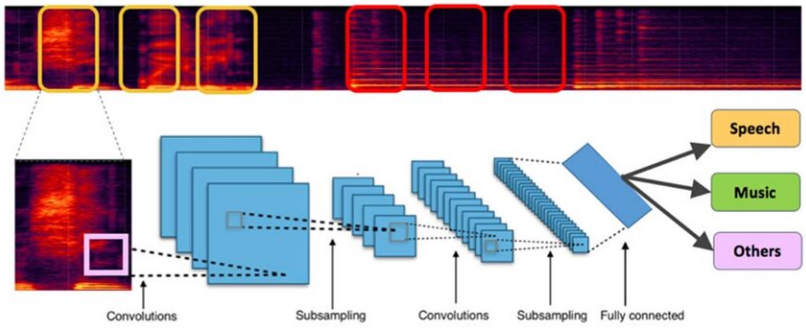



Figure 0.0.4: Παράδειγμα χρήσης αρχιτεκτονικής CNN σε ηχητικά δεδομένα [Cha+17].

Βασικές παράμετροι των συνελικτικών επιπέδων είναι το μέγεθος του φίλτρου ή πυρήνα (kernel), το βήμα της συνέλιξης (stride) και η μέθοδος συμπλήρωσης (padding), τα οποία ελέγχουν τις διαστάσεις των χαρακτηριστικών που εξάγονται από το επίπεδο. Μικρότερα φίλτρα (π.χ. 3×3) ανιχνεύουν πιο λεπτομερή χαρακτηριστικά, ενώ μεγαλύτερα φίλτρα (π.χ. 7×7) αποτυπώνουν ευρύτερα μοτίβα. Επιπλέον, μικρότερα βήματα παρέχουν υψηλότερη ανάλυση, ενώ μεγαλύτερα βήματα μειώνουν το μέγεθος των δεδομένων. Το padding επιτρέπει την πλήρη κάλυψη του σήματος από τα φίλτρα, προσθέτοντας μηδενικά στοιχεία στις άκρες της εισόδου.

Επίπεδα συγκέντρωσης (pooling layers) μειώνουν τις διαστάσεις των χαρακτηριστικών, διατηρώντας παράλληλα τη σημαντική πληροφορία. Τέλος, πλήρως συνδεδεμένα επίπεδα (fully connected layers) ενοποιούν τα εξαγόμενα χαρακτηριστικά για την τελική πρόβλεψη. Τεχνικές όπως υπερδειγματοληψία (upsampling), υποδειγματοληψία (downsampling) και συναρτήσεις ενεργοποίησης βελτιώνουν την απόδοση των CNNs σε διάφορες εφαρμογές.

### Δίκτυα Μακράς Βραχείας Μνήμης (Long Short-Term Memory Networks - LSTMs)

Τα Δίκτυα Μακράς Βραχείας Μνήμης (Long Short-Term Memory - LSTM) [HS97] αποτελούν μια εξελιγμένη μορφή των Αναδρομικών Νευρωνικών Δικτύων (Recurrent Neural Networks - RNNs), σχεδιασμένα να ξεπερ-

νούν το πρόβλημα vanishing gradient μέσω ενός μηχανισμού μνήμης που επιτρέπει τη διατήρηση πληροφοριών για μεγάλα χρονικά διαστήματα. Λόγω της ιδιότητάς τους αυτής, τα LSTM είναι κατάλληλα για την ανάλυση δεδομένων με χρονική εξάρτηση και βρίσκουν εφαρμογή σε πολλούς τομείς, όπως η αναγνώριση ομιλίας, η ανάλυση και η πρόβλεψη χρονοσειρών [Li+23; Lin+21].

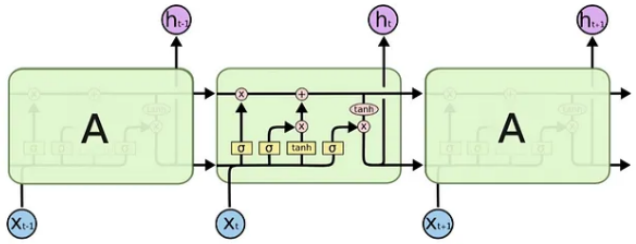



Figure 0.0.5: Αρχιτεκτονική ενός LSTM κελιού [Ola15].

Η αρχιτεκτονική τους βασίζεται σε "κελιά" μνήμης (memory cells) και πύλες ελέγχου (input, forget, output gates), οι οποίες διαχειρίζονται τη ροή πληροφοριών, επιλέγοντας ποιες τιμές να αποθηκεύσουν, να διατηρήσουν ή να διαγράψουν.

## Αυτοκωδικοποιητές (Autoencoders)

Οι αυτοκωδικοποιητές (autoencoders) είναι νευρωνικά δίκτυα σχεδιασμένα για εφαρμογές μη-επιβλεπόμενης μηχανικής μάθησης, όπου τα μοντέλα ανακαλύπτουν πρότυπα και δομές στα δεδομένα χωρίς τη χρήση επισημειώσεων. Αποτελούνται από δύο κύρια τμήματα, τον κωδικοποιητή (encoder) ο οποίος συμπιέζει τα δεδομένα εισόδου σε μία αναπαράσταση χαμηλότερης διάστασης ή λανθάνοντα χώρο (latent space), και τον αποκωδικοποιητή (decoder) που ανακατασκευάζει την αρχική είσοδο. Λόγω της ικανότητάς τους να ανακαλύπτουν σημαντικά χαρακτηριστικά στα δεδομένα, χρησιμοποιούνται ευρέως για εξαγωγή χαρακτηριστικών, ανίχνευση ανωμαλιών και ανακατασκευή δεδομένων χαμηλής ποιότητας [Ber+24].

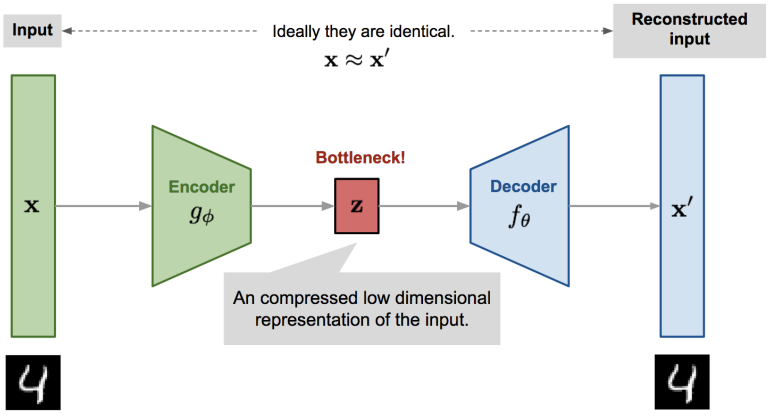



Figure 0.0.6: Παράδειγμα αρχιτεκτονικής autoencoder [Wen18].

Ο στόχος της εκπαίδευσης ενός autoencoder είναι η ελαχιστοποίηση της απώλειας ανακατασκευής (reconstruction loss), η οποία μετρά τη διαφορά μεταξύ της αρχικής εισόδου και της ανακατασκευής της από τον αποκωδικοποιητή. Μία σύνηθης μετρική απώλειας ανακατασκευής είναι το μέσο τετραγωνικό σφάλμα (Mean Squared Error - MSE) μεταξύ της εισόδου και της ανακατευασμένης εξόδου και ορίζεται ως εξής:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x - \hat{x})^2 \tag{0.0.2}$$

Όπου:

- $x$ είναι το δείγμα εισόδου.
- $\hat{x}$ είναι η ανακατασκευή της εισόδου από τον αποκωδικοποιητή.
- $N$ είναι ο αριθμός των δειγμάτων.

Διαφορετικές παραλλαγές των autoencoders έχουν αναπτυχθεί για την αντιμετώπιση διαφορετικών εφαρμογών, όπως οι Συνελικτικοί Autoencoders (CAEs) για εικόνες και ηχητικά δεδομένα και οι LSTM Autoencoders (LSTMAEs) για χρονοσειρές και ακολουθιακά δεδομένα.

#### Μεταβολικοί Αυτοκωδικοποιητές (Variational Autoencoders - VAEs)

Οι Μεταβολικοί Αυτοκωδικοποιητές (Variational Autoencoders - VAEs) διαφέρουν από τους κλασικούς autoencoders, καθώς αντί να αντιστοιχούν τα δεδομένα σε έναν σημείο στον λανθάνοντα χώρο, μαθαίνουν μια πιθανότητα κατανομής (Gaussian Distribution) για κάθε δείγμα. Ο κωδικοποιητής μαθαίνει τις παραμέτρους της κατανομής (μέση τιμή $\mu$ και διακύμανση $\sigma^2$), ενώ ο αποκωδικοποιητής ανακατασκευάζει δεδομένα από δειγματοληψία στον λανθάνοντα χώρο. Η απώλεια ενός VAE περιλαμβάνει δύο όρους:

- Απώλεια ανακατασκευής (Reconstruction Loss): Μετρά τη διαφορά μεταξύ εισόδου και εξόδου όπως στους κλασικούς autoencoders.
- Απόκλιση Kullback-Leibler (Kullback-Leibler - KL Divergence): Υπολογίζει την απόκλιση μεταξύ της λανθάνουσας κατανομής και μιας κανονικής κατανομής, βελτιώνοντας τη γενίκευση του μοντέλου.

Οι VAEs χρησιμοποιούνται ευρέως για την ανάλυση και την ανακατασκευή δεδομένων, και μπορούν να λειτουργήσουν ως γενετικά μοντέλα (generative models), δημιουργώντας νέα δείγματα παρόμοια με τα δεδομένα εκπαίδευσης [CCM24].

### Ανίχνευση Ανωμαλιών

Η ανίχνευση ανωμαλιών είναι μια σημαντική εφαρμογή της μηχανικής μάθησης που εντοπίζει αποκλίσεις από φυσιολογικά πρότυπα, συχνά σχετιζόμενες με βλάβες συστημάτων, απάτες ή προβλήματα υγείας. Οι autoencoders, ως μη-επιβλεπόμενη προσέγγιση, μαθαίνουν τη δομή των φυσιολογικών δεδομένων και εντοπίζουν ανωμαλίες μέσω απωλειών ανακατασκευής (π.χ. MSE, απόσταση Mahalanobis, απόκλιση KL).

Για την αξιολόγηση χρησιμοποιείται συνήθως η ROC-AUC μετρική, η οποία μετρά την ικανότητα διάκρισης μεταξύ φυσιολογικών και ανώμαλων δειγμάτων, αναλύοντας τη σχέση των πραγματικών ανωμαλιών που εντοπίζονται σωστά από το μοντέλο (True Positive Rate - TPR) και των φυσιολογικών δεδομένων που λανθασμένα ταξινομήθηκαν ως ανωμαλίες (False Positive Rate - FPR) σε διαφορετικά κατώφλια. Υψηλότερη τιμή ROC-AUC υποδηλώνει καλύτερη απόδοση του μοντέλου στην ανίχνευση ανωμαλιών.

## Σχετική Βιβλιογραφία

### Ψηφιακή Φαινοτυποποίηση

Η ψηφιακή φαινοτυποποίηση, η οποία αξιοποιεί την παθητική συλλογή δεδομένων από φορητές συσκευές, έχει αναδειχθεί ως ένα ισχυρό εργαλείο για την παρακολούθηση συμπεριφορικών και φυσιολογικών μοτίβων που σχετίζονται με την ψυχική υγεία. Διάφορες μελέτες έχουν αναδείξει τη συμβολή της στην πρόβλεψη υποτροπών μέσω επιβλεπόμενων και μη-επιβλεπόμενων μεθόδων μηχανικής μάθησης. Οι επιβλεπόμενες προσεγγίσεις, χρησιμοποίησαν συμπεριφορικά δεδομένα από smartphones, όπως κινητικότητα, μοτίβα επικοινωνίας και χρήση εφαρμογών, για να προβλέψουν υποτροπές στη σχιζοφρένεια, τη διπολική διαταραχή και την κατάθλιψη με ποσοστά ακρίβειας που φτάνουν το 80% [Bar+18; Osm+15; Ikä+24]. Παράλληλα, μη-επιβλεπόμενες μέθοδοι, όπως το σύστημα CrossCheck [Adl+20], αξιοποίησαν autoencoders (Σχήμα 0.0.9) για την ανίχνευση αποκλίσεων από τη φυσιολογική συμπεριφορά, εντοπίζοντας έγκαιρα σημάδια υποτροπής μέσω ανωμαλιών στην κινητικότητα, τον ύπνο και την κοινωνική αλληλεπίδραση. Επιπλέον, η ανάλυση χαρακτηριστικών της ομιλίας αποτελεί μια σημαντική προσέγγιση για την πρόβλεψη υποτροπών, καθώς οι διακυμάνσεις στην ομιλία μπορούν να λειτουργήσουν ως αντικειμενικοί δείκτες της ψυχικής κατάστασης. Σχετικές μελέτες αξιοποίησαν δεδομένα αυθόρμητης και καθοδηγούμενης ομιλίας από ηχογραφήσεις μέσω smartphones, εφαρμόζοντας τόσο παραδοσιακά μοντέλα μηχανικής μάθησης (όπως Support Vector Machines (SVMs) και Gaussian Mixture Models (GMMs)) όσο

και βαθιά νευρωνικά δίκτυα (όπως CNN-LSTM και Dense Neural Networks (DNNs)). Οι προσεγγίσεις αυτές είχαν ως αποτέλεσμα την εξαγωγή σημαντικών χαρακτηριστικών της ομιλίας και την ανίχνευση υποτροπών σε ασθενείς με διπολική διαταραχή και σχιζοφρένεια, επιτυγχάνοντας υψηλή ακρίβεια [Pan+18; HWS18; Gid+19].

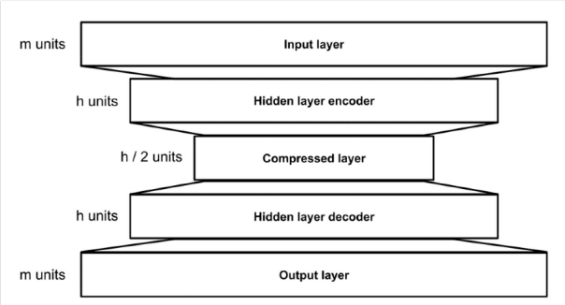



Figure 0.0.7: FNN-AE

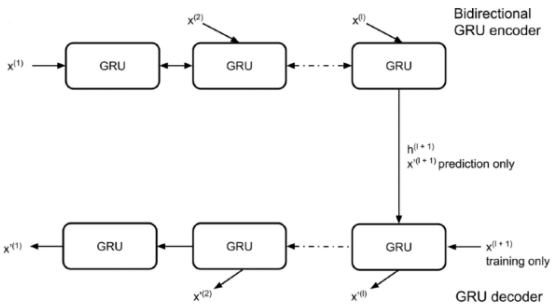



Figure 0.0.8: GRU-Seq2Seq-AE

Figure 0.0.9: Αρχιτεκτονικές autoencoders που χρησιμοποιήθηκαν στο σύστημα CrossCheck για την ανίχνευση ανωμαλιών [Adl+20].

Τα αποτελέσματα των ερευνών ήταν ενθαρρυντικά, αναδεικνύοντας την ικανότητα των μεθόδων ψηφιακής φαινοτυποποίησης στην πρόβλεψη υποτροπών και την παρακολούθηση της ψυχικής υγείας. Ωστόσο, παραμένουν προκλήσεις όπως τα μικρά σύνολα δεδομένων, η μεταβλητότητα μεταξύ ασθενών και η ανάγκη για την ανάπτυξη εξατομικευμένων μοντέλων πρόβλεψης, τα οποία απαιτούν περαιτέρω διερεύνση.

## Συνδιασμός Πολλαπλών Πηγών Δεδομένων (Multimodal Fusion)

Ο συνδυασμός πολλαπλών πηγών δεδομένων (multimodal fusion) έχει αποδειχθεί αποτελεσματική προσέγγιση στην ανίχνευση υποτροπών και παρακολούθηση ψυχικών διαταραχών, συνδυάζοντας ακουστικά, γραπτά, οπτικά, και συμπεριφορικά χαρακτηριστικά. Μελέτες που χρησιμοποιούν το σύνολο δεδομένων DAIC-WOZ [Rin+17] έχουν δείξει ότι ο συνδυασμός φωνητικών, οπτικών και γραπτών δεδομένων βελτιώνει την ανίχνευση της κατάθλιψης μέσω μοντέλω νευρωνικών δικτύων (VGGish [SZ14], BERT [Dev+19], CNNs, LSTMs, DNNs) [**Flores2022**; OZ22; LHL19; Alm+24].

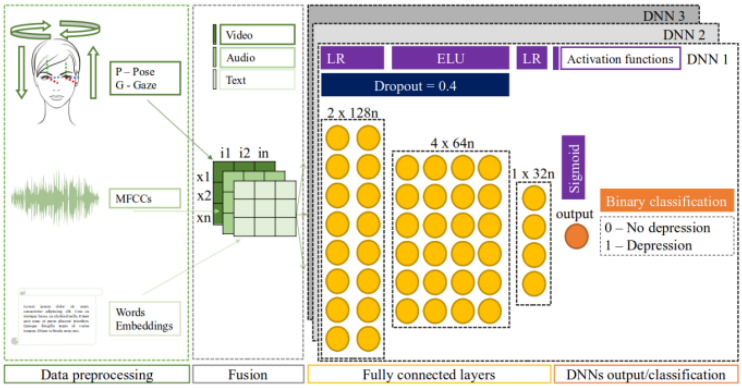



Figure 0.0.10: Προτεινόμενη αρχιτεκτονική για multimodal ανίχνευση ανωμαλιών σε ασθενείς με κατάθλιψη [Alm+24].

Παράλληλα, η ενσωμάτωση φυσιολογικών και συμπεριφορικών δεδομένων από smartphones έχει επιτρέψει την πρόβλεψη βαθμολογιών HDRS [Ham76] και YMRS [You+78] για διπολική διαταραχή [Su+21]. Αντίστοιχα, η χρήση του μοντέλου FusionNet απέδειξε ότι ο συνδυασμός δεδομένων κειμένου και οπτικών δεδομένων από μέσα κοινωνικής δικτύωσης βελτιώνει την ανίχνευση κατάθλιψης [Wan+22]. Ωστόσο, σημαντικές προκλήσεις παραμένουν, όπως η έλλειψη δεδομένων, η ανισορροπία μεταξύ κλάσεων, η ασυνέπεια στη χρονική αντιστοίχιση των χαρακτηριστικών και η υψηλή υπολογιστική πολυπλοκότητα των fusion τεχνικών. Επιπλέον, παρά την

πρόοδο του συνδυασμού ακουστικών, γραπτών και οπτικών δεδομένων, η συνδυασμένη ανάλυση φωνητικών και βιομετρικών δεδομένων παραμένει ανεξερεύνητη, παρουσιάζοντας σημαντικές ευκαιρίες για μελλοντική έρευνα.

## Ανίχνευση Ανωμαλιών σε Ηχητικά Δεδομένα

Η ανίχνευση ανωμαλιών σε ηχητικά δεδομένα αποτελεί πρόκληση λόγω της πολυπλοκότητας και της μεταβλητότητας των ηχητικών προτύπων, καθώς και του θορύβου που συχνά συνοδεύει τα ηχητικά σήματα. Η χρήση βαθιών autoencoder αρχιτεκτονικών, και ιδιαίτερα LSTM autoencoders, έχει αποδειχθεί αποτελεσματική στον εντοπισμό ανωμαλιών σε χρονικά εξαρτώμενα δεδομένα. Μελέτες έχουν αξιοποιήσει τέτοια μοντέλα για την ανίχνευση ανωμαλιών σε βιομηχανικά ηχητικά σήματα, επιτυγχάνοντας πολύ υψηλή ακρίβεια [Coe+22; BDI20; Mob+22]. Συνεπώς, δεδομένης της ικανότητάς τους να ανιχνεύουν αποκλίσεις σε πολύπλοκα ακουστικά μοτίβα, η εφαρμογή τους θα μπορούσε να επεκταθεί και στην ανίχνευση ανωμαλιών στην ομιλία, συμβάλλοντας στην παρακολούθηση της ψυχικής υγείας μέσω έγκαιρης ανίχνευσης αλλαγών στη φωνή και στη δομή της ομιλίας.

## Το Έργο e-Prevention

Το e-Prevention [Zla+22] αποτελεί τριετές έργο που αποσκοπεί στην ανάπτυξη προηγμένων και καινοτόμων ηλεκτρονικών υπηρεσιών για την ιατρική παρακολούθηση ψυχικών διαταραχών, συγκεκριμένα της διπολικής διαταραχής και της σχιζοφρένειας. Η προσπάθεια αυτή αποσκοπεί στον εντοπισμό δεικτών και χαρακτηριστικών που μπορούν να προβλέψουν μεταβολές της διάθεσης και των ψυχοπαθολογικών συμπτωμάτων των ασθενών, με στόχο την πρόληψη των υποτροπών και τη βελτίωση της ποιότητας ζωής τους.

Το ολοκληρωμένο σύστημα που υλοποιήθηκε κατά τη διάρκεια του έργου (Σχήμα 0.0.11), περιλαμβάνει έναν φορητό αισθητήρα smartwatch που παρακολουθεί και καταγράφει συνεχώς μια σειρά βιομετρικών (π.χ. μεταβλητότητα καρδιακού ρυθμού) και συμπεριφορικών δεικτών (π.χ. δεδομένα επιταχυνσιόμετρου και γυροσκοπίου). Επιπλέον, το σύστημα περιλαμβάνει μια φορητή συσκευή (tablet) η οποία έχει εγκατασταθεί στο σπίτι του ασθενούς και καταγράφει σύντομα οπτικοακουστικά αποσπάσματα των ασθενών καθώς συμμετέχουν σε συνεντεύξεις με κλινικούς ιατρούς. Αυτή η λειτουργία επιτρέπει τη συλλογή πολύτιμων δεδομένων, συμπεριλαμβανομένης της ομιλίας και των εκφράσεων του προσώπου.

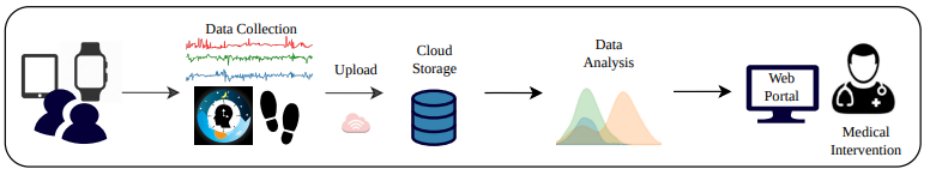



Figure 0.0.11: Το ολοκληρωμένο σύστημα του e-Prevention [Zla+22].

### Στατιστική Ανάλυση Δεδομένων Βιομετρικών Δεικτών

Βιομετρικά δεδομένα γυροσκοπίου, επιταχυνσιόμετρου και καρδιακού ρυθμού συλλέχθηκαν από το smartwatch, για 23 υγιείς εθελοντές και 24 ασθενείς με διπολική διαταραχή και σχιζοφρένεια. Για την στατιστική ανάλυση των δεδομένων χρησιμοποιήθηκαν τα χαρακτηριστικά που αναφέρονται στο Θεωρητικό Υπόβαθρο, με σκοπό να αναδειχθούν οι διαφορές μεταξύ των δύο ομάδων. Παρατηρήθηκαν σημαντικές διαφορές στα πρότυπα κίνησης κατά τη διάρκεια της ημέρας, καθώς και κατά τη διάρκεια του ύπνου, γεγονός που καθιστά τους βιομετρικούς δείκτες χρήσιμους για τη διάκριση μεταξύ των δύο ομάδων και την ανίχνευση ανωμαλιών.

### Ανίχνευση Υποτροπών με Χρήση Autoencoders

Για την ανίχνευση υποτροπών στους ασθενείς το έργο διερεύνησε διάφορες αρχιτεκτονικές autoencoders, χρησιμοποιώντας βιομετρικά και ηχητικά δεδομένα.

Στα βιομετρικά δεδομένα, τα εξατομικευμένα μοντέλα, όπου το κάθε μοντέλο εκπαιδεύτηκε για έναν συγκεκριμένο ασθενή, είχαν γενικά καλύτερη απόδοση σε σχέση με τα μοντέλα που εκπαιδεύτηκαν σε όλους τους ασθενείς. Συγκεκριμένα, οι Συνελικτικοί Autoencoders (CAEs) είχαν την καλύτερη απόδοση στα εξατομικευμένα πειράματα, ενώ οι Πλήρως Συνδεδεμένοι Autoencoders (Fully Connected AEs) ήταν πιο αποδοτικοί σε αυτά που αφορούσαν όλους τους ασθενείς.

Όσον αφορά τα δεδομένα ομιλίας, χρησιμοποιήθηκαν δεδομένα από 8 ασθενείς οι οποίοι είχα παρουσιάσει υποτροπή. Από τα δεδομένα αυτά εξήχθηκαν Mel-φασματογραφήματα και εκπαιδεύτηκαν CAE και CVAE μοντέλα. Όπως παρατηρήθηκε και στα βιομετρικά δεδομένα, τα εξατομικευμένα μοντέλα είχαν συγκριτικά καλύτερη απόδοση, ωστόσο τα CVAEs είχαν καλυτερη απόδοση σε σχέση με τα CAEs, στα πειράματα που αφορούσαν όλους τους ασθενείς.

Πραγματοποιήθηκαν επίσης πειράματα που συνδύαζαν τα βιομετρικά και ηχητικά δεδομένα, με την εκπαίδευση δύο CVAE μοντέλων για το κάθε είδος δεδομένων ξεχωριστά και τον συνδυασμό των μετρικών τους (KL Divergence). Η συνδυαστική ανάλυση των δεδομένων βελτίωσε περαιτέρω την ανίχνευση υποτροπών, γεγονός που καθιστά τα βιομετρικά δεδομένα χρήσιμα για την ανίχνευση ανωμαλιών σε συνδυασμό με τα ηχητικά δεδομένα.

#### Συμπεράσματα και Μελλοντικές Επεκτάσεις

Είναι εμφανές ότι το σύστημα e-Prevention έχει σημειώσει σημαντική πρόοδο στην αποτελεσματική πρόβλεψη πιθανών υποτροπών σε ασθενείς με διπολική διαταραχή και σχιζοφρένεια. Ωστόσο, υπάρχουν περιθώρια περαιτέρω βελτίωσης της απόδοσής του. Η επέκταση του συνόλου δεδομένων με τη συμμετοχή περισσότερων ασθενών θα μπορούσε να βελτιώσει τη γενίκευση των μοντέλων, ενώ η εξέταση διαφορετικών αρχιτεκτονικών autoencoders θα μπορούσε να προσφέρει νέες προοπτικές για την ανίχνευση ανωμαλιών. Επιπλέον, η συνδυαστική χρήση των βιομετρικών και ηχητικών δεδομένων θα μπορούσε να αναλυθεί περαιτέρω, με ένα κοινό μοντέλο που αξιοποιεί αμφότερα τα χαρακτηριστικά τους για την ανίχνευση υποτροπών. Συνεπώς, η παρούσα έρευνα στοχεύει στη διεύρυνση των δυνατοτήτων του e-Prevention, υλοποιώντας τις παραπάνω βελτιώσεις και διερευνώντας νέες αρχιτεκτονικές autoencoders με επιπλέον πηγές δεδομένων.

# Δημιουργία και Επεξεργασία Δεδομένων

## Ηχητικά Δεδομένα

Η ηχητική βάση δεδομένων του e-Prevention περιλαμβάνει ηχογραφήσεις συνεντεύξεων μεταξύ ασθενών και κλινικών ιατρών, καταγεγραμμένες μέσω της εφαρμογής e-Prevention από τον Μάιο του 2020 έως τον Δεκέμβριο του 2021. Αρχικά, η βάση δεδομένων περιείχε 474 συνεντεύξεις από 16 ασθενείς, ταξινομημένες στις εξής κατηγορίες με βάση την κατάσταση του ασθενή την περίοδο της συνέντευξης:

- **Καθαρές (Clean)**: Συνεντεύξεις στις οποίες ο ασθενής δεν είχε παρουσιάσει υποτροπή.
- **Προ-υποτροπής (Pre-Relapse)**: Συνεντεύξεις που προηγήθηκαν της υποτροπής μέχρι 28 ημέρες.
- **Υποτροπής (Relapse)**: Συνεντεύξεις που πραγματοποιήθηκαν κατά τη διάρκεια της υποτροπής.

Αμφότερα τα pre-relapse και relapse δεδομένα θεωρούνται ανωμαλίες στα πλαίσια των πειραμάτων. Επιπλέον, από τους 16 ασθενείς της βάσης, **8 ασθενείς** παρουσίασαν υποτροπή κατά τη διάρκεια της έρευνας.

Το πρώτο βήμα της δικής μας έρευνάς ήταν η επέκταση της βάσης δεδομένων με συνεντεύξεις καινούριων ασθενών που παρουσίασαν υποτροπή και ασθενών που είχαν δεδομένα μετά τις 31 Δεκεμβρίου 2021. Μετά την επέκταση, η βάση περιείχε 555 ηχογραφήσεις από 18 ασθενείς με συνεντεύξεις εώς και τον Μάιο του 2022, εκ των οποίων οι **9 ασθενείς** παρουσίασαν υποτροπή.

#### Προεπεξεργασία και Εξαγωγή Χαρακτηριστικών

Δεδομένου ότι οι ηχογραφήσεις των ασθενών προέρχονταν από βίντεο, πραγματοποιήθηκε εξαγωγή του ήχου και υποδειγματοληψία στα 16 kHz για να διατηρηθεί η ομοιομορφία μεταξύ όλων των ηχογραφήσεων.

Ο διαχωρισμός των ηχογραφήσεων της φωνής των ασθενών από των ιατρών (speaker diarization) πραγματοποιήθηκε με χρήση x-vector embeddings από το εργαλείο Kaldi [Sny+18; Pov+11]. Στη συνέχεια, από τα αποσπάσματα των ηχογραφήσεων των ασθενών εξήχθηκαν λογαριθμικά Mel-φασματογραφήματα με διάσταση 128×64 ανά δευτερόλεπτο.

Η επέκταση αύξησε τον συνολικό αριθμό των αποσπασμάτων σε 16.917, με 735 λεπτά ομιλίας, σε σύγκριση με τις αρχικά 14.562 αποσπάσματα και τα 635 λεπτά ομιλίας. Η ενημερωμένη βάση δεδομένων περιλαμβάνει πλέον 477 clean συνεντεύξεις, 27 pre-relapse συνεντεύξεις και 42 relapse συνεντεύξεις , σε αντίθεση με τις αρχικές 396 clean συνεντεύξεις, 26 pre-relapse συνεντεύξεις και 36 relapse συνεντεύξεις.

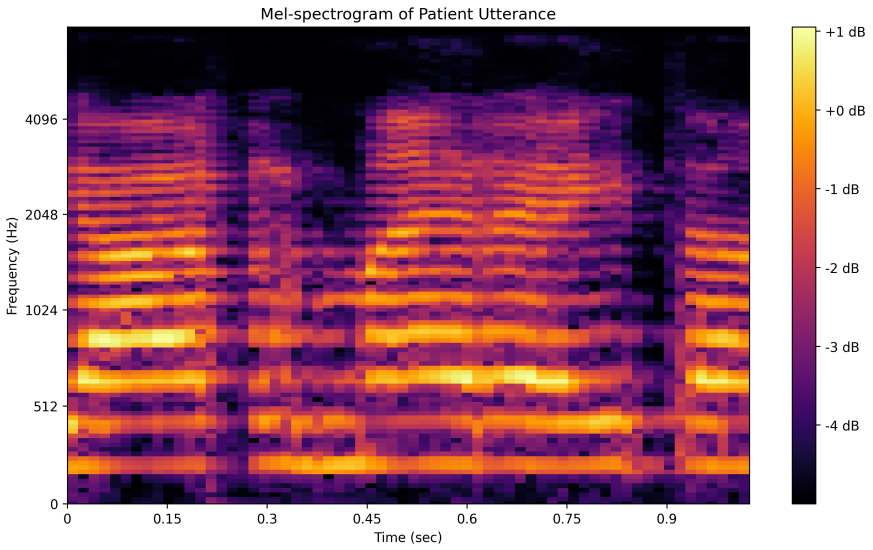



Figure 0.0.12: Παράδειγμα Mel-φασματογραφήματος από απόσμασμα ηχογραφημένης συνέντευξης ασθενούς.

Κατά την επέκταση της βάσης παρουσιάστηκαν κάποιες προκλήσεις όπως ασυνέπειες δεδομένων, σφάλματα στις επισημειώσεις των συνεντεύξεων και στο διαχωρισμό των ομιλητών, οι οποίες αντιμετωπίστηκαν με χειροκίνητες διορθώσεις για διατήρηση της συνέπειας της βάσης.

## Βιομετρικά Δεδομένα

Η βιομετρική βάση δεδομένων e-Prevention περιλαμβάνει δεδομένα από 24 ασθενείς με σχιζοφρένεια και διπολική διαταραχή και σχιζοσυναισθηματική διαταραχή, καταγεγραμμένα μέσω smartwatches [Mag+20]. Τα δεδομένα περιλαμβάνουν μετρήσεις καρδιακού ρυθμού, επιταχυνσιόμετρου και γυροσκοπίου, οι οποίες συλλέχθηκαν μέσω εφαρμογής, υλοποιημένη για τον σκοπό αυτό, και μεταφορτώθηκαν σε cloud server για ανάλυση.

### Αντιστοίχιση Ηχητικών και Βιομετρικών Δεδομένων

Για τις ανάγκες των multimodal πειραμάτων, επιλέχθηκαν οι 9 ασθενείς που παρουσίασαν υποτροπή, σύμφωνα με την ηχητική βάση δεδομένων, με στόχο τον συνδυασμό ηχητικών και βιομετρικών δεδομένων για τη βελτίωση της πρόβλεψης υποτροπών.

Αρχικά, επιλέχθηκαν βιομετρικά δεδομένα τα οποία είχαν καταγραφεί τη μέρα των συνεντεύξεων. Ωστόσο, ο αριθμός των δειγμάτων ήταν περιορισμένος με αποτέλεσμα να διερευνήσουμε τη χρήση δεδομένων που είχαν καταγραφεί γύρω από την ημερομηνία των συνεντεύξεων. Συγκεκριμένα, εξετάσαμε τα δεδομένα που είχαν καταγραφεί 3, 5 και 7 ημέρες πριν και μετά τις συνεντεύξεις, με στόχο τη βελτίωση της πρόβλεψης υποτροπών.

### Επεξεργασία Δεδομένων και Εξαγωγή Χαρακτηριστικών

Για κάθε αισθητήρα τα δεδομένα αποθηκεύτηκαν σε parquet αρχεία και έπειτα σε Pandas dataframes, με το καθένα να αντιστοιχεί σε συγκεκριμένη ημερομηνία. Η συχνότητα δειγματοληψίας του επιταχυνσιόμετρου και του γυροσκοπίου ήταν 20 Hz, ενώ του αισθητήρα του καρδιακού ρυθμού 5 Hz. Για την προεπεξεργασία των δεδομένων ακολουθήθηκε η εξής διαδικασία:

- **Διαθεσιμότητα Δεδομένων:** Για κάθε ημερομηνία, έγινε έλγχος για την ύπαρξη επαρκούς αριθμού δειγμάτων, δηλαδή τουλάχιστον 4 ωρών δεδομένων. Ο αριθμός αυτός επιλέχθηκε με γνώμονα την εξασφάλιση επαρκούς αριθμού δειγμάτων για την εκπαίδευση των μοντέλων και την εξαγωγή σημαντικών χαρακτηριστικών.
- **Τμηματοποίηση Δεδομένων**: Τα δεδομένα των αισθητήρων τμηματοποιήθηκαν σε διαστήματα 8 ωρών. Ο αριθμός αυτός επιλέχθηκε με γνώμονα την εξασφάλιση επαρκούς αριθμού δειγμάτων για την εκπαίδευση των μοντέλων και την εξαγωγή σημαντικών χαρακτηριστικών.

- **Διαγραφή Μη Έγκυρων Δεδομένων:** Δεδομένα με μη εγκυρες τιμές, οι οποίες συγκρίθηκαν με προκαθορισμένα όρια, διαγράφηκαν.

Για κάθε διάστημα 8 ωρών επιλέξαμε να εξάγουμε χαρακτηριστικά εντός διαστημάτων 5 λεπτών (96 διαστήματα των 5 λεπτών), σύμφωνα με προηγούμενη έρευνα [Ret+20]. Τα χαρακτηριστικά που εξήχθηκαν από τα δεδομένα των αισθητήρων είναι τα εξής:

- STE των σημάτων γυροσκοπίου κσι επιταχυνσιόμετρου.
- Μέσος καρδιακός ρυθμός και μέσος όρος των RR διαστημάτων (χρονικά χαρακτηριστικά HRV).
- Περιοδογράφημα Lomb-Scargle για την ανάλυση χαμηλών (0.04-0.15 Hz) και υψηλών συχνοτήτων (0.15-0.4 Hz).
- SD1 του γραφήματος Poincaré.
- Αριθμός των έγκυρων δειγμάτων ανά διάστημα 5 λεπτών.
- Ημιτονοειδής και συνημιτονοειδής αναπαράσταση του χρόνου για την αναπαράσταση των χρονικών μοτίβων στα δεδομένα.

Έπειτα από την εξαγωγή των χαρακτηριστικών, τιμές οι οποίες απουσίαζαν αντικαταστάθηκαν με τη μέση τιμή του εκάστοτε χαρακτηριστικού. Συνολικά, εξήχθησαν 10 χαρακτηριστικά, με αποτέλεσμα τη δημιουργία 96×10 πίνακα χαρακτηριστικών για κάθε διάστημα 8 ωρών.

Το Σχήμα 0.0.13 παρουσιάζει 3 από τα 10 χαρακτηριστικά που εξήχθησαν, συγκεκριμένα την ενέργεια των σημάτων γυροσκοπίου και επιταχυνσιόμετρου και τον μέσο καρδιακό ρυθμό ενός ασθενή κατά τη διάρκεια 8 ωρών.

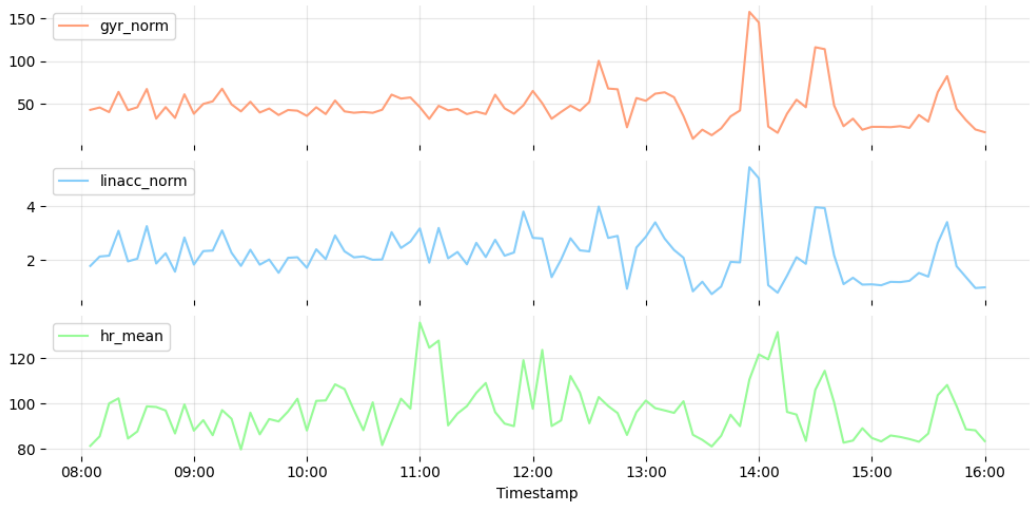



Figure 0.0.13: Παράδειγμα 3 χαρακτηριστικών που εξήχθησαν από τα βιομετρικά δεδομένα ενός ασθενή.

Όπως και στα ηχητικά δεδομένα, προκλήσεις όπως ασυνέπειες μεταξύ των ονομάτων των αρχείων, των ημερομηνιών, του αριθμού του ασθενούς και των δεδομένων των αισθητήρων αντιμετωπίστηκαν με χειροκίνητες διορθώσεις για τη διατήρηση της συνέπειας της βάσης.

## Πειράματα Ηχητικών Δεδομένων

Τα πρώτα πειράματα αυτής της έρευνας επικεντρώθηκαν στη συγκριτική ανάλυση autoencoder μοντέλων για την ανίχνευση υποτροπών σε ασθενείς με ψυχικές διαταραχές με δεδομένα αυθόρμητης ομιλίας. Αρχικά, αξιολογήθηκε η απόδοση των Convolutional Autoencoder (CAE) και Convolutional Variational Autoencoder

(CVAE) μοντέλων, στην νέα επεκταμένη ηχητική βάση δεδομένων. Στη συνέχεια, αναπτύχθηκαν και αξιολογήθηκαν, στην νέα βάση δεδομένων, autoencoders με LSTM αρχιτεκτονική (LSTMAE, LSTMVAE), συγκριτικά με τα συνελικτικά μοντέλα, με στόχο την εξέταση εάν η μοντελοποίηση χρονικών εξαρτήσεων μέσω των LSTM μπορεί να βελτιώσει την την ανίχνευση ανωμαλιών στην ομιλία.

## Μεθοδολογία

### Πειραματικά Σχήματα και Κανονικοποίηση Δεδομένων

Τα πειράματα των autoencoder μοντέλων αποτελούνταν από δύο κατηγορίες: εξατομικευμένα πειράματα, όπου τα μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν σε δεδομένα ενός ασθενή, και "καθολικά" (global) πειράματα, όπου τα μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν σε δεδομένα από όλους τους ασθενείς.

Στα καθολικά πειράματα, αξιολογήθηκαν δύο τεχνικές κανονικοποίησης: (i) κανονικοποίηση ανά ασθενή (per-patient), όπου τα δεδομένα κάθε ασθενούς κανονικοποιούνταν ξεχωριστά και (ii) καθολική (global) κανονικοποίηση όπου τα δεδομένα όλων των ασθενών κανονικοποιούνταν μαζί.

### Αρχιτεκτονικές Μοντέλων

- **Convolutional Autoencoder (CAE)**: Το μοντέλο, υλοποιημένο στα πλαίσια των πειραμάτων του e-Prevention [Gar+21; Zla+22], αποτελείται από 4 downsampling μπλοκ, όπου το καθένα περιλαμβάνει δισδιάστατα συνελικτικά επίπεδα, Max Pooling επίπεδα και ReLU συναρτήσεις ενεργοποίησης, τα οποία συμπιέζουν το αρχικό mel-φασματογράφημα διαστάσεων 128×64 σε έναν λανθάνοντα χώρο χαμηλότερης διάστασης. Για την ανακατασκευή του mel-φασματογραφήματος από τον decoder χρησιμοποιούνται 4 upsampling μπλοκ, με συμμετρική αρχιτεκτονική με τα downsampling μπλοκ, εκτός από τα Max Pooling επίπεδα που αντικαθίστανται από Upsampling επίπεδα.

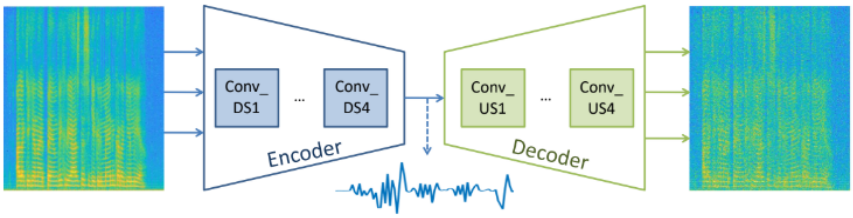



Figure 0.0.14: Προτεινόμενη αρχιτεκτονική CAE μοντέλου για ηχητικά δεδομένα [Gar+21].

- **Convolutional Variational Autoencoder (CVAE)**: Το μοντέλο αποτελεί επέκταση της αρχιτεκτονικής του CAE με πιθανοτική αναπαράσταση του λανθάνοντος χώρου, όπου ο encoder εξάγει τη μέση τιμή $\mu$ και τη λογαριθμική διακύμανση $\log \sigma^2$, τιμές οι οποίες χρησιμοποιούνται για τη δειγματοληψία αναπαραστάσεων (embeddings) από τον λανθάνοντα χώρο.

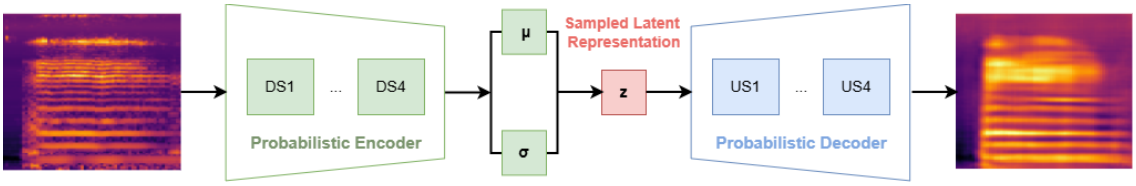



Figure 0.0.15: Προτεινόμενη αρχιτεκτονική CVAE μοντέλου για ηχητικά δεδομένα.

- **Long Short-Term Memory Autoencoder (LSTMAE)**: Η επιλογή υλοποίησης autoencoder αρχιτεκτονικής με LSTM επίπεδα είναι εμπνευσμένη από το έργο των Mobtahej et al. [Mob+22], που έδειξε σημαντική ικανότητα των LSTM να μοντελοποιούν χρονικές εξαρτήσεις σε ηχητικά δεδομένα με σκοπό την ανίχνευση ανωμαλιών. Το μοντέλο αποτελείται από έναν LSTM encoder που συμπιέζει mel-φασματογραφήματα διαστάσεων 64×128 και περιλαμβάνει επίπεδα κανονικοποίησης, Leaky ReLU

συνάρτηση ενεργοποίησης (η οποία διαφέρει από την ReLU συνάρτηση ενεργοποίησης στο γεγονός ότι επιτρέπει μικρές αρνητικές τιμές) και επίπεδα dropout, όπου ένας καθορισμένος αριθμός νευρώνων απενεργοποιείται τυχαία κατά την εκπαίδευση για την αποφυγή υπερεκπαίδευσης. Ο decoder αποτελείται από LSTM επίπεδα και αντίστοιχα επίπεδα κανονικοποίησης, Leaky ReLU και dropout.

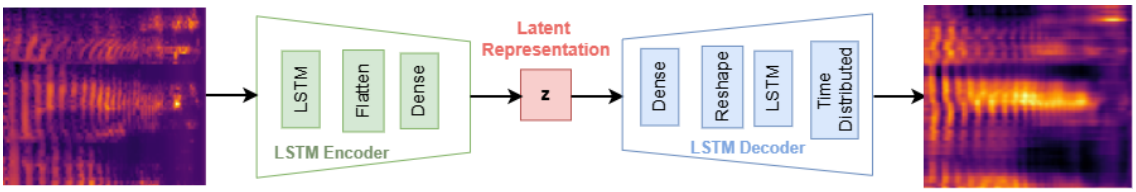



Figure 0.0.16: Προτεινόμενη αρχιτεκτονική LSTMAE μοντέλου για ηχητικά δεδομένα.

- **Long Short-Term Memory Variational Autoencoder (LSTMVAE)**: Αντίστοιχα με το CVAE, το LSTMVAE μοντέλο αποτελεί επέκταση της αρχιτεκτονικής του LSTMAE με πιθανοτική αναπαράσταση του λανθάνοντος χώρου.

### Εκπαίδευση και Αξιολόγηση

Για την εκπαίδευση των μοντέλων, χρησιμοποιήθηκε η μέθοδος της διασταυρούμενης επικύρωσης, με 5 επαναλήψεις (5-fold cross validation), και τα δεδομένα εκπαίδευσης αποτελούνταν αποκλειστικά από clean δεδομένα ομιλίας, έτσι ώστε τα μοντέλα να μάθουν τα βασικά πρότυπα της φυσιολογικής ομιλίας. Στα CAE και LSTMAE μοντέλα, χρησιμοποιήθηκε το MSE για την ελαχιστοποίηση των απωλειών, ενώ για τα CVAE και LSTMVAE, η KL-απόκλιση προστέθηκε στις απώλειες για την ρύθμιση του λανθάνοντα χώρου.

Τα μοντέλα αξιολογήθηκαν σε clean, pre-relapse και relapse δεδομένα. Συγκεκριμένα, για κάθε ημερομηνία συνέντευξης, τα μοντέλα παρήγαγαν ένα σκορ ανωμαλίας (anomaly score) για κάθε mel-φασματογράφημα της συνεδρίας, τα οποία στη συνέχεια συγκεντρώθηκαν χρονικά για να παραχθεί το τελικό anomaly score της συνεδρίας. Για τα CAE και LSTMAE χρησιμοποιήθηκε το MSE ως anomaly score ενώ για τα CVAE και LSTMVAE εφαρμόστηκαν τόσο το MSE όσο και η KL-απόκλιση. Η τελική αξιολόγηση, βασίστηκε στα μέσα (median) anomaly scores μεταξύ των 5 επαναλήψεων και το μέσο (mean) ROC-AUC score. Ωστόσο, για την παρουσίαση των αποτελεσμάτων στην παρούσα περίληψη θα χρησιμοποιηθεί το μέσο ROC-AUC score.

## Αποτελέσματα

### Σύγκριση Απόδοσης του CAE Μοντέλου στο Αρχικό και Επεκταμένο Σύνολο Δεδομένων

Στο Σχήμα 0.0.17 παρουσιάζεται η σύγκριση της απόδοσης του CAE μοντέλου στο αρχικό σύνολο δεδομένων του e-Prevention και στο σύνολο δεδομένων που επεκτείναμε. Ειδικότερα, απεικονίζονται τα ROC-AUC scores των εξατομικευμένων μοντέλων για κάθέναν από τους 9 ασθενείς, με τον ασθενή #8 να αποτελεί τη νέα προσθήκη, και ως εκ τούτου, να εμφανίζεται μόνο μετά την επέκταση. Επιπλέον, παρουσιάζεται το μέσο ROC-AUC score για τους 8 και τους 9 ασθενείς. Στις δύο τελευταίες στήλες, φαίνονται τα αποτελέσματα των καθολικών (global) CAE μοντέλων, στα οποία εφαρμόστηκαν οι δύο προσεγγίσεις κανονικοποίησης που αναλύθηκαν προηγουμένως: per-patient και global κανονικοποίηση. Επίσης, να σημειωθεί ότι τα αποτελέσματα του αρχικού συνόλου δεδομένων προκύπτουν από το άρθρο του έργου του e-Prevention [Zla+22].

Η ανάλυση των αποτελεσμάτων των εξατομικευμένων μοντέλων δείχνει βελτίωση της ικανότητας ανίχνευσης ανωμαλιών για τους περισσότερους ασθενείς. Συγκεκριμένα, τα ROC-AUC scores παρουσίασαν αύξηση για ασθενείς όπως οι #1, #7 και #9, ενώ η μέση τιμή ROC-AUC για τους 8 αρχικούς ασθενείς βελτιώθηκε από 0.667 σε 0.684, επιβεβαιώνοντας ότι η προσθήκη δεδομένων συνέβαλε στη βελτίωση της διάκρισης μεταξύ των καθαρών καταστάσεων και των καταστάσεων υποτροπής. Ωστόσο, για ορισμένους ασθενείς όπως ο #2 και #4, η απόδοση παρέμεινε σχεδόν αμετάβλητη, υποδηλώνοντας ότι η επίδραση της επέκτασης μπορεί να εξαρτάται από τα ατομικά χαρακτηριστικά της ομιλίας ή τον όγκο των προστιθέμενων δεδομένων. Η μεγαλύτερη βελτίωση παρατηρείται στα global πειράματα, όπου έχουμε σημαντική αύξηση των ROC-AUC scores από 0.531 σε 0.618 για την per-patient κανονικοποίηση και από 0.525 σε 0.633 για την global κανονικοποίηση, δείχνοντας ότι

το μοντέλο έγινε πιο ικανό στη διάκριση μεταξύ των καταστάσεων σε όλους τους ασθενείς. Συνεπώς, τα αποτελέσματα επιβεβαιώνουν ότι η επέκταση του συνόλου δεδομένων ενίσχυσε συνολικά την απόδοση του CAE μοντέλου.

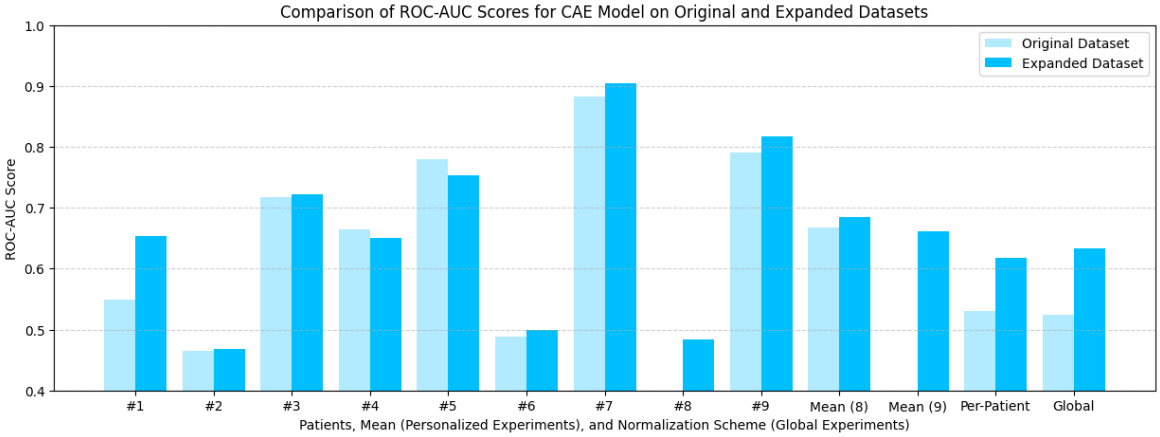



Figure 0.0.17: Σύγκριση των ROC-AUC scores του CAE μοντέλου στο αρχικό και επεκταμένο σύνολο δεδομένων για όλους τους ασθενείς στο εξατομικευμένο πειραματικό σχήμα, και τις δύο μεθόδους κανονικοποίσης (per-patient, global) του global πειραματικού σχήματος.

### Σύγκριση Απόδοσης του CVAE Μοντέλου στο Αρχικό και Επεκταμένο Σύνολο Δεδομένων

Στα Σχήματα 0.0.18 και 0.0.19 παρουσιάζεται η σύγκριση των ROC-AUC scores του CVAE μοντέλου στο αρχικό και επεκταμένο σύνολο δεδομένων, για τα MSE και KL anomaly scores αντίστοιχα, για όλα τους ασθενείς και πειραματικά σχήματα.

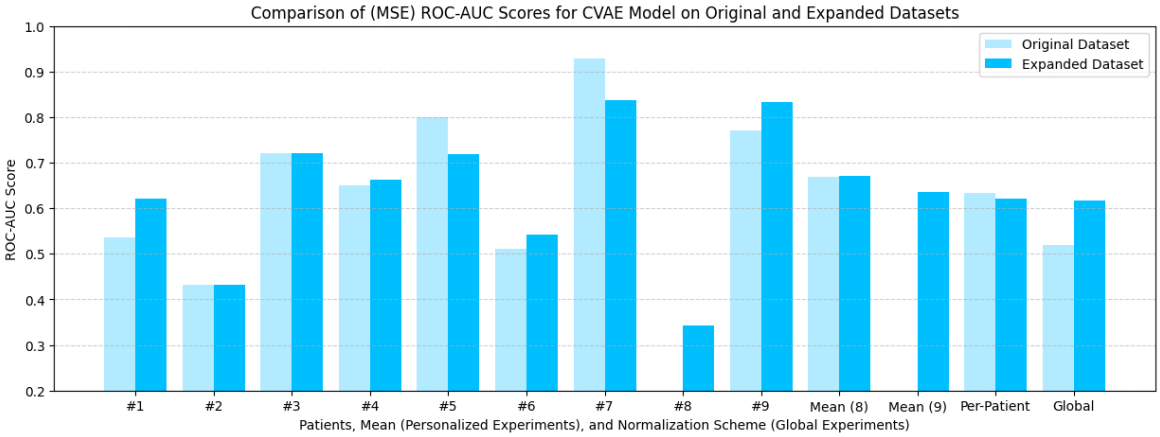



Figure 0.0.18: Σύγκριση των (MSE) ROC-AUC scores του CVAE μοντέλου στο αρχικό και επεκταμένο σύνολο δεδομένων για όλους τους ασθενείς στο εξατομικευμένο πειραματικό σχήμα, και τις δύο μεθόδους κανονικοποίσης (per-patient, global) του global πειραματικού σχήματος.

Από τα αποτελέσματα του CVAE μοντέλου στο επεκταμένο σύνολο δεδομένων, παρατηρούνται ήπιες βελτιώσεις στην ικανότητα ανίχνευσης ανωμαλιών, αν και η βελτίωση δεν είναι τόσο έντονη όσο στο CAE μοντέλο. Στα εξατομικευμένα πειράματα, το ROC-AUC score βελτιώθηκε ή παρέμεινε σχεδόν σταθερό για 6 από τους 8 ασθενείς του αρχικού συνόλου δεδομένων, ενώ τα μέσα MSE και KL ROC-AUC scores παρουσίασαν μικρές μεταβολές. Η απόδοση του μοντέλου για τον νέο ασθενή #8 ήταν αρκετά χαμηλή, γεγονός που επηρέασε αρνητικά τη συνολική απόδοση στο σύνολο των 9 ασθενών. Στα global πειράματα, παρατηρείται βελτίωση της

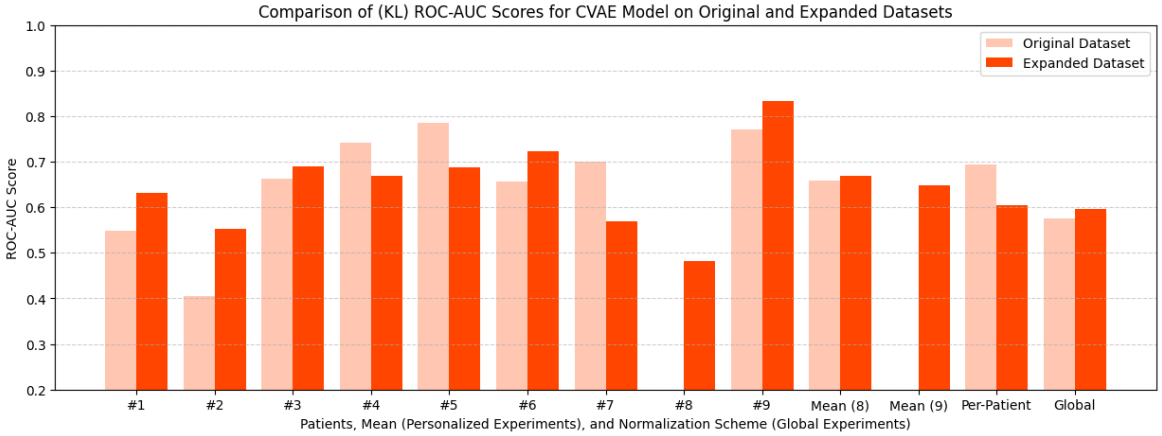



Figure 0.0.19: Σύγκριση των (KL) ROC-AUC scores του CVAE μοντέλου στο αρχικό και επεκταμένο σύνολο δεδομένων για όλους τους ασθενείς στο εξατομικευμένο πειραματικό σχήμα, και τις δύο μεθόδους κανονικοποίησης (per-patient, global) του global πειραματικού σχήματος.

απόδοσης για την global κανονικοποίηση, με το MSE ROC-AUC score να αυξάνεται από 0.519 σε 0.619, και το KL ROC-AUC score από 0.576 σε 0.594. Συνολικά, η επέκταση του συνόλου δεδομένων επηρέασε θετικά το CVAE μοντέλο, ωστόσο η αύξηση της απόδοσής του ήταν μικρότερη σε σύγκριση με το CAE μοντέλο. Αυτό ενδέχεται να οφείλεται στη μεγαλύτερη ευαισθησία του CVAE στη μεταβλητότητα των δεδομένων, η οποία πιθανώς αυξήθηκε μετά την επέκταση.

### Σύγκριση Απόδοσης των CAE και LSTMAE Μοντέλων στο Επεκταμένο Σύνολο Δεδομένων

Στο Σχήμα 0.0.20 παρουσιάζεται η σύγκρισης της απόδοσης των CAE και LSTMAE μοντέλων στις ίδιες περιπτώσεις που αναλύθηκαν προηγουμένως.

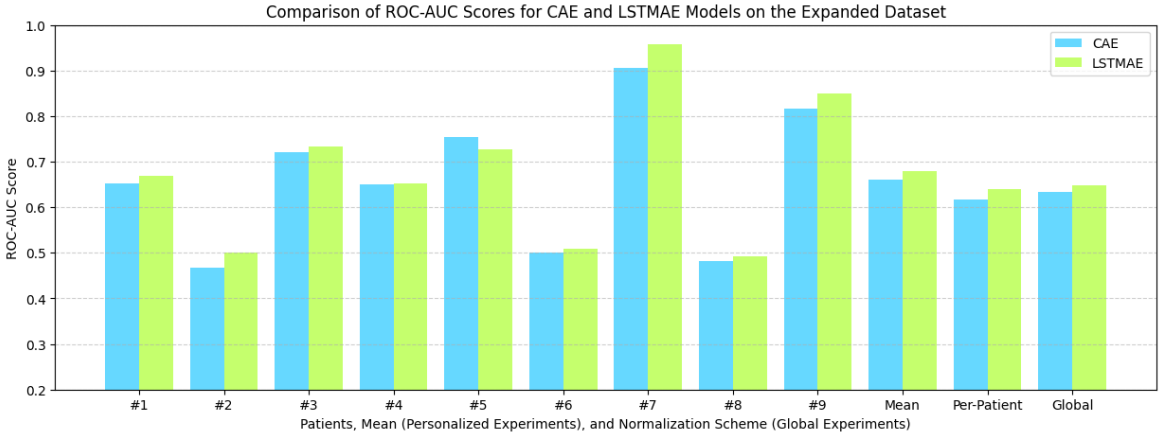



Figure 0.0.20: Σύγκριση των ROC-AUC scores των CAE και LSTMAE μοντέλων στο επεκταμένο σύνολο δεδομένων για όλους τους ασθενείς στο εξατομικευμένο πειραματικό σχήμα, και τις δύο μεθόδους κανονικοποίησης (per-patient, global) του global πειραματικού σχήματος.

Από τα αποτελέσματα του Σχήματος 0.0.20, παρατηρούμε ότι το LSTMAE υπερέχει στις περισσότερες περιπτώσεις, υποδεικνύοντας τα οφέλη της μοντελοποίησης χρονικών εξαρτήσεων, που προσφέρει η LSTM αρχιτεκτονική, στην ανίχνευση ανωμαλιών. Συγκεκριμένα, στα εξατομικευμένα πειράματα, το LSTMAE εμφάνισε υψηλότερα ROC-AUC scores για τους περισσότερους ασθενείς, με τη μέση τιμή να αυξάνεται από 0.661 σε 0.679,

ενώ παρατηρούνται αξιοσημείωτες βελτιώσεις για συγκεκριμένους ασθενείς, όπως ο ασθενής #7. Στα global πειράματα, το LSTMAE διατηρεί την υπεροχή του, με τα ROC-AUC scores να βελτιώνονται τόσο για την per-patient όσο και για την global κανονικοποίηση. Συνολικά, το LSTMAE, αξιοποιώντας τις χρονικές σχέσεις στα δεδομένα, φαίνεται να βελτιώνει την ανίχνευση υποτροπών στο εκτεταμένο σύνολο δεδομένων, ωστόσο, το CAE εξακολουθεί να αποτελεί αποτελεσματική προσέγγιση.

### Σύγκριση Απόδοσης των CVAE και LSTMAVE Μοντέλων στο Επεκταμένο Σύνολο Δεδομένων

Τα Σχήματα 0.0.21 και 0.0.22 παρουσιάζουν τη σύγκριση των MSE και KL ROC-AUC scores των CVAE και LSTMVAE μοντέλων στο επεκταμένο σύνολο δεδομένων.

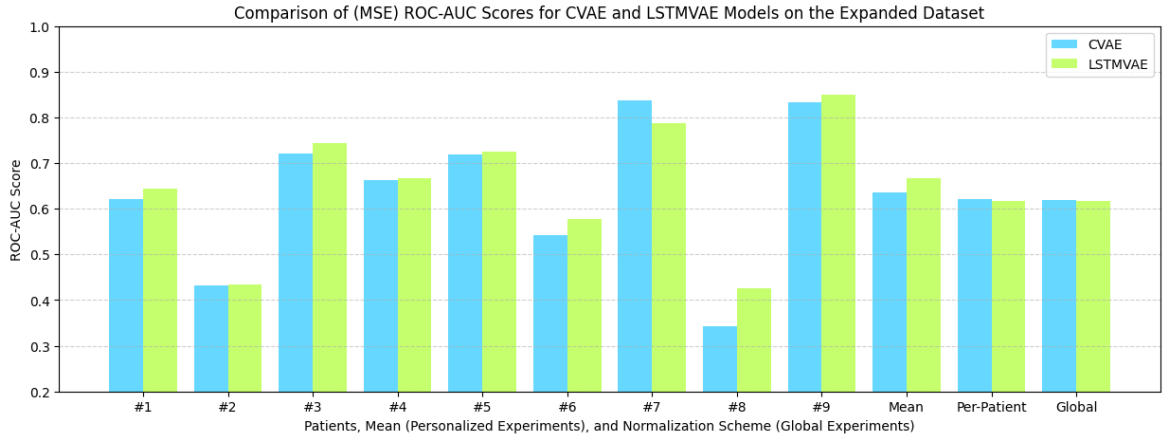



Figure 0.0.21: Σύγκριση των (MSE) ROC-AUC των scores των CVAE και LSTMVAE μοντέλων στο επεκταμένο σύνολο δεδομένων για όλους τους ασθενείς στο εξατομικευμένο πειραματικό σχήμα, και τις δύο μεθόδους κανονικοποίησης (per-patient, global) του global πειραματικού σχήματος.

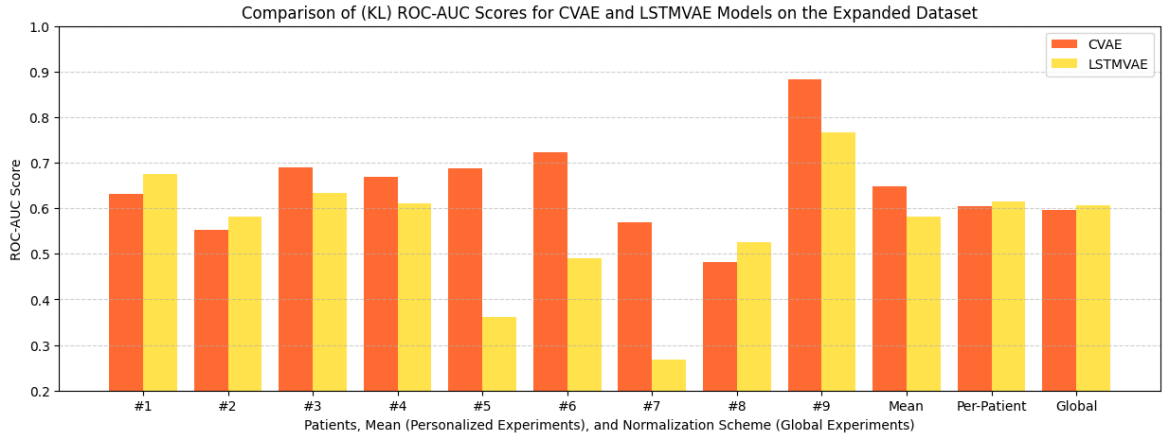



Figure 0.0.22: Σύγκριση των (KL) ROC-AUC των scores των CVAE και LSTMVAE μοντέλων στο επεκταμένο σύνολο δεδομένων για όλους τους ασθενείς στο εξατομικευμένο πειραματικό σχήμα, και τις δύο μεθόδους κανονικοποίησης (per-patient, global) του global πειραματικού σχήματος.

Η σύγκριση των LSTMVAE και CVAE δείχνει ότι το CVAE αποδίδει καλύτερα στη διάκριση ανώμαλων καταστάσεων μέσω KL-απόκλισης, ενώ το LSTMVAE υπερέχει ελαφρώς στην ανίχνευση ανωμαλιών μέσω του MSE. Στα εξατομικευμένα πειράματα, το LSTMVAE εμφάνισε υψηλότερα MSE ROC-AUC scores, αλλά το CVAE είχε σημαντικά καλυτερη απόδοση όσον αφορά την KL-απόκλιση. Στα global πειράματα, τα KL ROC-AUC scores του LSTMVAE ήταν ελαφρώς υψηλότερα, αν και τα MSE ROC-AUC scores ήταν κοντά σε αυτά

του CVAE. Συνολικά, φαίνεται πως αν και το LSTMVAE είναι ικανό στην ανίχνευση υποτροπών, το CVAE αποδίδει καλύτερα στο συγκεκριμένο σύνολο δεδομένων.

Συνοψίζοντας, η επέκταση του συνόλου δεδομένων βελτίωσε την ανίχνευση ανωμαλιών σε όλα τα μοντέλα, με διαφορετικό βαθμό βελτίωσης ανάλογα με την αρχιτεκτονική και το πειραματικό πλαίσιο. Το CAE και το LSTMAE παρουσίασαν σημαντική αύξηση στα ROC-AUC scores του, με το LSTMAE να υπερέχει, αξιοποιώντας της ακολουθιακές και χρονικές εξαρτήσεις των δεδομένων. Το CVAE αναδείχθηκε ως προτιμότερη variational προσέγγιση του προβλήματος, ωστόσο το LSTMVAE παρουσίασε συγκρίσιμη επίδοση, γεγονός που το καθιστά μια εναλλακτική επιλογή.

# Πειράματα Πολλαπλών Πηγών Δεδομένων (Multimodal)

Το κύριο μέρος της έρευνας αφορά τον συνδυασμό των ηχητικών και βιομετρικών δεδομένων αποσκοπώντας στη βελτίωση της ανίχνευσης υποτροπών σε ασθενείς με ψυχικές διαταραχές. Για τον σκοπό αυτό, αναπτύχθηκαν δύο μοντέλα συνδυαστικών (joint) autoencoders, καθένα με ξεχωριστούς κλάδους για τα ηχητικά και βιομετρικά δεδομένα, με τον συνδυασμό τους (fusion) να πραγματοποιείται στον λανθάνοντα χώρο ώστε να δημιουργηθεί μια κοινή αναπαράσταση. Επιπλέον, πραγματοποιήθηκαν πειράματα για να αξιολογηθεί η ανεξάρτητη συμβολή κάθε πηγής δεδομένων και η επίδραση της προ-εκπαίδευσης των μοντέλων στην απόδοσή τους.

## Μεθοδολογία

### Αντιστοίχιση Ηχητικών και Βιομετρικών Δεδομένων

Για την εκπαίδευση και αξιολόγηση των joint μοντέλων, η ακριβής αντιστοίχιση των ηχητικών και βιομετρικών δεδομένων ήταν κρίσιμη. Αρχικά, τα βιομετρικά δεδομένα αντιστοιχίστηκαν στις ακριβείς ημερομηνίες των συνεντεύξεων, δημιουργώντας ένα μικρό σύνολο δεδομένων "day-of" που περιλάμβανε τέσσερις ασθενείς με επαρκή και έγκυρα δεδομένα. Ωστόσο, το περιορισμένο μέγεθος αυτού του συνόλου δεδομένων δεν επέτρεπε αξιόπιστη εκπαίδευση των μοντέλων. Για την αντιμετώπιση αυτού του ζητήματος, συμπεριλήφθηκαν επιπλέον βιομετρικά δεδομένα μέσα σε χρονικά παράθυρα γύρω από τις ημερομηνίες των συνεντεύξεων, οδηγώντας στη δημιουργία των συνόλων δεδομένων 3-day, 5-day και 7-day. Ο Πίνακας 1 παρουσιάζει τα δημογραφικά στοιχεία, πληροφορίες σχετικά με τα σύνολα δεδομένων day-of, 3-day, 5-day και 7-day, μετά την προ-επεξεργασία και εξαγωγή χαρακτηριστικών.

| **Datasets** | **Day-of** | **3-day** | **5-day** | **7-day** |
|---|---|---|---|---|
| **Demographics** | | | | |
| Male/Female | 2/2 | 2/3 | 2/3 | 3/4 |
| Age (years) | 31 ± 8.7 | 30.2 ± 8 | 30.2 ± 8 | 28 ± 7.6 |
| Education (years) | 14 ± 2 | 14.4 ± 2 | 14.4 ± 2 | 13.7 ± 2 |
| Illness duration (years) | 8.8 ± 9 | 7.2 ± 8.6 | 7.2 ± 8.6 | 6.4 ± 7.4 |
| **Recorded Data** | | | | |
| Num. of Days Recorded (total) | 66 | 102 | 124 | 158 |
| Num. of Days Recorded (mean ± std) | 16.5 ± 5 | 20.4 ± 7.5 | 24.8 ± 7.2 | 22.6 ± 10.8 |
| Num. of Hours Recorded (total) | 888 | 1,752 | 2,400 | 3,280 |
| Num. of Hours Recorded (mean ± std) | 222 ± 45.7 | 350.4 ± 89.6 | 480 ± 96.7 | 468.6 ± 185 |
| Num. of 5-min intervals (total) | 10,656 | 21,024 | 28,800 | 39,360 |
| Num. of 5-min intervals (mean ± std) | 2,664 ± 548.9 | 4,204.8 ± 1,074.9 | 5,760 ± 1171 | 5,622.9 ± 2,219.5 |

Table 1: Σύγκριση δημογραφικών στοιχείων, πληροφοριών για την ασθένεια και καταγεγραμμένων βιομετρικών δεδομένων για τους ασθενείς στα σύνολα δεδομένων day-of, 3-day, 5-day, και 7-day, μετά την προ-επεξεργασία και εξαγωγή χαρακτηριστικών.

Η αντιστοίχιση των δύο τύπων δεδομένων παρουσίασε προκλήσεις λόγω των θεμελιωδών διαφορών στη χρονική τους δομή και στα χαρακτηριστικά, γεγονός που καθιστούσε δύσκολη την απευθείας αντιστοίχιση, καθώς δεν υπήρχε απευθείας σχέση 1-προς-1 μεταξύ των δύο τύπων δεδομένων. Επιπλέον, ο αριθμός των mel-φασματογραφημάτων ανά ημερομηνία συνέντευξης ήταν σημαντικά μεγαλύτερος από τον αριθμό των διαθέσιμων

βιομετρικών δεδομένων. Για να αντιμετωπιστεί αυτή η διαφορά, κάθε mel-φασματογράφημα αντιστοιχίστηκε σε πολλαπλά στιγμιότυπα των βιομετρικών δεδομένων από την ίδια ημερομηνία. Αυτή η διαδικασία εξασφάλισε επαρκή ποικιλία στα στοιχισμένα δεδομένα, διατηρώντας παράλληλα τη χρονική συσχέτιση μεταξύ των δύο.

### Αρχιτεκτονικές Joint Autoencoder

Οι joint autoencoder αρχιτεκτονικές που αναπτύχθηκαν για τα multimodal πειράματα συνδυάζουν δύο ξεχωριστούς κλάδους: έναν για τα ηχητικά δεδομένα και έναν για τα βιομετρικά, χρησιμοποιώντας διαφορετικούς autoencoders για κάθε τύπο δεδομένων.

Ο ηχητικός κλάδος περιλαμβάνει τις δύο αρχιτεκτονικές autoencoder, τον CAE και LSTMAE, που αξιολογήθηκαν στα πειράματα για την ανίχνευση ανωμαλιών σε δεδομένα ομιλίας.

Ο κλάδος των βιομετρικών δεδομένων βασίζεται σε ένα CAE μοντέλο (Σχήμα 0.0.23), το οποίο αναπτύχθηκε με βάση την καλύτερη αρχιτεκτονική των πειραμάτων του e-Prevention για ανίχνευση υποτροπων σε βιομετρικά δεδομένα [Zla+22]. Το μοντέλο δέχεται ως είσοδο τα βιομετρικά διανύσματα χαρακτηριστικών διάστασης $96 \times 10$, τα οποία παρουσιάστηκαν αναλυτικά παραπάνω. Ο αρχιτεκτονική του encoder περιλαμβάνει 4 συνελικτικά επίπεδα, επίπεδα κανονικοποίησης παρτίδας (batch normalization) και LeakyReLU συναρτήσεις ενεργοποίησης, ενώ ο decoder εφαρμόζει αντίστροφη διαδικασία με Upsampling και Dense επίπεδα ώστε να ανακατασκευάσει τα αρχικά βιομετρικά χαρακτηριστικά.

Συνολικά, ο encoder του κάθε κλάδου παράγει μία συμπιεσμένη αναπαράσταση των δεδομένων εισόδου του, η οποία στη συνέχεια συνδυάζεται για να δημιουργηθεί ένας κοινός λανθάνοντας χώρος από τον οποίο οι αντίστοιχοι decoders θα προσπαθήσουν να ανακατασκευάσουν την είσοδο με την πλεονάζουσα πληροφορία που προσέφερε ο διαφορετικός τύπος δεδομένων. Στα Σχήματα 0.0.24 και 0.0.25 παρουσιάζονται οι αρχιτεκτονικές των CAE-CAE και LSTMAE-CAE μοντέλων αντίστοιχα.

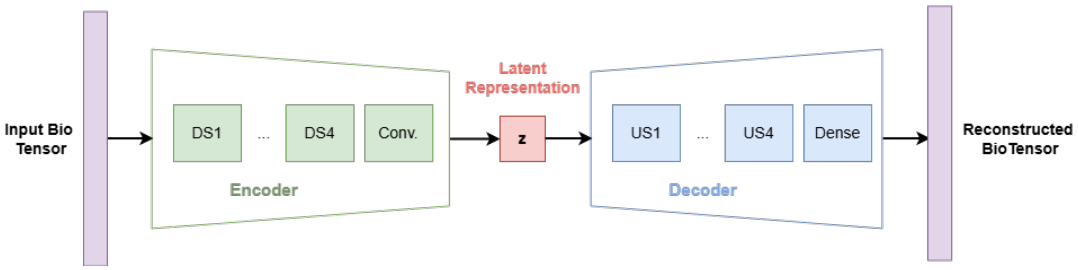



Figure 0.0.23: Επισκόπηση της προτεινόμενης CAE για τον κλάδο των βιομετρικών δεδομένων.

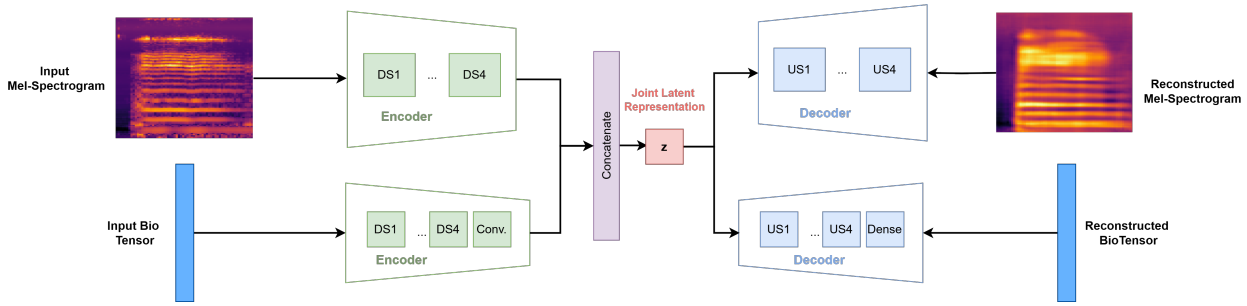



Figure 0.0.24: Επισκόπηση της προτεινόμενης joint αρχιτεκτονικής CAE-CAE.

### Εκπαίδευση και Αξιολόγηση

Η εκπαίδευση των joint autoencoder μοντέλων ακολούθησε την ίδια βασική μεθοδολογία με τα προηγούμενα πειράματα, με την προσθήκη βαρών στους κλάδους των joint μοντέλων που καθορίζουν τη συνεισφορά του κάθε τύπου δεδομένων. Για την αξιολόγηση, αρχικά εκπαιδεύτηκαν τα μοντέλα του κάθε τύπου δεδομένων (unimodal) σε κάθε σύνολο δεδομένων με σκοπό τη χρήση των αποτελεσμάτων της απόδοσής του ως σημείο αναφοράς στην αξιολόγηση των multimodal μοντέλων. Οι μετρικές αξιολόγησης ήταν τα MSE anomaly scores

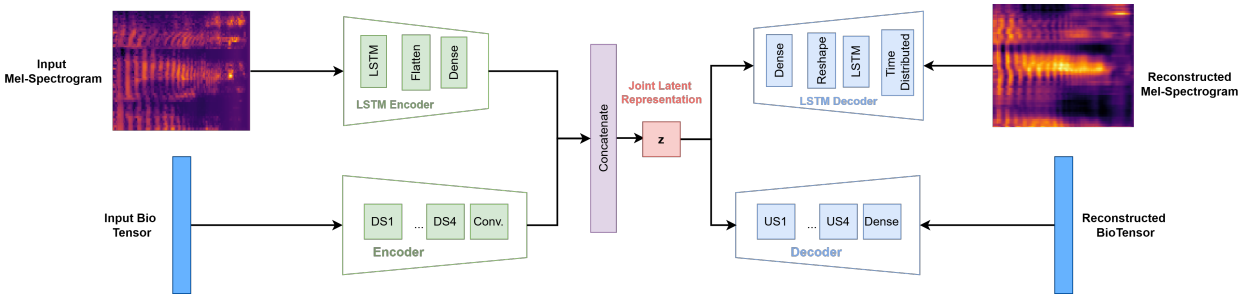



Figure 0.0.25: Επισκόπηση της προτεινόμενης joint αρχιτεκτονικής LSTMAE-CAE.

και ROC-AUC scores των κλάδων, αλλά και ένα συνδυαστικό MSE και ROC-AUC score, που προκύπτει από την πρόσθεση των επιμέρους scores πολλαπλασιασμένα με το βάρος του κάθε κλάδου, με σκοπό την αξιολόγηση της συνολικής απόδοσης του μοντέλου.

## Αποτελέσματα

### Joint CAE-CAE Μοντέλο

Στο Σχήμα 0.0.26 παρουσιάζονται τα ROC-AUC αποτελέσματα του joint CAE-CAE μοντέλου στα εξατομικευμένα πειράματα, για κάθε σύνολο δεδομένων. Αρχικά, παρατηρούμε ότι στο day-of σύνολο δεδομένων υπάρχει σημαντική ανισορροπία στην απόδοση των unimodal μοντέλων, και κατ'επέκταση και στους αντίστοιχους κλάδους. Το γεγονός αυτό ήταν ο λόγος που αποφασίσαμε να συμπεριλάβουμε δεδομένα εντός ενός χρονικού παραθύρου γύρω από την ημέρα των συνεντεύξεων, το οποίο μπορούμε να παρατηρήσουμε ότι ήταν αποτελεσματικό αφού φαίνεται πως η ανισορροπία στη συνεισφορά του κάθε τύπου δεδομένων μειώθηκε. Ωστόσο, ακόμη και με την ύπαρξη της ανισορροπίας, μπορούμε να δούμε ότι ο συνδυασμός των δύο πηγών δεδομένων βελτίωσε την απόδοση του κάθε κλάδου ξεχωριστά στο day-of σύνολο δεδομένων, αλλά η συνολική απόδοση είναι βελτιωμένη συγκριτικά μόνο με το βιομετρικό μοντέλο. Για τα υπόλοιπα σύνολα δεδομένων, παρατηρούμε ότι το joint μοντέλο υπερείχε σταθερά έναντι των unimodal μοντέλων, με τον ηχητικό και τον βιομετρικό κλάδο να επιδεικνύουν βελτιωμένη ικανότητα ανίχνευσης ανωμαλιών. Επομένως, τα αποτελέσματα αυτά δείχνουν ότι ο συνδυασμός ηχητικών και βιομετρικών δεδομένων είναι ικανός να ενισχύσει την ικανότητα ανίχνευσης υποτροπών σε ασθενείς με ψυχικές διαταραχές.

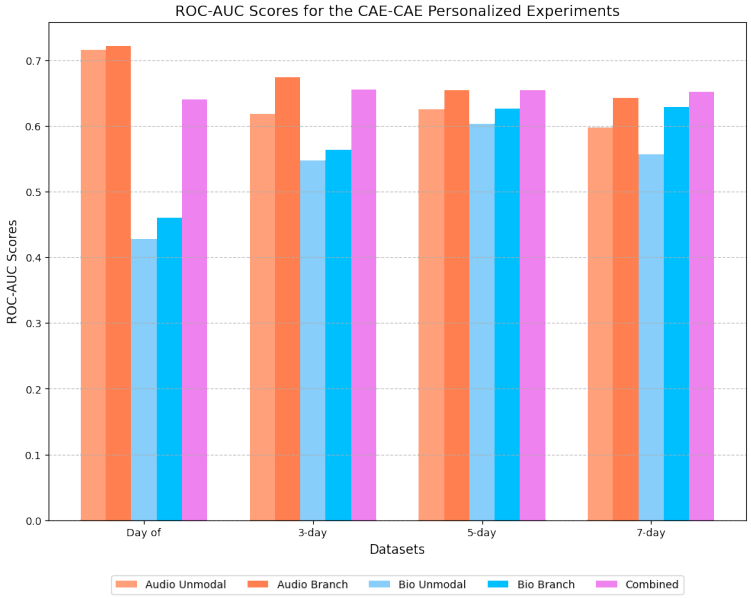



Figure 0.0.26: Μέσα ROC-AUC scores για το joint CAE-CAE μοντέλο στα εξατομικευμένα πειράματα.

Το Σχήμα 0.0.27 παρουσιάζει τα ROC-AUC scores του joint CAE-CAE μοντέλου στα global πειράματα με per-patient κανονικοποίηση, ενώ το Σχήμα 0.0.28 παρουσιάζει τα αντίστοιχα αποτελέσματα με global κανονικοποίηση.

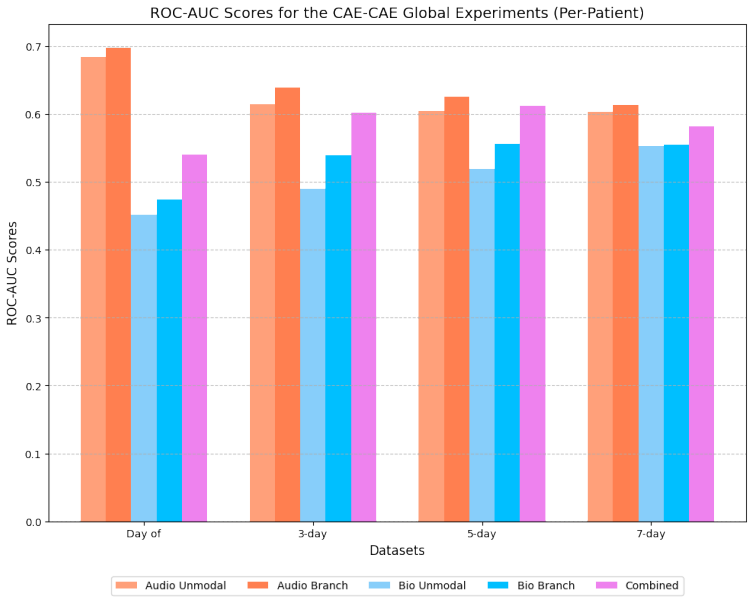



Figure 0.0.27: Μέσα ROC-AUC scores για το joint CAE-CAE μοντέλο στα global πειράματα με per-patient κανονικοποίηση.

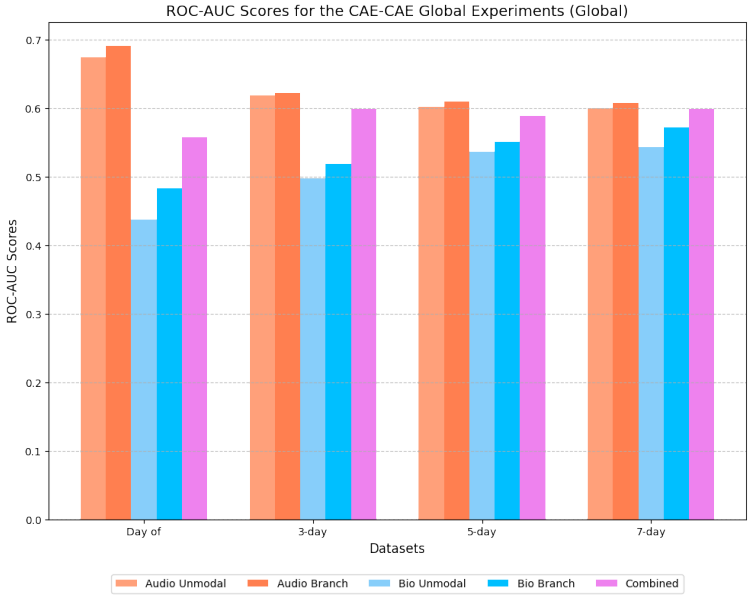



Figure 0.0.28: Μέσα ROC-AUC scores για το joint CAE-CAE μοντέλο στα global πειράματα με global κανονικοποίηση.

Στα πειράματα με per-patient κανονικοποίηση, το joint μοντέλο CAE-CAE παρουσίασε μέτριες βελτιώσεις σε σχέση με τα unimodal μοντέλα, με τον ηχητικό και τον βιομετρικό κλάδο να επιτυγχάνουν σταθερά υψηλότερες βαθμολογίες ROC-AUC. Ωστόσο, οι βελτιώσεις ήταν λιγότερο έντονες από αυτές των εξατομικευμένων πειραμάτων. Οι συνδυασμένες τιμές ROC-AUC παρουσίασαν διακυμάνσεις μεταξύ των συνόλων δεδομένων, με το 5-day σύνολονα παρουσιάζει τη μεγαλύτερη βελτίωση έναντι και των δύο unimodal μοντέλων, ενώ στα σύνολα

3 και 7 ημερών το joint μοντέλο υπερείχε κυρίως του βιομετρικού unimodal μοντέλου. Τα αποτελέσματα αυτά δείχνουν ότι η per-patient κανονικοποίηση επωφελείται από τον συνδυασμό, αλλά η απόδοση είναι μειωμένη σε σύγκριση με τα εξατομικευμένα μοντέλα, πιθανώς λόγω της διαφοροποίησης μεταξύ των ασθενών.

Στη global κανονικοποίηση, οι βελτιώσεις στην ανίχνευση ανωμαλιών φαίνεται να είναι πιο περιορισμένες, με τους κλάδους να παρουσιάζουν μικρές βελτιώσεις στην απόδοση σε σύγκριση με τα unimodal μοντέλα. Επιπλέον, η συνολική απόδοση να είναι υψηλότερη συγκριτικά μόνο με το βιομετρικό unimodal μοντέλο, γεγονός που υποδηλώνει ότι ο συνδυασμός των δεδομένων είχε περιορισμένο αντίκτυπο σε αυτή την κανονικοποίηση. Τα αποτελέσματα αυτά δείχνουν ότι η απόδοση του joint μοντέλου στην global κανονικοποίηση είναι μεν βελτιωμένη σε σύγκριση με τα unimodal μοντέλα, αλλά δεν αξιοποιούνται πλήρως οι δυνατότητες της multimodal προσέγγισης, όπως στην per-patient κανονικοποίηση.

Συνολικά, ο συνδυασμός των δύο τύπων δεδομένων αποδείχθηκε επωφελής για την ανίχνευση υποτροπών, με τα εξατομικευμένα μοντέλα να παρουσιάζουν τις πιο σημαντικές βελτιώσεις. Τα ευρήματα αυτά υπογραμμίζουν την αξία της multimodal προσέγγισης για την εξατομικευμένη παρακολούθηση της ψυχικής υγείας και την ανάγκη περαιτέρω βελτιστοποίησης των global μοντέλων, ετσι ώστε να βελτιωθεί η ικανότητα γενίκευσης σε ποικίλους ασθενείς.

### Joint LSTMAE-CAE Μοντέλο

Στο Σχήμα 0.0.29 παρουσιάζονται τα ROC-AUC scores του joint LSTMAE-CAE μοντέλου στα εξατομικευμένα πειράματα, για κάθε σύνολο δεδομένων. Όπως και στα πειράματα του CAE-CAE μοντέλου, η ανισορροπία μεταξύ της απόδοσης των ηχητικών και βιομετρικών μοντέλων είναι εμφανής. Ωστόσο, στα μεγαλύτερα σύνολα δεδομένων, το LSTMAE-CAE παρουσίασε βελτιώσεις στην ανίχνευση καταστάσεων υποτροπής συγκριτικά με τα unimodal μοντέλα, με τα σύνολα δεδομένων των 3 ημερών να παρουσιάζει την υψηλότερα συνολική βελτίωση. Επιπλέον, σε όλα τα σύνολα δεδομένων τα ROC-AUC scores των ηχητικών και βιομετρικών κλάδων είναι υψηλότερες από αυτές των αντίστοιχων unimodal μοντέλων. Συνεπώς, τα αποτελέσματα του LSTMAE-CAE μοντέλου στα εξατομικευμένα πειράματα ενισχύει την αξία της multimodal προσέγγισης για την ανίχνευση υποτροπών, ειδικά για μεμονωμένους ασθενείς.

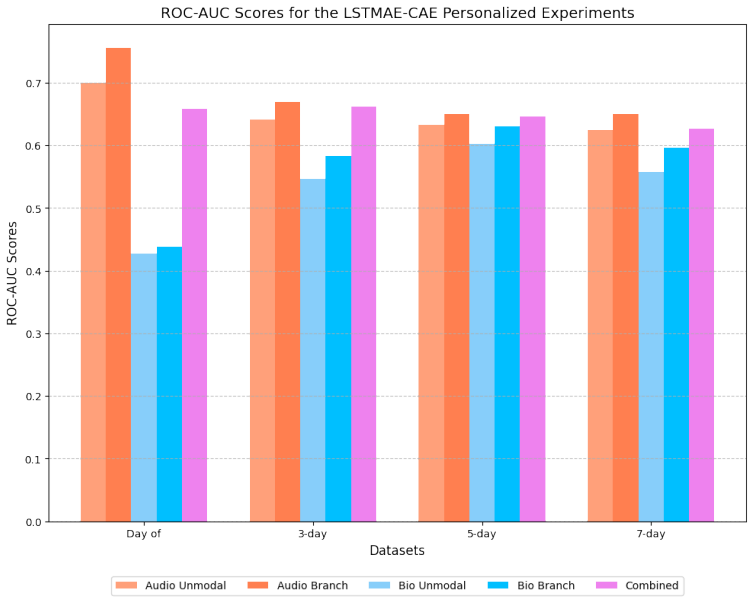



Figure 0.0.29: Μέσα ROC-AUC scores για το joint LSTMAE-CAE μοντέλο στα εξατομικευμένα πειράματα.

Το Σχήμα 0.0.30 παρουσιάζει τα ROC-AUC scores του joint CAE-CAE μοντέλου στα global πειράματα με per-patient κανονικοποίηση, ενώ το Σχήμα 0.0.31 παρουσιάζει τα αντίστοιχα αποτελέσματα με global κανονικοποίηση.

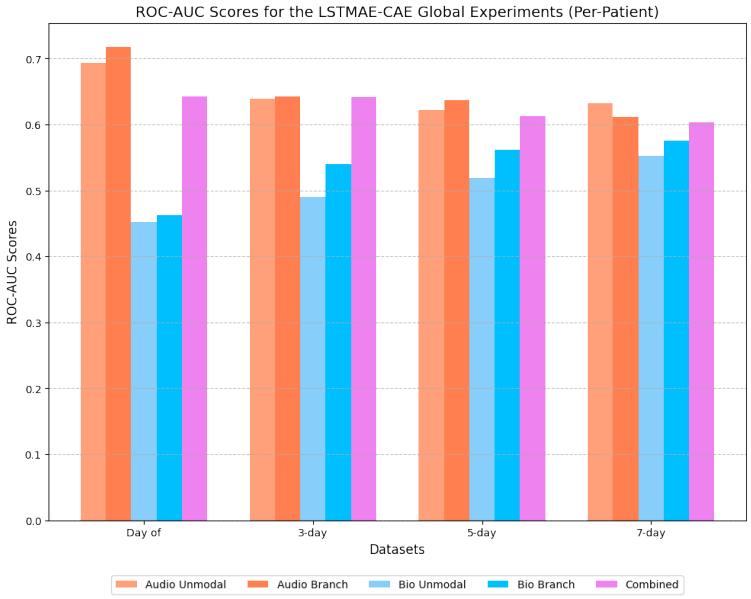



Figure 0.0.30: Μέσα ROC-AUC scores για το joint LSTMAE-CAE μοντέλο στα global πειράματα με per-patient κανονικοποίηση.

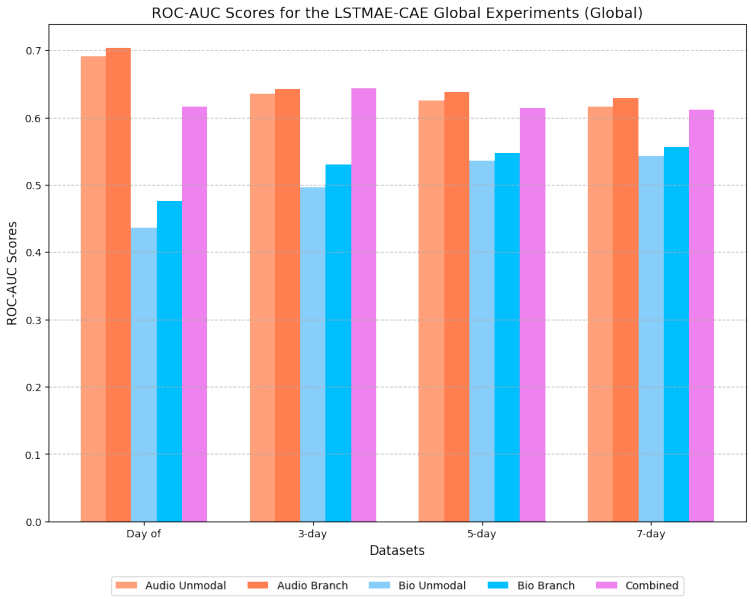



Figure 0.0.31: Μέσα ROC-AUC scores για το joint LSTMAE-CAE μοντέλο στα global πειράματα με global κανονικοποίηση.

Από τα αποτελέσματα των global πειραμάτων με per-patient κανονικοποίηση παρατηρούμε, όπως και στο CAE-CAE μοντέλο, μικρότερες βελτιώσεις των κλάδων σε σχέση με τα unimodal μοντέλα. Η συνολική απόδοση του μοντέλου είναι μεικτή σε σύγκριση με τα εξατομικευμένα πειράματα και συγκρίσιμη κυριώς με το βιομετρικό unimodal μοντέλο. Στη global κανονικοποίηση, το joint μοντέλο παρουσίασε επίσης λιγότερο έντονες βελτιώσεις σε σχέση με τα unimodal μοντέλα. Αν και οι τιμές ROC-AUC ήταν ελαφρώς υψηλότερες από τα αντίστοιχα unimodal μοντέλα, η συνδυασμένη επίδοση του μοντέλου έδειξε βελτιωμένα αποτελέσματα μόνο στο σύνολο δεδομένων των 3 ημερών. Επομένως, ο συνδυασμός των δύο τύπων δεδομένων είναι χρήσιμος για την ανίχνευση υποτροπών σε όλους τους ασθενείς, όμως τα εξατομικευμένα μοντέλα παραμένουν ανώτερα σε απόδοση.

Συνολικά, η απόδοση των δύο μοντέλων φαίνεται να είναι παρόμοια, με μικρές διαφορές στα εξατομικευμένα πειράματα, όπου το CAE-CAE μοντέλο παρουσίασε καλύτερη απόδοση, ενώ το LSTMAE-CAE μοντέλο παρουσίασε καλύτερη απόδοση στα global πειράματα. Τα ευρήματα και των δύο υπογραμμίζουν ο συνδυασμός βιομετρικών και ηχητικών δεδομένων είναι πράγματι αποτελεσματικός, και είναι ικανός να προσφέρει έγκαιρη ανίχνευση υποτροπών σε ασθενείς με ψυχικές διαταραχές. Ταυτόχρονα, καθίσταται σαφές ότι η global προσέγγιση απαιτεί περαιτέρω βελτιστοποίηση, ώστε να επιτευχθεί η αποτελεσματική γενίκευση των μοντέλων σε δεδομένα πολλών ασθενών.

### Αξιολόγηση της Συνεισφοράς του Κάθε Κλάδου

Τέλος, για την αξιολόγηση της συνεισφοράς του κάθε κλάδου στο συνολικό joint μοντέλο αποφασίσαμε να "απενεργοποιήσουμε" τον έναν από τους δύο κλάδους στη διαδικασία της πρόβλεψης. Η απενεργοποίηση επιτεύχθηκε μηδενίζοντας την είσοδό του κλάδου, δηλαδή θέτοντας όλα τα δεδομένα εισόδου του συγκεκριμένου κλάδου σε μηδενικές τιμές. Αυτό είχε ως αποτέλεσμα να μην παρέχεται καμία πληροφορία του κλάδου αυτού στο joint μοντέλο, αναγκάζοντάς το να βασιστεί αποκλειστικά στον ενεργό κλάδο για την ανίχνευση ανωμαλιών. Οι Πίνακες 2-5 παρουσιάζουν τα ROC-AUC αποτελέσματα των πειραμάτων απενεργοποίησης του κάθε κλάδου στα εξατομικευμένα και global μοντέλα.

| **All Patients** | **ROC-AUC** | | |
|---|---|---|---|
| | **Audio Unimodal** | **Audio Branch (Bio Disabled)** | **Audio Branch** |
| **Mean** | 0.598±0.064 | 0.620±0.108 | 0.643±0.079 |

Table 2: Σύγκριση των ROC-AUC scores του εξατομικευμένου ηχητικού μοντέλου, του ηχητικού κλάδου του joint μοντέλου με τον βιομετρικό κλάδο απενεργοποιημένο και ενεργοποιημένο αντίστοιχα.

| **All Patients** | **ROC-AUC** | | |
|---|---|---|---|
| | **Bio Unimodal** | **Bio Branch (Audio Disabled)** | **Bio Branch** |
| **Mean** | 0.557±0.136 | 0.602±0.108 | 0.629±0.120 |

Table 3: Σύγκριση των ROC-AUC scores του εξατομικευμένου βιομετρικού μοντέλου, του βιομετρικού κλάδου του joint μοντέλου με τον ηχητικό κλάδο απενεργοποιημένο και ενεργοποιημένο αντίστοιχα.

| **Norm.** | **ROC-AUC** | | |
|---|---|---|---|
| | **Audio Unimodal** | **Audio Branch (Bio Disabled)** | **Audio Branch** |
| **Per-Patient** | 0.603±0.059 | 0.607±0.093 | 0.614±0.048 |
| **Global** | 0.600±0.060 | 0.602±0.056 | 0.607±0.053 |

Table 4: Σύγκριση των ROC-AUC scores του global ηχητικού μοντέλου, του ηχητικού κλάδου του joint μοντέλου με τον βιομετρικό κλάδο απενεργοποιημένο και ενεργοποιημένο αντίστοιχα.

| **Norm.** | **ROC-AUC** | | |
|---|---|---|---|
| | **Bio Unimodal** | **Bio Branch (Audio Disabled)** | **Bio Branch** |
| **Per-Patient** | 0.553±0.029 | 0.555±0.040 | 0.555±0.032 |
| **Global** | 0.543±0.050 | 0.550±0.057 | 0.572±0.039 |

Table 5: Σύγκριση των ROC-AUC scores του global βιομετρικού μοντέλου, του βιομετρικού κλάδου του joint μοντέλου με τον ηχητικό κλάδο απενεργοποιημένο και ενεργοποιημένο αντίστοιχα.

Τα πειράματα πραγματοποιήθηκαν για το CAE-CAE μοντέλο στο σύνολο δεδομένων των 7 ημερών και τα αποτελέσματα, στα εξατομικευμένα και στα global πειράματα, έδειξαν, όπως φαίνεται από τους Πίνακες 2-5, ότι τα ROC-AUC scores του ηχητικού κλάδου, όταν ο βιομετρικος ήταν απενεργοποιημένος, ήταν υψηλότερα από το αντίστοιχο unimodal μοντέλο αλλά μειωμένα σε σχέση με αυτά του ηχητικού κλάδου όταν ο βιομετρικός είναι ενεργοποιημένος. Η ίδια συμπεριφορά παρατηρείται αντίστοιχα για τον βιομετρικό κλάδο. Η ανάλυση αυτή επιβεβαιώνει ότι το joint μοντέλο πράγματι αξιοποιεί αποτελεσματικά τα δεδομένα και από τις δύο πηγές, οδηγώντας σε ακριβέστερη ανίχνευση υποτροπών.

Συνοψίζοντας, όλα τα παραπάνω πειράματα ανέδειξαν τη σημασία του συνδυασμού ηχητικών και βιομετρικών δεδομένων στην ανίχνευση υποτροπών, με τα joint μοντέλα να ξεπερνούν σταθερά τις επιδόσεις των unimodal προσεγγίσεων. Η εξατομικευμένη προσέγγιση επιβεβαιώθηκε ως η πιο αποτελεσματική, καθώς τα μοντέλα που εκπαιδεύτηκαν σε δεδομένα συγκεκριμένων ασθενών απέδωσαν καλύτερα από αυτά που εκπαιδεύτηκαν σε όλους τους ασθενείς, αναδεικνύοντας τη σημασία της εξατομικευμένης παρακολούθησης της ψυχικής υγείας.

# Συμπεράσματα και Μελλοντικές Επεκτάσεις

Η παρούσα διπλωματική εστιάζει στην ανάλυση ηχητικών σημάτων και βιομετρικών δεδομένων για τη βελτίωση της πρόβλεψης υποτροπών σε ασθενείς με διπολική διαταραχή και σχιζοφρένεια. Αρχικά, επεκτείναμε την ηχητική βάση δεδομένων του e-Prevention με πρόσθετα δεδομένα ασθενών και περισσότερες συνεντεύξεις που αντιστοιχούν σε περιόδους υποτροπής. Η αξιολόγηση των ήδη υλοποιημένων μοντέλων CAE και CVAE του e-Prevention σε αυτό το νέο σύνολο δεδομένων επιβεβαίωσε ότι η αύξηση του αριθμού των δεδομένων οδήγησε σε βελτίωση της απόδοσης των μοντέλων, με το CAE να παρουσιάζει τη μεγαλύτερη βελτίωση. Παράλληλα, το CVAE επηρεάστηκε θετικά από την αύξηση των δεδομένων, ωστόσο λίγοτερο σε σχέση με το CAE, γεγονός που πιθανώς να οφείλεται στη μεγαλύτερη ευαισθησία του CVAE στη μεταβλητότητα των δεδομένων. Για τη βελτίωση της ανίχνευσης ανωμαλιών, προτείναμε μοντελοποίηση των ακολουθιακών εξαρτήσεων των δεδομένων υλοποιώντας LSTM autoencoders (LSTMAE και LSTMVAE). Τα αποτελέσματα έδειξαν ότι το LSTMAE υπερείχε του CAE σε όλα τα πειραματικά σενάρια, ενώ το LSTMVAE είχε παρόμοια απόδοση με το CVAE, ωστόσο το CVAE παραμένει η πιο αξιόπιστη προσέγγιση για variational autoencoders. Η σημαντικότερη συμβολή της έρευνας ήταν ο συνδυασμός βιομετρικών και ηχητικών δεδομένω για τη βελτίωση της πρόβλεψης υποτροπής. Αναπτύχθηκαν joint autoencoder μοντέλα, που ενσωματώνουν τα CAE και LSTMAE μοντέλα για τα ηχητικά δεδομένα και ένα CAE μοντέλο για τα βιομετρικά δεδομένα. Τα μοντέλα αξιολογήθηκαν σε σύνολα δεδομένων που περιελάμβαναν βιομετρικά δεδομένα από την ίδια ημέρα των συνεντεύξεων, καθώς και από χρονικά παράθυρα γύρω από την ημέρα των συνεντεύξεων. Τα αποτελέσματα έδειξαν ότι τα joint μοντέλα ξεπέρασαν τις επιδόσεις των ηχητικών και βιομετρικών μοντέλων ξεχωριστά, με τα εξατομικευμένα μοντέλα να επιτυγχάνουν τη μεγαλύτερη ακρίβεια. Συγκρίνοντας τα μοντέλα CAE-CAE και LSTMAE-CAE, διαπιστώθηκε ότι το CAE-CAE ήταν πιο αποδοτικό στα εξατομικευμένα πειράματα, ενώ το LSTMAE-CAE παρουσίασε ένα μικρό πλεονέκτημα στα global πειράματα, γεγονός που υποδηλώνει ότι η επιλογή του κατάλληλου μοντέλου εξαρτάται από τις απαιτήσεις του εκάστοτε προβλήματος. Τέλος, τα πειράματα απενεργοποίησης των κλάδων επιβεβαίωσαν ότι τα joint μοντέλα αξιοποιούν αποτελεσματικά και τις δύο πηγές δεδομένων, βελτιώνοντας την ανίχνευση υποτροπών.

Οι μελλοντικές επεκτάσεις της έρευνας θα μπορούσαν να επικεντρωθουν στα εξής:

- Περαιτέρω διεύρυνση του συνόλου δεδομένων με την ένταξη περισσότερων ασθενών και επιπλέον περιπτώσεων υποτροπής, με σκοπό την ενίσχυση της γενίκευσης των μοντέλων και την αύξηση της αξιοπιστίας των αποτελεσμάτων.
- Για τα variational autoencoder μοντέλα, μπορούν να διερευνηθούν πρόσθετες τεχνικές κανονικοποίησης, όπως η προσαρμοστική κανονικοποίηση, η χρήση βαρών προσαρμοσμένα σε κάθε ασθενή ή περαιτέρω προσαρμογή (fine-tuning) των μοντέλων για συγκεκριμένους ασθενείς όπου η απόδοση δεν βελτιώθηκε επαρκώς.
- Εισαγωγή τεχνικών χρονικής μοντελοποίησης, όπως οι μηχανισμοί προσοχής (attention mechanisms), οι οποίες θα μπορούσαν να βελτιώσουν περαιτέρω την απόδοση των LSTM autoencoder μοντέλων.
- Διερεύνηση της αντιστοίχισης των ηχητικών και βιομετρικών δεδομένων και τεχνικών επαύξησης των δεδομένων (data augmentation) για την αντιμετώπιση της διαφοράς στην απόδοση των δύο κλάδων.

# Chapter 1

# Introduction

Mental health disorders are complex and multifaceted conditions that pose significant challenges to individuals, families, and healthcare providers. These disorders are often characterized by reoccuring relapses, which can have profound impacts on patients' well-being and functional capacity. Relapses are not only distressing but also costly, often requiring hospitalization and leading to disruptions in personal and professional life. Addressing these challenges necessitates approaches that go beyond traditional methods of care and prioritize early detection and prevention. Advances in technology have created new opportunities to address the intricacies of mental health management. By harnessing diverse types of information, it is possible to gain deeper insights into the subtle changes that often precede relapse. Patterns in speech, physical activity, and physiological signals, for example, offer valuable clues about an individual's mental state. These data-driven approaches have the potential to fill the gaps left by traditional methods, which rely heavily on clinical assessments and self-reported symptoms.

This thesis seeks to address these challenges by exploring advanced methods for relapse prediction in patients with bipolar disorder and schizophrenia spectrum disorders (SSD). By evaluating and improving models that detect behavioral and physiological changes, it aims to provide valuable insights that could support earlier and more effective interventions. Through this research, the thesis contributes to the broader goal of developing tools and strategies that improve mental health outcomes and facilitate the lives of individuals living with these complex conditions.

## 1.1 Relapse in Bipolar Disorder and Schizophrenia

Mental health disorders, particularly bipolar disorder and schizophrenia spectrum disorders (SSD) represent a significant global health challenge due to their prevalence, chronicity, and profound impact on patients' quality of life. Schizophrenia affects approximately 24 million people worldwide, about 0.32% of the global population [Wor22], while bipolar disorder impacts an 2.4% of the global population [Zho+24]. These conditions are associated with high rates of reoccuring relapse, which often necessitate hospitalization, disrupt social and occupational functioning, and increase the risk of mortality. For instance, studies indicate that up to 52% of individuals with schizophrenia who have already been hospitalized experience a relapse within the first year after discharge [BMV12]. Similarly, 25% of individuals with bipolar disorder experience relapses severe enough to require hospitalization or result in acute episodes, with 40% experiencing multiple relapses within a five-year period [Het+23].

Bipolar disorder is characterized by recurrent episodes of depression and mania. Depressive episodes are marked by feelings of deep sadness, fatigue, and loss of interest in daily activities, often accompanied by suicidal thoughts. In contrast, manic episodes include hyperactivity, impulsive decision-making and a reduced need for sleep [Nie+23]. These behaviors can lead to risky actions and require immediate medical intervention. Relapse in either phase is common and can significantly impact an individual's daily functioning, often leading to long-term disability. [JA03; Gra+16].

Schizophrenia is a chronic psychotic disorder with symptoms including hallucinations, delusions, cognitive deficits, emotional dysregulation and social withdrawal [VK09]. Psychotic relapses, which frequently occur in the course of the illness, can lead to prolonged hospital stays, further cognitive and functional decline and diminished opportunities for recovery [LDG13]. Furthermore, schizophrenia is associated with a significantly increased risk of premature mortality, driven primarily by suicide and comorbid cardiovascular conditions [LNM14].

During manic and psychotic episodes, individuals with bipolar disorder often exhibit accelerated speech, along with increased variability in pitch and volume. In contrast, depressive episodes are characterized by slower, monotonic speech that often lacks emotional expression. Similarly, in schizophrenia, patients during relapse tend to demonstrate reduced prosody, lower vocal intensity, and an increased number of pauses. These vocal characteristics can serve as significant indicators for evaluating the state of these disorders, as changes in speech patterns often reflect the emotional state of the patient [LBG20; Fau+16].

In addition to vocal features, bipolar disorder and schizophrenia affect various biometric characteristics of patients. Levels of physical activity and movement fluctuate significantly during periods of relapse. Manic episodes in bipolar disorder are often marked by increased agitation and hyperactivity, whereas depressive episodes frequently lead to reduced activity and lethargy [Max+16]. Similarly, in schizophrenia, psychotic

episodes are associated with hyperactivity, often manifesting as irregular, repetitive, and rapid physical movements [WM17]. Another critical aspect of both disorders is the dysfunction of the autonomic nervous system (ANS). Heart Rate Variability (HRV), a measure of ANS function, can also be used as an indicator of fluctuations in the mental state of patients. Individuals with bipolar disorder and schizophrenia often exhibit reduced HRV during depressive episodes or heightened emotional distress, indicating lower adaptability to stressors. Conversely, higher HRV is typically associated with relaxation, reduced stress, and emotional balance, which correspond to stable mental health states [Hen+10].

Traditional methods for assessing bipolar disorder and schizophrenia primarily rely on clinical interviews, self-reported questionnaires, and standardized diagnostic tools such as the DSM-5 criteria [Ame13], the Positive and Negative Syndrome Scale (PANSS) [KFO87a] for schizophrenia, and the Young Mania Rating Scale (YMRS) [You+78] for bipolar disorder. These evaluations are inherently subjective, relying heavily on patient recall and clinician interpretation. This makes it difficult to capture subtle, continuous changes that may precede relapse, leading to missed opportunities for early intervention.

Consequently, continuous and accurate monitoring of patients with bipolar disorder and schizophrenia is vital. Existing approaches often fall short in detecting relapses in a timely manner and managing the fluctuations in mental state. The complexity of these disorders, combined with their significant impact on multiple aspects of a patient's life, underscores the need for complementary methods that utilize vocal and biometric markers to facilitate early detection and relapse prevention.

## 1.2 Digital Phenotyping

Digital phenotyping is an emerging approach that involves the use of passive data from digital devices, such as smartphones and wearable sensors, to capture and analyze behavioral and physiological patterns (Fig. 1.2.1). It enables continuous, objective and unobtrusive monitoring of individuals' daily lives, providing insights into their mental and physical health. Unlike traditional relapse evaluations, digital phenotyping allows for the detection of subtle, evolving patterns in behavior or physiology that may precede clinical symptoms of relapse. These behavioral markers can be particularly valuable for mental disorders such as bipolar disorder and schizophrenia, where early intervention can significantly improve outcomes [OR16; MZS17].

Digital phenotyping has found applications across a variety of health domains. In cardiovascular health, wearable devices monitor heart rate variability (HRV) to detect arrhythmias, assess stress levels, or predict risks of heart disease [SG17a]. Similarly, continuous glucose monitoring systems paired with smartphones enable real-time tracking of blood sugar levels for diabetes management [Cap+19]. Neurological conditions such as Parkinson's disease are monitored through wearable devices that track tremors and gait abnormalities, while epilepsy management has been enhanced by tools that detect seizure activity [JMM18]. Other notable applications include maternal health, where wearables track vital signs to detect pregnancy complications like preeclampsia [DAS22].

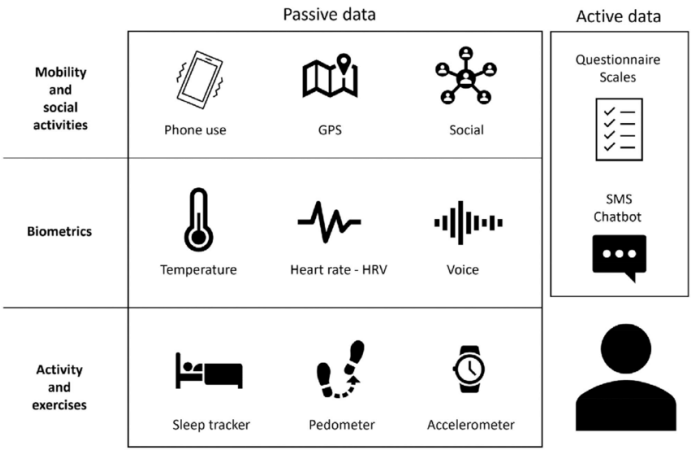



Figure 1.2.1: Overview of digital phenotyping [Mou+21].

In the context of mental health, digital phenotyping has emerged as a powerful tool for understanding and managing conditions such as bipolar disorder, schizophrenia, depression, and anxiety. Continuous monitoring of speech, physical activity, and physiological metrics has shown promise in identifying behavioral and emotional changes that precede clinical symptoms of relapse or deterioration. By providing real-time, passive data, digital phenotyping enhances the ability to detect early warning signs of relapse, ultimately enabling more personalized and proactive mental health care [Pan+18; Bar+18].

## 1.3 Goals and Contributions

This study builds upon the e-Prevention project [Zla+22], which leverages digital phenotyping for early relapse detection in patients with bipolar disorder and schizophrenia spectrum disorders (SSD). The project developed an integrated system for long-term monitoring using wearables and video recordings, applying machine learning techniques to detect behavioral and physiological markers of relapse. However, its primary focus was on analyzing biometric and speech data separately, without fully utilizing the potential of multimodal fusion for relapse prediction. The primary goal of this research is to further evaluate and extend the models developed in the e-Prevention project by incorporating advancements in data processing and machine learning techniques. Additionally, the study seeks to explore the integration of audio and biometric data within a multimodal framework to enhance relapse prediction performance.

More specifically, this research extends the field of digital phenotyping and mental health monitoring through the following contributions:

- Expansion of the e-Prevention audio database to create a more diverse dataset, in order to improve model evaluation and generalization.
- Reassessment of the CAE and CVAE models developed during the e-Prevention project on the expanded dataset, ensuring their performance on detecting relapse across existing and additional patients.
- Development and evaluation of LSTM-based autoencoders for relapse detection from speech data, comparing their effectiveness to convolutional models and examining how their ability to capture temporal dependencies in speech signals can improve prediction accuracy.
- Development of joint autoencoder models for feature-level fusion of audio and biometric data, enhancing relapse prediction accuracy and demonstrating the benefits of multimodal approaches in mental health monitoring.

## 1.4 Thesis Outline

This thesis is organized into the following chapters:

- **Chapter 1: Introduction**

  This chapter provides an overview of the research context, highlighting the challenges of relapse detection in bipolar disorder and schizophrenia. It introduces the e-Prevention project, the objectives and the contributions of this thesis.

- **Chapter 2: Theoretical Background**

  This chapter reviews the key methodologies and machine learning concepts relevant to this research. It covers audio and biometric signal representations, neural network architectures, autoencoder approaches and the concept of anomaly detection.

- **Chapter 3: Literature Review**

  This chapter analyzes current state-of-the art approaches to relapse detection in mental health using behavioral and physiological data. It covers studies on digital phenotyping, multimodal fusion, and techniques for anomaly detection in audio data. Finally, it presents on the e-Prevention project, its methodologies, and the models developed for audio and biometric data.

- **Chapter 4: Data and Preprocessing**

  This chapter describes the datasets used in the study, including the expansion of the e-Prevention audio database and the processing of biometric data. It details data preprocessing techniques, feature extraction, and the alignment of audio and biometric data using temporal windows.

- **Chapter 5: Audio Experiments**

  This chapter focuses on the methodologies, results, and discussions related to audio-only experiments. It presents the architectures used for detecting anomalies in speech data, including typical and variational autoencoders consisting of convolutional and LSTM-based blocks, and evaluates their performance on the expanded audio dataset.

- **Chapter 6: Multimodal Experiments**

  This chapter presents the methodologies, results, and discussions for experiments that integrate biometric and audio data. It introduces the autoencoder model used for the biometric data and the joint autoencoder models developed for multimodal relapse detection. Additionally, it examines the contribution of each modality through ablation experiments.

- **Chapter 7: Conclusion and Future Work**

  The final chapter summarizes the key contributions of the thesis and discusses potential directions for future research that can be conducted based on the presented work.

# Chapter 2

# Theoretical Background

## 2.1 Audio Signal Representations and Features

Audio signals are a critical source of information for analyzing speech patterns and identifying behavioral or emotional changes. To utilize these signals effectively in machine learning tasks, raw data must be transformed into representations that highlight key features.

### 2.1.1 Time Domain-Representations

In the time domain, an audio signal is represented as a continuous waveform $s(t)$, where $t$ denotes time in seconds, and the amplitude of the signal varies with time.

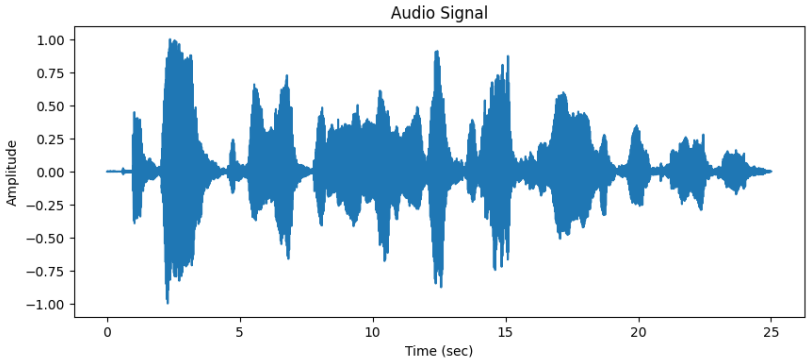



Figure 2.1.1: Audio signal waveform.

To process these signals digitally, they must be sampled, converting the continuous signal into a discrete-time signal:

$$s[n] = s(nT_s), \quad T_s = \frac{1}{f_s} \tag{2.1.1}$$

Where:

- $f_s$ is the sampling frequency in Hz.
- $T_s$ is the sampling period in seconds.
- $n$ is the sample index.

The sampling process is governed by the sampling frequency $f_s$ which determines the number of samples taken per second. According to the Nyquist theorem, the sampling rate must be at least twice the highest frequency present in the signal to avoid aliasing (i.e., deformation of the analog signal).

### 2.1.2 Frequency-Domain Representations

While the time-domain representation of a signal captures its amplitude variation over time, it does not provide information directly about the frequency content. The Fourier Transform is used to convert the time-domain signal into the frequency domain:

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft} dt \tag{2.1.2}$$

Where:

- $S(f)$ is the Fourier Transform of the signal $s(t)$, representing the signal's frequency content.
- $f$ is the frequency in Hz.
- $e^{-j2\pi ft}$ represents oscillations at frequency $f$.

The Discrete Fourier Transform (DFT) analyzes the frequency components of a discrete-time signal $s[n]$ over $N$ samples:

$$S[k] = \sum_{n=0}^{N-1} s[n] e^{-j2\pi kn/N} \tag{2.1.3}$$

Where:

- $k$ corresponds to specific frequency components present in the discrete-time signal $s[n]$.
- $S[k]$ represents the frequency-domain signal at frequency index $k$.
- $N$ is the number of samples.

For practical use with finite and non-stationary signals, the Short-Time Fourier Transform (STFT) is applied. The STFT analyzes short segments of the signal using a sliding window:

$$STFT\{s[n]\}(m,k) = S(m,k) = \sum_{n=-\infty}^{\infty} s[n] w[n-m] e^{-j2\pi nk/N} \tag{2.1.4}$$

Where:

- $S(m,k)$ is the STFT of the signal.
- $w[n]$ is a window function (e.g., Hamming or Hann).
- $m$ is the frame index, indicating the position of the window.
- $k$ is the frequency index.

The STFT produces a time-frequency representation, capturing how the frequency content of the signal evolves over time.

### 2.1.3 Spectral Representations

The spectrogram is a widely used time-frequency representation that visualizes how the power of a signal varies across different frequencies over time. Mathematically, the spectrogram is computed as follows:

$$\text{spectrogram}(m,\omega) = |STFT\{s[n]\}(m,\omega)|^2 \tag{2.1.5}$$

Where:

- $|STFT\{s[n]\}(m,\omega)|^2$ is the power spectral density (PSD) of the signal.

The spectrogram reveals how energy is distributed over time and frequency but still represents frequencies on a linear scale, unlike the human auditory system, which perceives pitch logarithmically.

### 2.1.4 Mel-Spectrograms

The Mel-spectrogram adjusts the spectrogram's frequency representation to match human auditory perception. Humans are more sensitive to lower frequencies and perceive higher frequencies on a logarithmic scale. This transformation uses the Mel-scale, which maps linear frequencies $f$ to perceptually meaningful Mel-frequencies $m(f)$ as follows:

$$m(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.1.6}$$

To compute a mel-spectrogram:

1. The STFT is computed for each frame to obtain the magnitude spectrum:

$$|S(\tau, \omega)|^2 \tag{2.1.7}$$

2. A set of triangular filters is applied to the frequency bands of the spectrogram to map them onto the Mel-scale:

$$M(m, k) = \sum_{\omega} |S(m, \omega)|^2 H_k(\omega) \tag{2.1.8}$$

Where:

- $M(m, k)$ is the mel-spectrogram.
- $H_k(\omega)$ is the triangular filter centered at frequency band $k$.

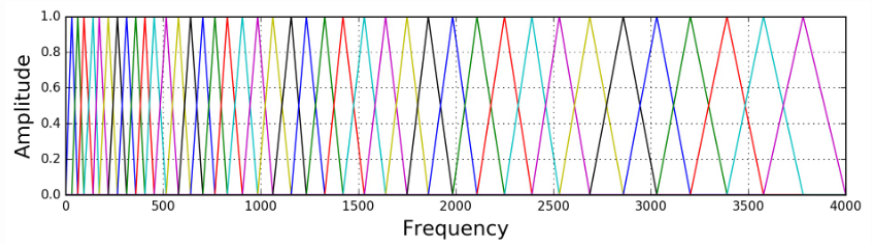



Figure 2.1.2: Filter bank on a Mel-scale [Fay16].

3. The logarithm of the Mel-scaled spectrogram is computed to compress its dynamic range:

$$\text{Log-Mel-Spectrogram}(m, k) = \log(M(m, k) + \epsilon) \tag{2.1.9}$$

Where:

- $\epsilon$ is a small constant to avoid logarithmic instability for near-zero values.

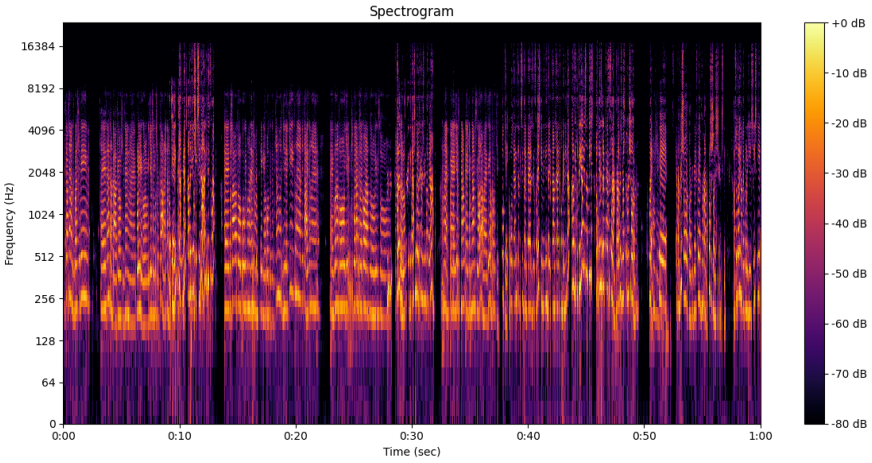



Figure 2.1.3: Mel-spectrogram of audio signal.

## 2.2 Biometric Signal Representations and Features

Biometric signals derived from sensors provide crucial insights into physiological states and behavioral patterns. These signals are represented as time-series data and processed into time-domain, frequency-domain, and non-linear features for analysis. In this section, we present an overview of common biometric signal representations that have been used in related studies [Zla+22; Ret+20; Mag+20; Fil+20].

### 2.2.1 Time-Domain Features

Time-domain features are computed directly from the raw signal and describe statistical characteristics of the data.

#### Short-Time Energy (STE)

Short-Time Energy (STE) measures the localized energy of a signal within overlapping windows, capturing temporal variations in intensity. It is a fundamental technique in short-time analysis, used to examine non-stationary signals by segmenting them into overlapping frames. This allows for tracking changes in signal intensity, variability, and periodicity over time.

STE is computed as the sum of squared signal samples within a window of length $N$:

$$\text{STE} = \sum_{n=0}^{N-1} x^2[n] \tag{2.2.1}$$

#### Mean and Variability

Statistical features, such as the mean and variability, are fundamental descriptors of biometric signals. These features summarize the central tendency and spread of the data, offering insights into the overall behavior and fluctuations within the signal.

The mean is computed as the average of the signal samples:

$$\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \tag{2.2.2}$$

The variability is computed as the standard deviation of the signal samples:

$$\text{Var(x)} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2} \tag{2.2.3}$$

For biometric signals representing time intervals (e.g. NN intervals in Heart Rate Variability (HRV)), the mean NN interval and standard deviation of NN intervals (SDNN) are commonly extracted features [SG17b].

### 2.2.2 Frequency-Domain Features

Frequency-domain analysis decomposes biometric signals into their spectral components to understand how power is distributed across various frequency bands. The Lomb-Scargle Periodogram [Sca82] is a particularly useful method estimating the PSD and detecting periodic patterns in HRV and other physiological signals.

#### Lomb-Scargle Periodogram

The periodogram estimates the PSD of a signal by squaring the magnitude of its DFT. The Lomb-Scargle Periodogram is a powerful extension of the periodogram for analyzing unevenly spaced time-series data, which is common in physiological recordings like HRV. It is computed as follows:

$$P_{LS}(\Omega) = \frac{1}{2} \left\{ \frac{\left[\sum_{n=0}^{N-1} x[n] \cos(\Omega(t_n - \tau))\right]^2}{\sum_{n=0}^{N-1} \cos^2(\Omega(t_n - \tau))} + \frac{\left[\sum_{n=0}^{N-1} x[n] \sin(\Omega(t_n - \tau))\right]^2}{\sum_{n=0}^{N-1} \sin^2(\Omega(t_n - \tau))} \right\}, \tag{2.2.4}$$

Where:

- $\tau$ is given by:

$$\tau = \frac{1}{2\Omega} \tan^{-1} \left( \frac{\sum_{n=0}^{N-1} \sin(2\Omega t_n)}{\sum_{n=0}^{N-1} \cos(2\Omega t_n)} \right). \tag{2.2.5}$$

- $\Omega$ is the angular frequency in radians per second (rad/s).
- $t_n$ is the time at which the signal is sampled.

#### HRV Frequency Bands

In the context of HRV, the PSD is analyzed within specific frequency bands [SG17a], which are defined as follows:

- **Ultra-Low Frequency (ULF)**: $\leq 0.003$ Hz.

  Reflects slow physiological processes such as circadian rhythms.

- **Very Low Frequency (VLF)**: $0.003 - 0.04$ Hz.

  Associated with regulatory mechanisms, including thermoregulation and the renin-angiotensin system. Low VLF power is linked to increased mortality risk and inflammation.

- **Low Frequency (LF)**: $0.04 - 0.15$ Hz.

  Represents both sympathetic and parasympathetic nervous system activity under resting conditions. Influenced by baroreceptor activity and breathing rhythms.

- **High Frequency (HF)**: $0.15 - 0.4$ Hz.

  Reflects parasympathetic activity, often related to respiratory sinus arrhythmia (RSA). Corresponds to heart rate changes during inhalation and exhalation.

The LF/HF Ratio, a commonly used metric, provides an estimate of the balance between sympathetic and parasympathetic nervous system activities [SG17a].

### 2.2.3 Non-Linear Features

Non-linear features capture the complexity and irregularity of signals, providing deeper insights into the underlying dynamics that are not easily represented in the time or frequency domains.

#### Poincaré Plot Analysis

Poincaré analysis is a widely used non-linear method for visualizing and quantifying variability in biometric signals, particularly in HRV. The Poincaré plot is a scatterplot of consecutive intervals, where each point represents a pair of successive intervals $(NN_i, NN_{i+1})$.

The geometry of the Poincaré plot is often quantified using two standard descriptors: SD1 and SD2. These metrics are derived by fitting an ellipse to the scatterplot, where SD1 and SD2 correspond to the lengths of the minor and major semi-axes of the fitted ellipse, respectively [BPK01; PG07].

- **SD1 (Short-Term Variability)**: Captures the spread of points perpendicular to the line of identity $(y = x)$, reflecting rapid, short-term fluctuations primarily associated with parasympathetic activity.

$$SD1 = \frac{\sqrt{2}}{2} Var(NN_i - NN_{i+1}) \tag{2.2.6}$$

- **SD2 (Long-Term Variability)**: Captures the spread of points along the line of identity, reflecting slower, long-term fluctuations associated with both sympathetic and parasympathetic activity.

$$SD2 = \sqrt{2Var(NN_i)^2 - SD1^2} \tag{2.2.7}$$

- The $SD1/SD2$ Ratio provides insights into the balance between short-term and long-term variability, with lower values indicating a shift towards sympathetic dominance [BPK01].

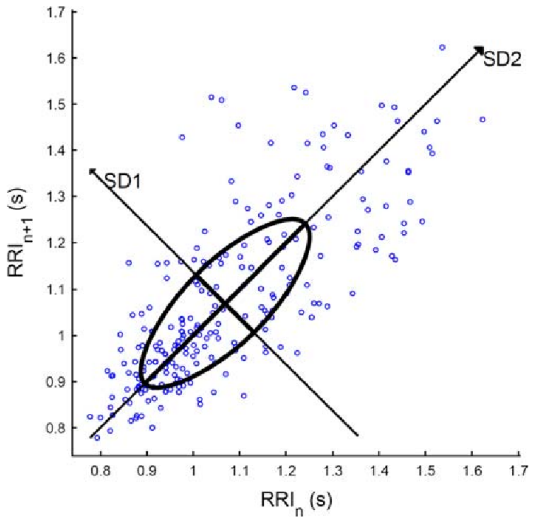



Figure 2.2.1: Example of Poincaré plot of HRV signal [Cho+09].

### Entropy

Entropy quantifies the unpredictability or irregularity of a time-series signal, making it a robust tool for assessing complexity. In biometric signals, entropy measures are utilized to capture the irregularity of physiological processes.

- **Approximate Entropy (ApEn)**: Quantifies the regularity of patterns in a time series; higher values indicate greater complexity.
- **Sample Entropy (SampEn)**: An improved version of ApEn, less dependent on data length and excluding self-matching patterns.
- **Shannon Entropy**: Measures the uncertainty or information content of a signal, reflecting the diversity of signal values.
- **Permutation Entropy**: Evaluates the complexity of a signal by analyzing the ordinal patterns of its values, robust to noise.
- **Multiscale Entropy (MSE)**: Captures the complexity of a signal across multiple time scales, providing insights into the dynamics of physiological systems.

### Fractal Dimension

Fractals are complex geometric shapes that exhibit self-similarity across different scales, meaning their structure looks similar regardless of the level of magnification. Defined mathematically, a fractal dimension is a measure that describes how the detail in a pattern changes with scale, providing a non-integer value that characterizes the object's complexity [Man82]. In biometric signals, the fractal dimension is used to quantify the irregularity and self-similarity of physiological processes.

Common fractal dimension measures for time-series signals include:

- **Higuchi Fractal Dimension (HFD)**: Estimates the fractal dimension directly from time-series data by calculating the length of the curve at various scales, providing insight into the signal's complexity [Hig88].
- **Multiscale Fractal Dimension (MFD)**: Extends fractal analysis across multiple temporal scales to capture variations in signal complexity at different resolutions [Mar94].

## 2.3 Machine Learning

Machine learning (ML) is a branch of artificial intelligence (AI) focused on developing algorithms that allow systems to learn from data and improve their performance on specific tasks without explicit instructions. It forms the foundation of many modern applications, from predictive analytics to autonomous systems.

### 2.3.1 Types of Machine Learning

Machine learning techniques are broadly classified into three main categories based on the availability of labeled data and the learning objective:

- **Supervised Learning**: In supervised learning, the model is trained on labeled data, where each input is paired with a corresponding output. The objective is to learn a mapping function $f(x) \to y$ that accurately predicts the output for new inputs. Common supervised learning tasks include regression, which involves predicting a continuous value or vector from given inputs, and classification, where the model assigns inputs to discrete categories or classes.
- **Unsupervised Learning**: Unsupervised learning focuses on discovering hidden patterns, underlying structures, and relationships in data without relying on labeled inputs. It allows models to explore and organize data autonomously, making it particularly useful for extracting meaningful insights from large, unstructured datasets. Common unsupervised learning tasks include clustering, dimensionality reduction and anomaly detection.
- **Reinforcement Learning**: Reinforcement learning is a subfield of machine learning where an agent learns to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties based on its actions, allowing it to learn the optimal strategy for maximizing cumulative rewards over time. Reinforcement learning is commonly used in robotics, gaming, and autonomous systems.

### 2.3.2 Key Concepts in Machine Learning

#### Feature Engineering

Feature engineering is the process of selecting, transforming, and creating new features from raw data to improve the performance of machine learning models. It involves identifying relevant features, handling missing values and scaling numerical data to ensure that the model can effectively learn from the input data.

#### Model Training and Optimization

Training a machine learning model involves optimizing its parameters to minimize a loss function, which measures the difference between the model's predictions and the ground truth. A well-designed training process is essential to achieve high performance and generalization. Key components of model training include the following:

**Loss Function**: The loss function quantifies how well the model's predictions align with the actual values. It guides the optimization process by providing a metric to minimize. Common loss functions include:

- **Mean Squared Error (MSE)** for regression:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.3.1}$$

- **Cross-Entropy Loss** for binary or multi-class classification:

$$\text{CE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2.3.2}$$

Where:

- $y_i$ is the actual value, which is a continuous value for regression problems, and a binary label (0 or 1) for classification problems.
- $\hat{y}_i$ is the predicted value, which is a continuous value in regression problems, and a probability score or discrete class label in classification problems, depending on the model's output type.

**Optimization Algorithms**: Optimization algorithms, such as stochastic gradient descent (SGD), are used to update the model's parameters iteratively, reducing the loss and improving its predictive accuracy. The parameter updates are calculated using the gradient of the loss function with respect to the parameters:

$$\theta_{t+1} = \theta_t - \alpha \nabla L(\theta_t) \tag{2.3.3}$$

Where:

- $\theta_t$ represents the model parameters at iteration $t$.
- $\alpha$ is the learning rate that controls the step size of the updates.
- $\nabla L(\theta_t)$ is the gradient of the loss function with respect to the parameters.

**Regularization**: Regularization techniques are used to prevent overfitting, where the model performs exceptionally well on the training data but fails to generalize to unseen data. Common regularization methods include L1 and L2 regularization, which add penalty terms to the loss function to discourage complex models. The regularization strength is controlled by the hyperparameter λ, which determines the trade-off between model complexity and performance.

- **L1 Regularization**: Adds the absolute value of the weights to the loss function:

$$L_{L1} = L + \lambda \sum_{i=1}^{N} |\theta_i| \tag{2.3.4}$$

- **L2 Regularization**: Adds the squared weights to the loss function:

$$L_{L2} = L + \lambda \sum_{i=1}^{N} \theta_i^2 \tag{2.3.5}$$

**Hyperparameter Tuning**: Hyperparameters are parameters that control the learning process. Common hyperparameters include the learning rate, batch size (number of samples processed in each iteration) and model architecture parameters. Hyperparameter tuning involves selecting the optimal values for these parameters to improve the model's performance.

**Validation**: Validation is used to evaluate the model's performance on unseen data during training. Cross-validation techniques, such as k-fold cross-validation, split the data into training and validation sets multiple times to obtain more reliable performance estimates.

**Early Stopping**: Early stopping is a regularization technique that stops training when the chosen monitoring metric (e.g., validation loss) stops decreasing, preventing the model from overfitting.

### Model Evaluation

Model evaluation is a critical step in the ML pipeline to assess a model's performance and ensure its ability to generalize to unseen data. Common evaluation metrics for binary classification tasks include accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (ROC-AUC). Additionally, for regression tasks, metrics such as MSE (2.3.1) or the mean absolute error (MAE) are commonly used.

- **Accuracy**: The proportion of correctly classified samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Number of Predictions}} \tag{2.3.6}$$

- **Precision**: The proportion of true positive predictions out of all positive predictions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.3.7}$$

- **Recall**: The proportion of true positive predictions out of all actual positive samples. Also known as sensitivity or true positive rate (TPR):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.3.8}$$

- **F1 Score**: The harmonic mean of precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3.9}$$

- **ROC-AUC**: Measures a model's ability to distinguish between classes by evaluating the trade-off between the True Positive Rate (TPR), which represents the proportion of actual positives correctly identified, and the False Positive Rate (FPR), which indicates the proportion of actual negatives incorrectly classified as positives, with the AUC (Area Under the Curve) quantifying overall performance across different classification thresholds.

Where:

- TP is the number of true positive predictions.
- TN is the number of true negative predictions.
- FP is the number of false positive predictions.
- FN is the number of false negative predictions.

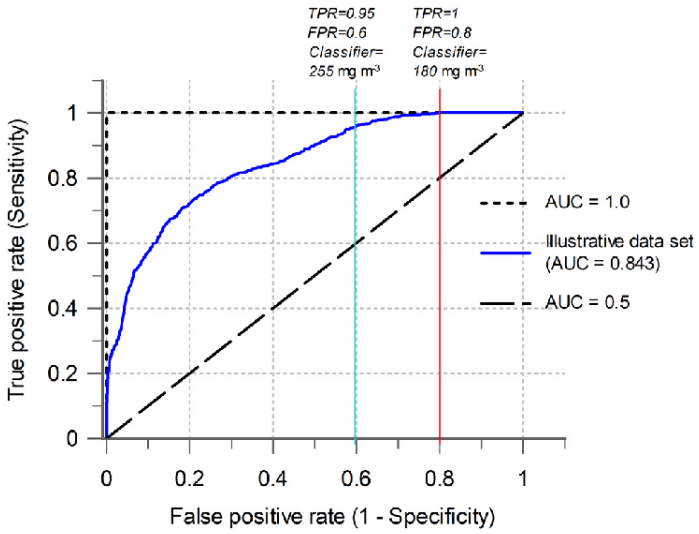



Figure 2.3.1: Example of ROC curve with ROC-AUC score [DG21].

## 2.4 Neural Network Architectures

Neural networks are computational models inspired by the structure and function of biological neurons. They consist of interconnected layers of neurons that process inputs and learn to map them to desired outputs. A typical neural network comprises an input layer, one or more hidden layers, and an output layer. Each neuron applies a weighted sum of its inputs, adds a bias, and passes the result through an activation function

to introduce non-linearity. This structure allows neural networks to model complex, non-linear relationships within data [Bis94].

Mathematically, the output of a neuron can be expressed as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.4.1}$$

Where $f$ is the activation function, $x$ is the input vector, $w$ is the weight vector and $b$ is the bias.

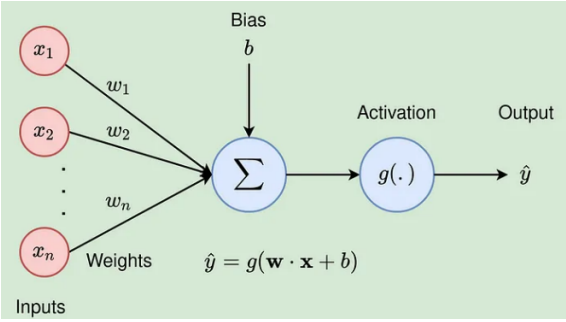



Figure 2.4.1: Example of a simple neural network architecture [Raj23].

### Activation Functions

An activation function is a vital component of a neural network that introduces non-linearity, enabling the model to learn complex patterns and distinguishing it from a simple linear model. The functions are applied to the weighted sum of inputs and biases to determine the output of a neuron. Common activation functions include:

- **Sigmoid**: The sigmoid function maps the input to a range between 0 and 1. It is commonly used in the output layer of binary classification models.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.4.2}$$

- **Tanh (Hyperbolic Tangent)**: The tanh function maps the input to a range between -1 and 1. It is commonly used in the hidden layers of neural networks.

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.4.3}$$

- **ReLU (Rectified Linear Unit)**: The ReLU function sets all negative values to zero and passes positive values unchanged. It is commonly used in deep learning models due to its simplicity and efficiency.

$$\text{ReLU}(x) = \max(0, x) \tag{2.4.4}$$

- **Leaky ReLU**: The Leaky ReLU function allows a small gradient for negative values, preventing the dying neuron problem [Maa13].

$$\text{LeakyReLU}(x) = \max(\alpha x, x) \quad \text{where} \quad \alpha \in (0, 1) \tag{2.4.5}$$

### Training and Optimization

Neural networks incorporate many of the techniques mentioned in Section 2.3.2. These include the use of loss functions, such as MSE (2.3.1) and cross-entropy loss (2.3.2), to measure prediction error, optimization algorithms like stochastic gradient descent (SGD) to iteratively minimize this error and regularization techniques such as L1 (2.3.4) and L2 (2.3.5) penalties to prevent overfitting.

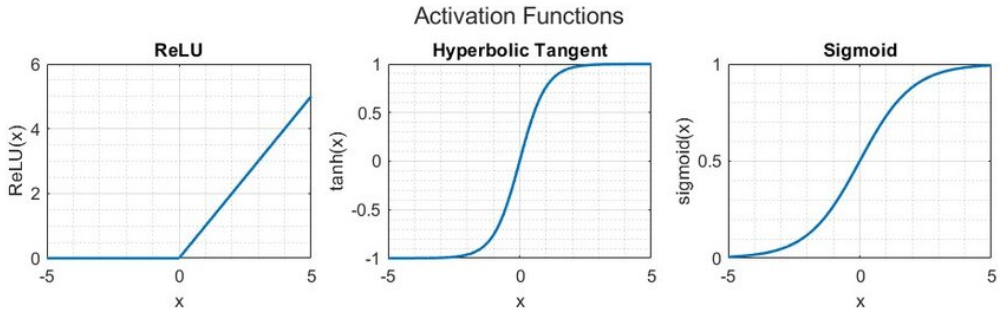



Figure 2.4.2: Common activation functions used in neural networks [İşb+23].

Concepts like hyperparameter tuning, cross-validation, and early stopping are also applied to neural networks, ensuring robust training and generalization. Building on these general techniques, neural networks incorporate additional mechanisms that are fundamental to their architecture and training:

- **Backpropagation**: The backpropagation algorithm calculates the gradient of the loss function with respect to each parameter in the network using the chain rule of differentiation. This enables the optimization algorithm to update weights and biases efficiently, layer by layer. The process involves the forward pass and the backward pass phases.

  **Forward Pass**: During the forward pass the inputs are propagated through the network and the predictions are computed from the output layer. Then, the loss is calculated by comparing the predictions with the ground truth. This value will be backpropagated through the network to update the weights.

  **Backward Pass**: During the backward pass, the gradients of the loss function with respect to the weights and biases are computed by applying the chain rule. These gradients are then used to update the weights and biases, reducing the loss and improving the model's performance.

- **Dropout**: Dropout is a regularization technique where randomly selected neurons, along with their connections, are set to zero during training. This prevents overfitting and improves the networks's generalization performance by preventing it from relying too heavily on specific neurons. The dropout rate is a hyperparameter that defines the probability of a neuron being dropped out during training.

- **Batch Normalization**: Batch normalization is a technique that normalizes the input of each layer to have zero mean and unit variance. For input $x$, the batch normalization operation is defined as:

$$\hat{x} = \frac{x - \mu}{\sigma} \quad y = \gamma\hat{x} + \beta \tag{2.4.6}$$

  Where $\mu$ is the mean and $\sigma$ is the standard deviation of the batch, and $\gamma$ and $\beta$ are learnable parameters.

- **Layer Normalization**: Layer normalization is an alternative to batch normalization, particularly useful for sequential data like text or time series. Instead of normalizing across the batch, it normalizes across the features of each input sample [BKH16]. The normalized output is computed as:

$$\hat{x_i} = \frac{x_i - \mu}{\sigma} \quad y = \gamma\hat{x_i} + \beta \tag{2.4.7}$$

  Where $\mu$ is the mean and $\sigma$ is the standard deviation of the input sample.

### 2.4.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have revolutionized the fields of image and audio processing by leveraging their ability to learn hierarchical patterns in data. Unlike traditional fully connected neural networks, CNNs take advantage of the spatial and local structure of input data. By using convolutional operations, these networks focus on small, localized regions of the input, which makes them highly efficient for complex and high-dimensional data, enabling their application in tasks such as object detection, image classification, speech recognition, and time-series analysis [MIB20; HGD17].

A CNN's architecture consists of convolutional, pooling layers, fully connected layers and activation functions, with each component contributing to the network's ability to learn complex hierarchical features. Furthermore, CNNs often incorporate techniques such as upsampling and downsampling, padding, and stride to control the spatial dimensions of the data and the size of the learned features.

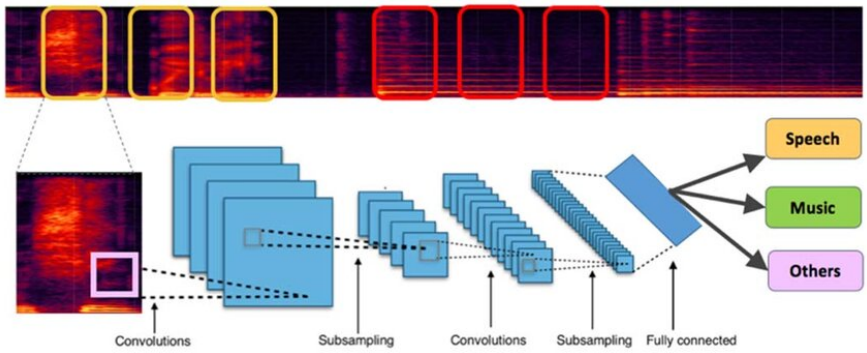



Figure 2.4.3: Example of a CNN architecture for audio task [Cha+17].

### Convolutional Layers

Convolutional layers are the fundamental building blocks of CNNs. These layers perform the convolution operation, a mathematical process that extracts local features from the input data. In a convolutional layer, a small matrix called a filter or kernel slides over the input data. At each position, the filter computes the dot product of its weights with the overlapping region of the input. This process is mathematically represented as:

$$(y * k)[i, j] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x[i + m, j + n] \cdot k[m, n] \tag{2.4.8}$$

Where:

- $x$ is the input data.
- $k$ is the kernel (filter) matrix.
- $y$ is the output.
- $M$ and $N$ are the dimensions of the kernel.
- $i$ and $j$ are the positions of the filter on the input.

**Key Parameters of Convolutional Layers**

- **Kernel Size**: Defines the dimension of the filters. Common kernel sizes include 3x3, 5x5, and 7x7. Smaller kernels capture fine-grained details, while larger kernels capture more global patterns.
- **Stride**: Determines the step size of the filter as it moves across the input (e.g. a stride of 1 moves the filter one sample at a time). A larger stride results in a smaller filter output.
- **Padding**: Refers to the addition of zeros around the input data to preserve its spatial dimensions after convolution.

During training, the weights of these filters are optimized, allowing the network to adaptively learn the most relevant features for the task. The output, known as a feature map, highlights the presence of features detected by the filter across different regions of the input. A single convolutional layer typically contains multiple filters, each producing its own feature map. These feature maps are stacked together to form the layer's output, which serves as the input to the next layer. This allows CNNs to capture diverse patterns at different spatial locations.

The benefits of convolutional layers include parameter sharing, which reduces the number of learnable parameters, and translation invariance, which enables the network to detect features regardless of their location in the input.

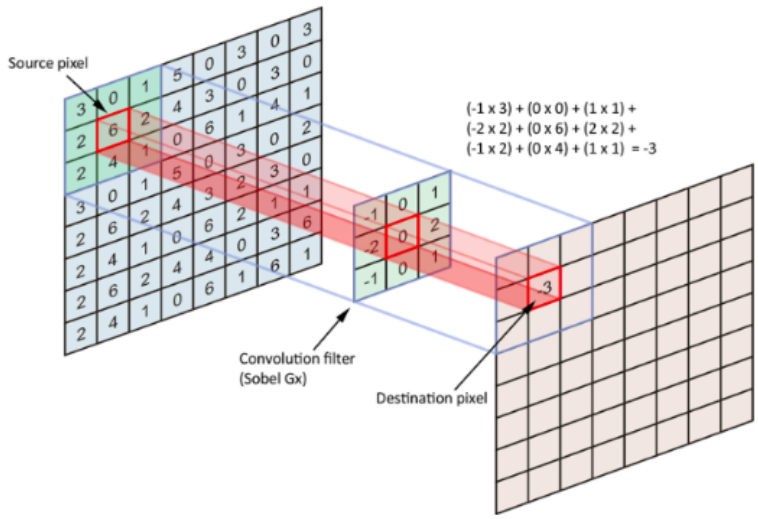



Figure 2.4.4: Visualization of convolution layer and demonstration of how the filter moves across the input image and performs the convolution operation [BI21].

### Pooling Layers

Pooling layers are designed to reduce the spatial dimensions of feature maps while retaining their most important information. By summarizing the presence of features in localized regions, pooling layers help to achieve spatial invariance, reduce computational complexity, and prevent overfitting. Pooling normally operates on small, non-overlapping regions of a feature map and applies a specified function to aggregate the values within each region. The most common pooling operations are:

- **Max Pooling**: Extracts the maximum value from each region with the purpose of capturing the most prominent feature in each region.
- **Average Pooling**: Extracts the average value from each region, providing a more smooth representation of the feature map.

The key parameters of pooling layers include the pooling size, which defines the dimensions of the pooling regions, and the stride, which determines the step size of the pooling operation. Most commonly, the stride is set equal to the pooling size.

### Fully Connected Layers

Fully connected (FC) serve as the final layers in a CNN and are responsible for combining the high-level features learned by the convolutional and pooling layers to make predictions. These layers connect every neuron in one layer to every neuron in the next layer, allowing the network to capture global patterns in the data by integrating all the features extracted by earlier layers. The output of the FC layers is computed by the equation (2.4.1).

### Additional Techniques

- **Upsampling and Downsampling**: Upsampling and downsampling are techniques used to modify the spatial dimensions of data. Upsampling increases the spatial resolution, while downsampling, on the other hand, is primarily performed through pooling layers, which reduce the spatial dimensions while retaining essential features.
- **Activation Functions**: The most common activation functions used in CNNs are ReLU and its variants, such as Leaky ReLU. After the FC layers, a softmax activation function is often used in the

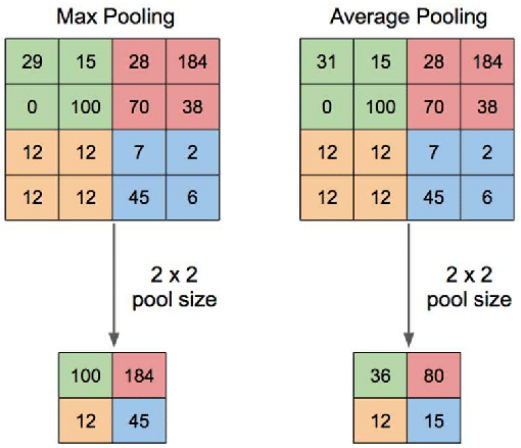



Figure 2.4.5: Visualization of max and average pooling operations [YIS19].

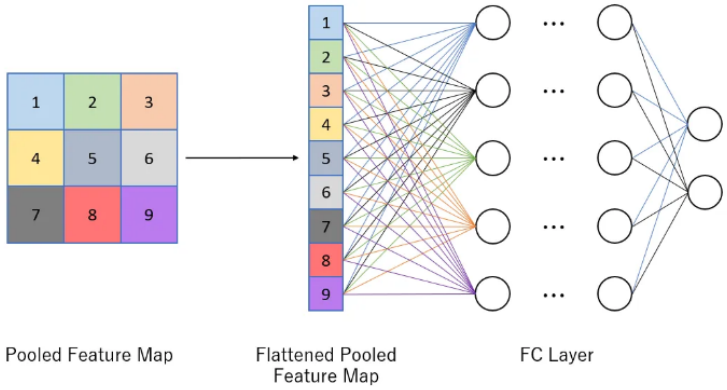



Figure 2.4.6: Visualization of fully connected layer in a neural network [HGD17].

output layer for classification tasks.

### 2.4.2 Long Short-Term Memory Networks (LSTMs)

#### Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are a class of neural networks designed to handle sequential data by maintaining a memory of previous inputs through recurrent connections. Unlike traditional feedforward networks, RNNs process inputs sequentially, making them well-suited for tasks where the order of data points is critical, such as time-series analysis, natural language processing, and speech recognition.

The key feature of RNNs is their hidden state, which serves as a dynamic memory, capturing information from previous time steps. At each time step, the network updates its hidden state based on the current input and the previous hidden state:

$$h_t = f(W_h \cdot h_{t-1} + W_x \cdot x_t + b) \tag{2.4.9}$$

Where:

- $h_t$ is the hidden state at time $t$.
- $W_h$ and $W_x$ are the weight matrices for the hidden state and input, respectively.
- $b$ is the bias term.

- $f$ is the activation function, commonly a Tanh or ReLU function.

The output of an RNN at each time step is typically computed as:

$$y_t = g(W_y \cdot h_t + c) \tag{2.4.10}$$

Where:

- $y_t$ is the output at time $t$.
- $W_y$ is the weight matrix that maps the hidden state to the output.
- $c$ is the bias term.
- $g$ is an activation function.

Despite their potential, traditional RNNs suffer from two major limitations:

- **Vanishing Gradient Problem**: Gradients become very small during backpropagation through long sequences, making it difficult for the network to learn long-term dependencies.
- **Exploding Gradient Problem**: Conversely, RNNs can also suffer from the exploding gradient problem, where the gradients grow exponentially during training, leading to unstable learning.

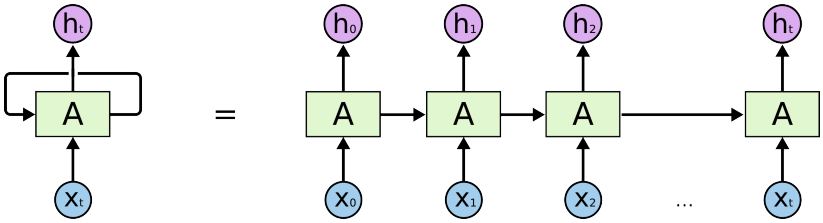



Figure 2.4.7: Basic architecture of an RNN [Ola15].

### Long Short-Term Memory Networks (LSTMs)

Long Short-Term Memory (LSTM) networks, introduced by Hochreiter and Schmidhuber [HS97], address the vanishing gradient problem in standard RNNs by incorporating gates that regulate the flow of information. Unlike traditional RNNs, which struggle to retain information over long sequences, LSTMs introduce memory cells that store information over extended periods and gating mechanisms (input, forget, and output gates) that selectively update, retain, or discard information. This structure allows LSTMs to effectively capture long-term dependencies in sequential data, making them well-suited for tasks like time-series forecasting, natural language processing, and speech recognition [Li+23; Lin+21].

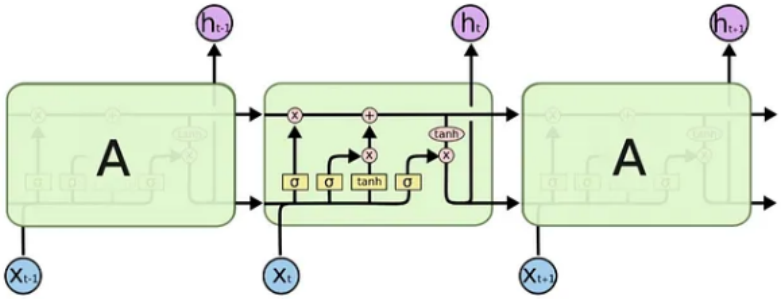
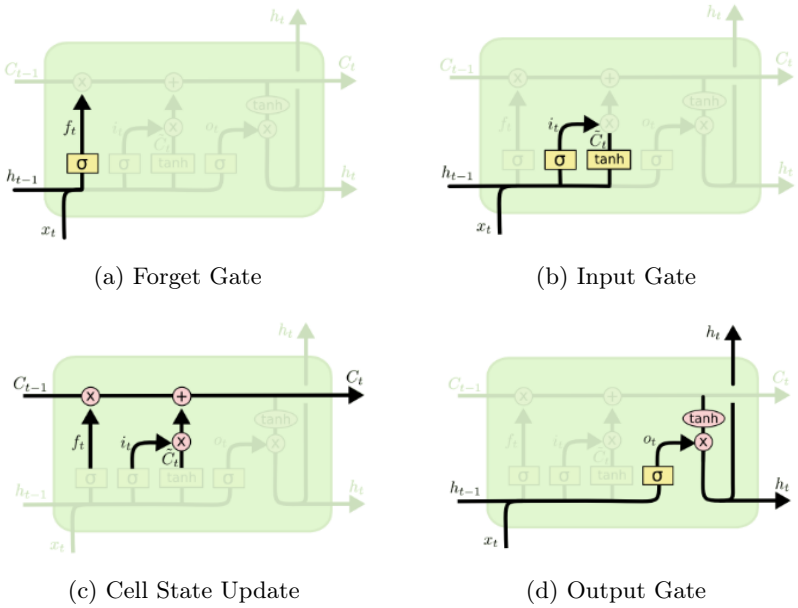



Figure 2.4.8: Internal architecture of an LSTM cell [Ola15].

At the core of an LSTM is the memory cell, which maintains information over long periods. To control the flow of information, LSTMs uses the following gates:



(a) Forget Gate (b) Input Gate

(c) Cell State Update (d) Output Gate

Figure 2.4.9: Visualization of LSTM cell components [Ola15].

### Forget Gate

The forget gate determines which information to discard from the previous cell state $c_{t-1}$. The output is a value between 0 (forget) and 1 (retain) for each element in the cell state. The forget gate is computed as:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.4.11}$$

Where is $f_t$ is the forget gate output, $x_t$ is the current input and $\sigma$ is the sigmoid activation function.

### Input Gate

The input gate decides what new information to add to the cell state. It works in conjunction with a candidate value $\tilde{c}_t$ which represents potential updates to the cell state. The input gate is computed as:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.4.12}$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2.4.13}$$

Where $i_t$ is the input gate output.

### Cell State Update

The cell state $C_t$ is updated by combining the retained information from the previous cell state $f_t \cdot C_{t-1}$ with the new information $i_t \cdot \tilde{C}_t$:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{2.4.14}$$

### Output Gate

The output gate determines what information from the cell state should be included in the hidden state $h_t$. The output gate is computed as:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.4.15}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{2.4.16}$$

Where $o_t$ is the output of the output gate.

### Variants of LSTMs

LSTM networks have inspired several variants to address specific challenges and enhance their functionality, including:

- **Bidirectional LSTMs (BiLSTMs)**: BiLSTMs process input sequences in both forward and backward directions, enabling the model to capture dependencies from past and future contexts simultaneously. This is particularly useful for tasks where full sequence information is essential, such as speech recognition and text analysis.
- **Stacked LSTMs**: Stacked LSTMs consist of multiple layers of LSTMs stacked on top of one another. This deep architecture enables the network to learn complex temporal patterns, with lower layers capturing short-term dependencies and higher layers modeling long-term relationships.
- **Gated Recurrent Units (GRUs)**: GRUs are a simplified version of LSTMs that combine the forget and input gates into a single update gate. This reduces the number of parameters and computations, making GRUs more computationally efficient than LSTMs. GRUs are particularly useful for tasks where memory efficiency is a priority.

## 2.5 Autoencoders

Autoencoders (AEs) are a class of neural networks designed for unsupervised learning tasks, primarily focusing on learning efficient representations of input data. Their architecture consists of two main components: an encoder that compresses the input into a lower-dimensional latent representation and a decoder that reconstructs the input from this representation. By minimizing the difference between the original input and its reconstruction, autoencoders aim to capture the most important features of the data.

The primary goal of autoencoders is to learn compressed, meaningful representations, which makes them valuable for tasks like dimensionality reduction, feature extraction, and anomaly detection. Unlike traditional dimensionality reduction techniques like Principal Component Analysis (PCA), autoencoders can learn non-linear relationships in data, providing greater flexibility and adaptability.

Autoencoders also serve as foundational architectures for advanced applications such as data denoising, generative modeling, and pre-training for other machine learning tasks. Over time, specialized variations like variational autoencoders (VAEs) have been developed to address specific challenges and expand their application domains [Ber+24].

### Encoder

The encoder function $g(\cdot)$ is parameterized by $\phi$ and maps the input data $x$ to a low-dimensional latent representation $z$, often referred to as the bottleneck layer, which captures the essential features of the input data. The bottleneck layer can be represented as:

$$z = g_\phi(x) \tag{2.5.1}$$

The encoder can consist of multiple layers, allowing for more complex representations of the input data.

### Decoder

The decoder function $f(\cdot)$ is parameterized by $\theta$ and reconstructs the input data $\hat{x}$ from the latent representation $z$. The reconstruction is a non-linear mapping:

$$\hat{x} = f_\theta(z) = f_\theta(g_\phi(x)) \tag{2.5.2}$$

The encoder and decoder are typically symmetrical in structure, but asymmetrical architectures are sometimes used depending on the task.

### Loss Function

The parameters $(\phi, \theta)$ of the autoencoder are learned by minimizing a loss function that measures the difference between the input data $x$ and the reconstructed output $\hat{x}$, in order to achieve $x \approx f_\theta(g_\phi(x))$. Most commonly, the MSE (2.3.1) is used as the loss function, but other metrics like cross-entropy can be used depending task.

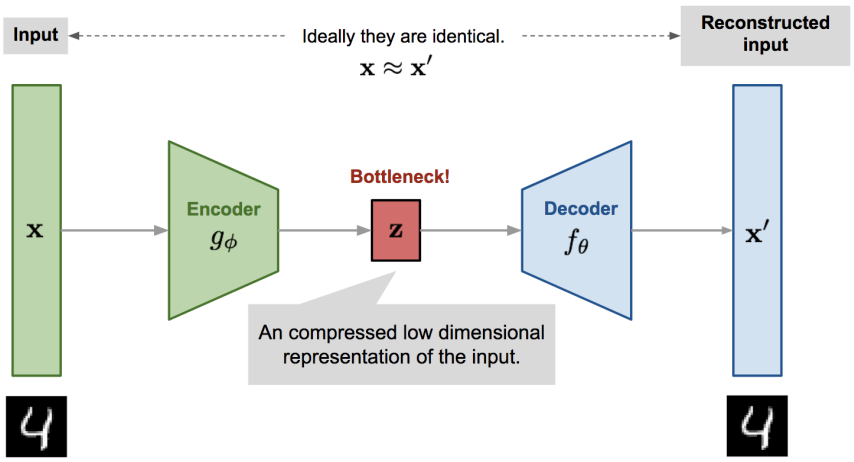



Figure 2.5.1: Example of a basic autoencoder architecture [Wen18].

### 2.5.1 Types of Autoencoders

#### Vanilla Autoencoders

Vanilla Autoencoders are the simplest and most basic form of autoencoders. They typically consist of one or more fully connected (dense) layers in both the encoder and decoder. These models are primarily used for tasks such as dimensionality reduction, feature learning, and data reconstruction. Due to their straightforward architecture, vanilla perform well on simple, structured data, but may struggle with more complex patterns.

#### Convolutional Autoencoders (CAEs)

Convolutional Autoencoders (CAEs) adapt the autoencoder architecture to handle spatially structured data, such as images, audio spectrograms, and other grid-like data. Instead of fully connected layers, CAEs use convolutional layers in the encoder and decoder to capture local patterns and spatial hierarchies.

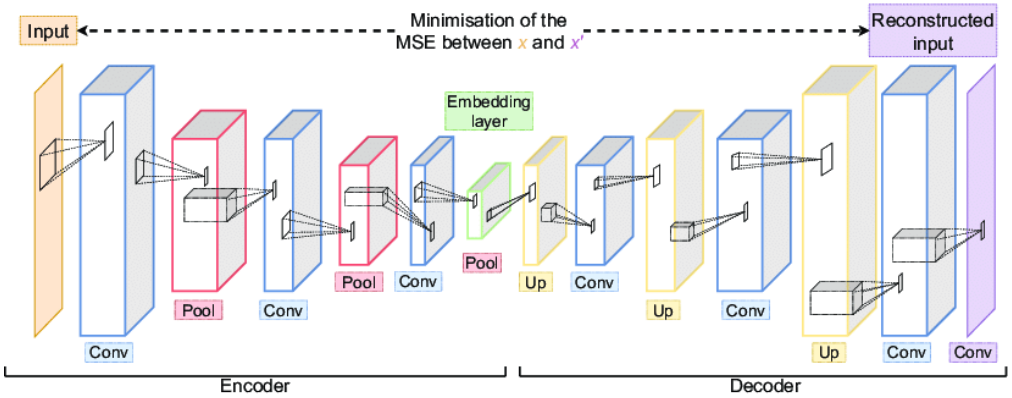



Figure 2.5.2: Example of a CAE architecture [Gou+21].

The encoder of a CAE consists of convolutional layers to extract spatial features, progressively reducing the input's spatial dimensions through pooling or strided convolutions. The decoder uses transposed convolutional layers to upsample the latent representation and reconstruct the input.

#### LSTM Autoencoders (LSTMAEs)

LSTM Autoencoders (LSTMAEs) are designed for sequential data, such as time series, text, and speech.

They combine the autoencoder architecture with LSTM cells to capture temporal dependencies and patterns in the data. An LSTM network is used in both the encoder and decoder to process sequential data, with the encoder compressing the input sequence into a latent representation using the last hidden state output and then the decoder then reconstructs the original sequence from this latent representation.

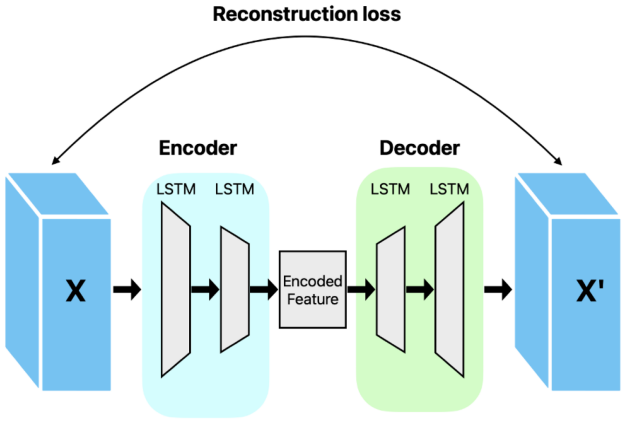



Figure 2.5.3: Example of a LSTMAE architecture [Lee+24].

### 2.5.2 Variational Autoencoders (VAEs)

Variational autoencoders (VAEs) are a type of autoencoder that learns a probabilistic latent space representation of the input data. They are trained to maximize the likelihood of the input data under the learned latent space distribution, which is typically a Gaussian distribution. Additionally, VAEs are designed to generate new data points by sampling from the learned latent space, making them suitable for generative modeling tasks [CCM24].

**Architecture**

- **Encoder**: Maps the input $x$ to a distribution in the latent space, producing the mean $\mu$ and variance $\sigma^2$ of the distribution.
- **Latent space sampling**: Samples latent representations $z$ from the learned distribution using the reparameterization trick:

$$z = \mu + \sigma \cdot \epsilon \tag{2.5.3}$$

  Where $\epsilon \sim \mathcal{N}(0, 1)$ is a random noise from a standard normal distribution.
- **Decoder**: Reconstructs the input from the sampled latent representation.

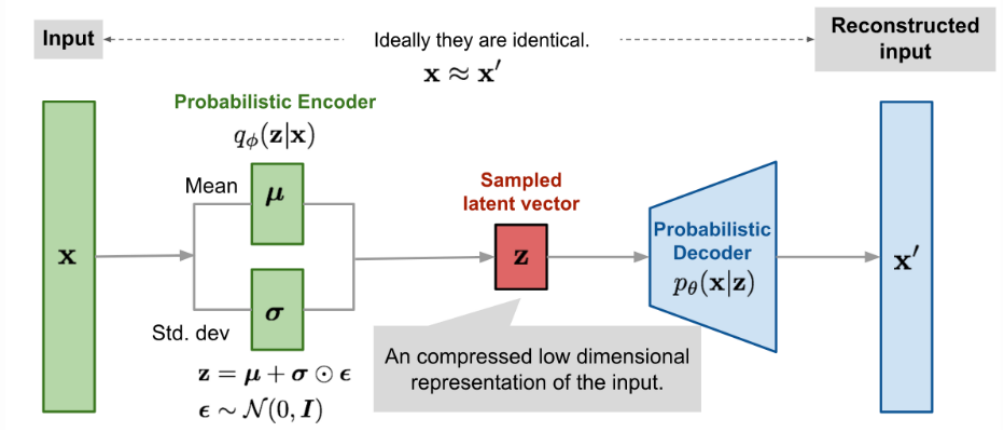



Figure 2.5.4: Example of a VAE architecture [Wen18].

### Loss Function

The loss function of a VAE consists of two components: the reconstruction loss and the KL divergence loss. Similar to traditional autoencoders, the reconstruction loss measures the difference between the input and the reconstructed output, while the KL divergence loss regularizes the latent space distribution to be close to a standard Gaussian distribution.

The Kullback-Leibler (KL) Divergence is a measure of how one probability distribution $q(z|x)$, the posterior distribution learned by the VAE, differs from a second probability distribution $p(z)$, typically a standard normal distribution. The KL divergence between two distributions $q(z|x)$ and $p(z)$, for the case of a spherical isotropic Gaussian distribution, is given by:

$$D_{KL}(q(z|x)||p(z)) = \frac{1}{2}\sum_{i=1}^{K}(\sigma_i^2 + \mu_i^2 - \log(\sigma_i^2) - 1) \tag{2.5.4}$$

Where:

- $\mu_i$ and $\sigma_i$ are the mean and standard deviation of the latent distribution for the $i$-th dimension.
- $K$ is the dimensionality of the latent space.

The total loss of a VAE is the sum of the reconstruction loss and the KL divergence loss:

$$\mathcal{L}_{VAE} = \mathcal{L}_{reconstruction} + \beta \cdot D_{KL}(q(z|x)||p(z)) \tag{2.5.5}$$

Where $\beta$ is a hyperparameter, introduced in β-VAEs [Hig+17], that controls the trade-off between the two losses.

## 2.6 Anomaly Detection

Anomaly detection is a critical area in machine learning that focuses on identifying patterns or observations that deviate significantly from the expected behavior or norm. These anomalies, often referred to as outliers, can represent rare but important events, such as system failures, fraud, or health-related abnormalities. The ability to detect such deviations is essential across various domains, including finance, healthcare, manufacturing, and cybersecurity.

Traditional anomaly detection methods include statistical approaches (e.g., Z-score analysis), clustering techniques and supervised ML models, such as Support Vector Machines (SVMs). However, with the rise of autoencoder architectures, anomaly detection has seen significant advancements in recent years. Autoencoders, in particular, have proven to be an effective unsupervised learning approach in capturing complex patterns and identifying anomalies in high-dimensional data.

A typical autoencoder-based anomaly detection system includes the following steps:

- **Learning Normal Patterns:** The autoencoder is trained solely on data representing normal conditions, allowing the model to learn to encode and reconstruct the structure of normal inputs.
- **Defining Anomalies:** Once trained, the model reconstructs new unseen data samples, and an anomaly score is computed based on the difference between the original and reconstructed input. Common scores include MSE between input and reconstruction, Mahalanobis distance of reconstruction errors, or KL divergence in the latent space.
- **Thresholding for Outlier Detection:** A threshold is defined to distinguish anomalies from normal samples. This threshold can be set empirically or based on statistical analysis of reconstruction errors. In practice, models are evaluated using ROC-AUC scores, which provides a threshold-independent performance metric.

# Chapter 3

# Literature Review

## 3.1 Relapse Detection in Mental Health

### 3.1.1 Digital Phenotyping

As mentioned in 1.2, digital phenotyping is an emerging approach that takes advantage of passive data collection from smartphones and wearables to monitor behavioral and physiological patterns, offering new possibilities for mental health care. It enables the continuous and unobtrusive tracking of individuals, capturing real-time insights into mood, behavior, and cognitive function. Various physiological signals, such as heart rate variability, movement patterns from gyroscope and accelerometer sensors, and behavioral data from smartphones, including phone usage duration and social communication metrics, provide valuable indicators of stress levels, mood states, and cognitive decline [Adl+20; Ikä+24; Osm+15]. By providing objective and scalable data, digital phenotyping addresses limitations in traditional subjective clinical evaluations such as interviews, self-reports, and questionnaires, paving the way for more effective mental health monitoring and early intervention [AMC17].

#### Supervised Machine Learning Approaches

Several studies have demonstrated the potential of digital phenotyping for predicting relapse and monitoring mental health conditions through passive smartphone data collection. A pilot study on digital phenotyping [Bar+18] has demonstrated the potential of smartphone-based phenotyping to predict relapse in schizophrenia patients. Using the Beiwe app, the researchers collected passive data, such as mobility and communication patterns, and active self-reported symptom data from seventeen (17) participants over three months. The findings revealed that behavioral anomalies, detected through smartphone use, increased by 71% in the two weeks leading to relapse. This approach highlighted the potential of smartphones as low-cost, scalable tools for real-time mental health monitoring, offering critical early warning signs that could prompt timely interventions.

The study by Osmani et al. [Osm+15] involved twelve (12) patients with bipolar disorder who were monitored for an average of 12 weeks using smartphones equipped with accelerometers, GPS, and voice analysis capabilities. The study demonstrated that sensor data, such as physical activity and location patterns, could predict the state of the patient with up to 81% accuracy while the fusion of phone and sensor modalities could detect episode changes with 94% precision and 96% recall, by using traditional classification methods such as Naive Bayes and k-Nearest Neighbors.

Similarly, the study by Ikaheimonen et al. [Ikä+24] explored the potential of passive smartphone data to predict and monitor depression symptoms. Thirty-two (32) behavioral markers were identified, including GPS location, app usage, communication logs, and battery levels, collected from ninety-nine (99) participants participants over one year. By combining this data with biweekly survey scores, the study used supervised ML approaches, such as XGBoost [CG16], achieving up to 82% accuracy in detecting depression and 75% accuracy in predicting depression state transitions. Significant predictors included screen-on time, app usage, and total distance traveled, highlighting behavioral patterns associated with depressive states.

Although the results were promising, these studies faced several challenges. Small sample sizes and short monitoring periods limited the generalizability of the findings, emphasizing the need for larger, long-term datasets to validate the results. Additionally, the last study [Ikä+24] highlighted the need for developing personalized models, applying temporal modeling techniques, and integrating deep learning methods to enhance the robustness and performance of predictive models.

#### Unsupervised Machine Learning Approaches

The CrossCheck system [Adl+20] represents a significant advancement in leveraging smartphone-based passive sensing for monitoring and managing symptoms of serious mental illnesses, particularly Schizophrenia Spectrum Disorders (SSDs). CrossCheck collected continuous behavioral data from participants using passive smartphone sensors alongside occasional self-reported assessments. The primary objective was to explore how unobtrusive digital phenotyping could detect early warning signs of psychotic relapse, enabling timely clinical interventions. The behavioral data collected from sixty (60) individuals with SSDs, eighteen (18) of whom experienced relapses, included the following:

- **Physical Activity:** Measured using accelerometer data.
- **Phone Usage:** Patterns of app usage, text messaging, and screen activity.
- **Social Interaction:** Frequency and duration of calls and detected conversations.
- **Geolocation:** Movement patterns and time spent at different locations.
- **Sleep Metrics:** Estimation of sleep onset, duration, and wake times.

Daily behavioral features were extracted from the passive data, such as mean acceleration, call activity (number and duration of incoming, outgoing, missed, rejected, and blocked calls), conversation frequency and duration, location data (time spent in primary and secondary locations, total distance traveled), screen activity (frequency and duration of phone use), sleep metrics (duration, onset, wake time), and text messaging behavior (number of received, sent, drafted, failed, and queued messages).

The study employed autoencoders for anomaly detection, leveraging their ability to reconstruct patterns of "healthy" behavior and flag deviations as potential anomalies. Specifically, a FNN and a GRU-based sequence-to-sequence (Seq2Seq) AE, as shown in Figure 3.1.3, were trained on the data to detect behavioral anomalies indicative of relapse. Both models were trained exclusively on data from "healthy" periods to establish a baseline of normal behavior. The reconstruction errors were then analyzed to detect anomalies in the days leading up to relapse. A baseline comparison with the Local Outlier Factor (LOF) algorithm [Bre+00] provided additional context for the models' performance, highlighting the robustness of neural network approaches in identifying subtle behavioral shifts.

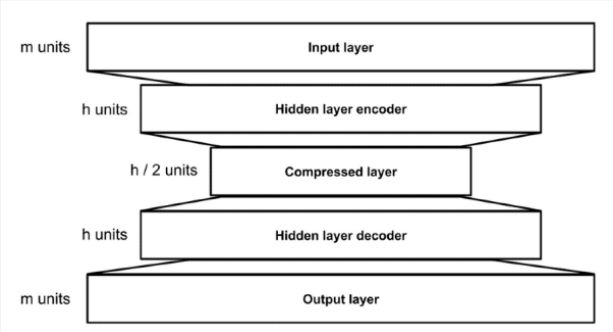



Figure 3.1.1: FNN-AE

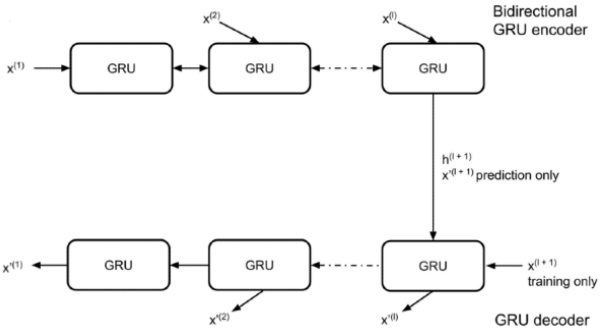



Figure 3.1.2: GRU-Seq2Seq-AE

Figure 3.1.3: Autoencoder architectures used in the CrossCheck system for anomaly detection [Adl+20].

The models were optimized using varying hidden layer configurations and training dataset sizes, with the FNN-AE achieving the best results with 40 hidden units and 80% training data, achieving a median sensitivity of 0.25 and specificity of 0.88.

CrossCheck demonstrated the potential of passive sensing to identify behavioral changes preceding relapse. For instance, significant reductions in physical activity and altered communication patterns often marked the near-relapse period. However, challenges such as missing data, variability in individual behaviors and the need for personalized models were identified as areas for future research.

### 3.1.2 Speech-Based Relapse Detection

As a natural and unobtrusive medium, speech serves as a rich source of markers that reflect cognitive, emotional, and behavioral states, with features such as pitch, intensity, and prosody revealing mood fluctuations and can be used to detect early signs of relapse in mental health disorders [LBG20; Fau+16].

### Detecting Manic State of Bipolar Disorder Based on Support Vector Machine and Gaussian Mixture Model Using Spontaneous Speech

The study by Pan et al. [Pan+18] investigated the use of machine learning techniques to detect manic states in patients with bipolar disorder based on their spontaneous speech. The study involved twenty-one (21) hospitalized patients, all diagnosed with bipolar disorder and and having experienced a manic episode during the course of the study. Speech data from conversations between patients and clinicians was collected using a smartphone, ensuring that the calls took place in a controlled, sound-insulated environment. Each patient provided multiple recordings, capturing both manic and euthymic states.

The extracted speech features were categorized into two groups: (i) prosodic, which included pitch and formants and (ii) spectral, such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs) and Gammatone Frequency Cepstral Coefficients (GFCCs).

A Support Vector Machine (SVM) and a Gaussian Mixture Model (GMM) were trained on the speech features to classify manic and euthymic states. The models were evaluated in two settings:

- **Single-Patient Detection**: The SVM model outperformed the GMM, achieving an average accuracy of 88.56% compared to GMM's 84.46%. This demonstrated the efficacy of SVM in scenarios with smaller datasets and individualized patterns.
- **Multiple-Patient Detection**: The GMM excelled in this context, with an accuracy of 72.27%, significantly higher than SVM's 60.87%. GMM's generative nature made it better suited for capturing variability across a larger, diverse dataset.

The results highlighted the distinct advantages of both models: SVM for personalized, small-scale applications, and GMM for broader, population-level predictions. The study emphasized the potential of speech analysis as a non-invasive tool for monitoring mood states in bipolar disorder, paving the way for integrating machine learning into clinical practices for mood disorder management. The researchers also noted the need to explore additional speech features and mood states.

### Attention-based Convolutional Neural Network and Long Short-term Memory for Short-term Detection of Mood Disorders based on Elicited Speech Responses

Another novel study by Huang et al. [HWS18] proposed a methodology for the short-term detection of mood disorders, including unipolar depression and bipolar disorder, using elicited speech responses. Participants watched six (6) emotion-eliciting videos that triggered emotions such as happiness, sadness, and anger, followed by interviews where their speech responses were recorded and analyzed.

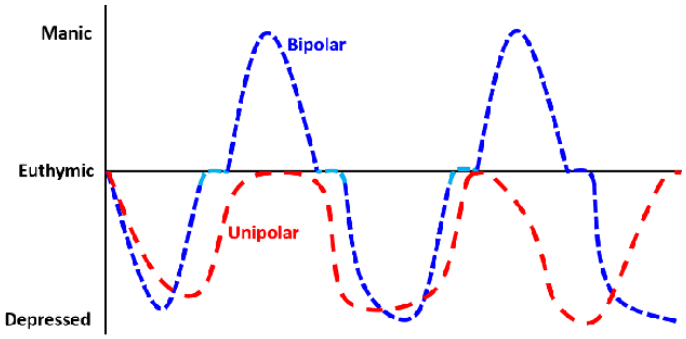



Figure 3.1.4: Example of mood shifts in unipolar depression and bipolar disorder [Wu+23].

Two databases were collected and utilized for training and evaluation in this study: (i) the CHI-MEI Mood Disorder Database, which contained elicited speech responses from forty-five (45) participants (fifteen (15) with bipolar disorder, fifteen (15) with unipolar depression, and fifteen (15) healthy controls) and (ii) the Multimedia Human–Machine Communication (MHMC) Emotion Database, a labeled dataset of emotional

expressions. Due to the complexity of labeling the diverse emotions in the CHI-MEI database, the MHMC dataset was used to train a model for generating emotion profiles (EPs).

For the CHI-MEI database, speech data were collected after participants viewed 6 carefully selected emotion-eliciting videos, each designed to evoke happiness, sadness, anger, fear, surprise, or disgust. Participants then answered five (5) emotion-related questions in an interview setting. To bridge the gap between the labeled emotion data (MHMC) and the unlabeled mood disorder data (CHI-MEI), the Hierarchical Spectral Clustering (HSC) algorithm [Liu+13] was employed. This process adapted the emotion database into the mood disorder space by iteratively aligning clusters of features, addressing the inherent bias in transferring knowledge between domains.

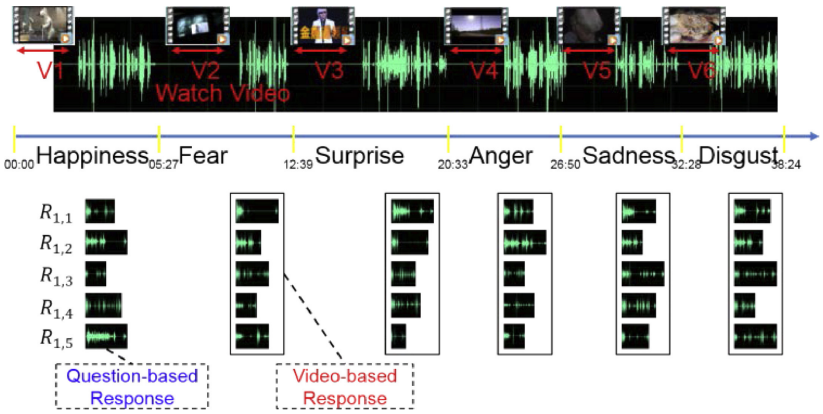



Figure 3.1.5: Example of mood disorder database structure for each speech response after watching corresponding emotion-eliciting video [HWS18].

### Modeling Emotional Profiles

To extract localized emotional features from each speech response, the study implemented an Attention-based CNN. The CNN architecture included convolutional layers to capture short-term emotional characteristics within speech signals, pooling layers for dimensionality reduction and fully connected layers to output the Emotion Profiles (EPs). These EPs provided a vector representation of the local intensity and variation of emotions expressed during each response. An attention mechanism enhanced the CNN by weighting parts of the speech response that were most relevant to the emotional context, ensuring that crucial features were emphasized during the EP generation process.

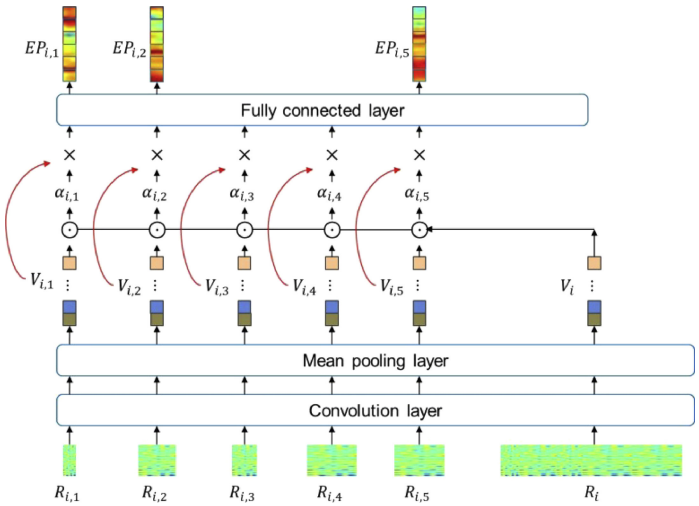



Figure 3.1.6: Proposed architecture of the Attention-based CNN for generating EPs [HWS18].

### Temporal Analysis of Emotion Profiles

While CNNs captured the local features of individual speech responses, the study recognized the importance of modeling the temporal progression of emotions across all six responses. To achieve this, an LSTM network was employed. The LSTM captured the sequence of EPs, modeling the evolution of emotional patterns over time. Prior to feeding the EP sequences into the LSTM, an Multilayer Perceptron (MLP)-based attention model was employed to refine the weights of the video-based responses. The MLP assigned importance to each video-based response (consisting of concatenated question-based EPs) by evaluating their relevance to mood disorder detection. Finally, the LSTM processed the weighted EP sequences, generating a final prediction of the participant's mood disorder state.

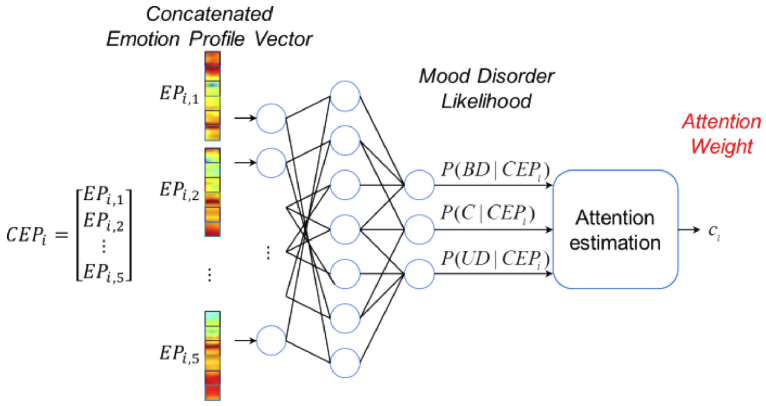



Figure 3.1.7: Proposed architecture of the MLP-based attention model [HWS18].

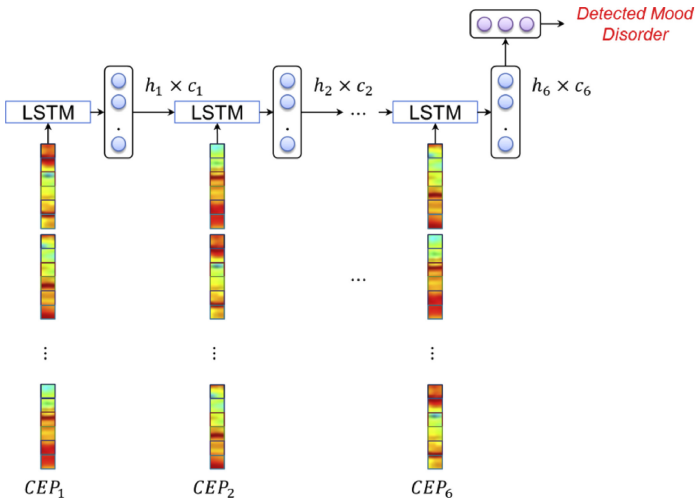



Figure 3.1.8: Proposed architecture of the LSTM-based mood disorder detection model with attention [HWS18].

The proposed methodology combining Attention-based CNNs and LSTM networks demonstrated a superior performance for mood disorder detection, achieving an overall accuracy of 75.56%. This significantly outperformed traditional classifiers such as SVMs and CNN-based models without temporal analysis. The study also found that emotion-eliciting videos designed to provoke anger, sadness, and disgust resulted in more accurate mood disorder detection compared to other emotions like fear or happiness. Additionally, the attention-enhanced LSTM network provided a critical advantage by modeling the temporal progression of emotional states across responses, validating the importance of sequence-level analysis.

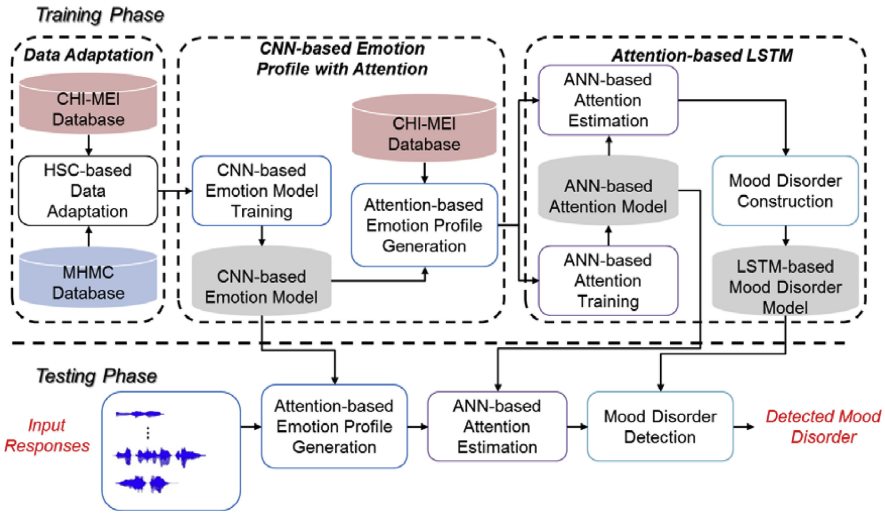



Figure 3.1.9: Overview of the proposed mood disorder detection system framework [HWS18].

### Detecting Abnormal Mood using Everyday Smartphone Conversations

In the study by Gideon et al. [Gid+19], reserachers proposed a novel approach for detecting abnormal mood states in individuals with bipolar disorder using everyday smartphone conversations. Additionally, this research is part of the PRIORI (Predicting Individual Outcomes for Rapid Intervention) project, which focuses on passive mood monitoring using natural phone conversations, emphasizing the prediction of when clinical intervention is necessary.

The study utilized the PRIORI dataset [Kho+18], which included 51,970 phone calls (approximately 3,997 hours) from fiftly-one (51) individuals with bipolar disorder and nine (9) healthy controls. Conversations were passively recorded using smartphones equipped with the PRIORI app, and data were labeled using clinical mood assessments such as the YMRS [You+78] and Hamilton Depression Rating Scale (HDRS) [Ham76].

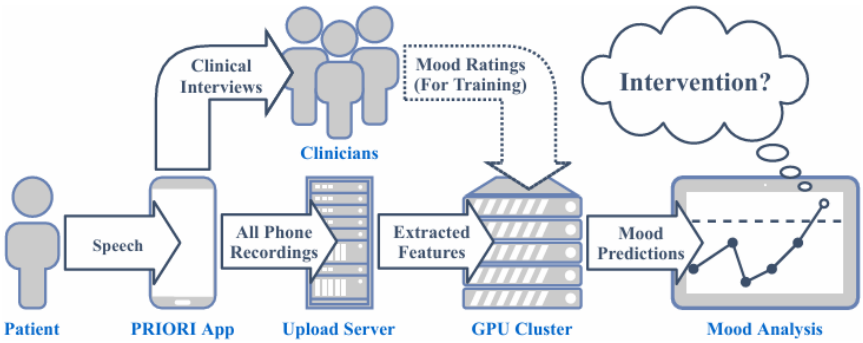



Figure 3.1.10: Overview of the proposed system [Gid+19].

After data collection, the following features were extracted from the phone conversations:

- **Emotion Features:** Emotion-related data were extracted from speech using a Multiclass Adversarial Discriminative Domain Generalization (MADDoG) model [GMP19] trained on multiple emotion datasets. This model identified levels of emotional intensity (activation) and positivity or negativity (valence) in the speaker's voice. These emotion levels were summarized into statistics like averages, maximums, and ranges across each call or day, providing a detailed picture of how emotions changed during conversations.
- **Transcript Features:** Linguistic features, such as word choice, sentence structure, speaking speed

and pauses were analyzed using an automatic speech recognition (ASR) tool. The use of emotional words, grammar patterns and speech timing were also analyzed.

**Temporal Normalization (TempNorm)**

The study introduced an algorithm called Temporal Normalization (TempNorm) to adapt to individual mood baselines and detect mood abnormalities by learning and adapting to an individual's unique mood baseline over time. The algorithm starts with a general population baseline and updates it dynamically using an exponentially weighted moving average (EMA) and exponentially weighted moving variance (EMVar). A half-life parameter controls how quickly the algorithm adapts: shorter half-lives make it more sensitive to recent changes, while longer half-lives create a more stable baseline by incorporating a larger history of data.

A Dense Neural Network (DNN) was employed to predict mood abnormality ratings based on TempNorm-processed features. The DNN incorporated a TempNorm layer within its architecture to align feature spaces with personalized mood baselines.

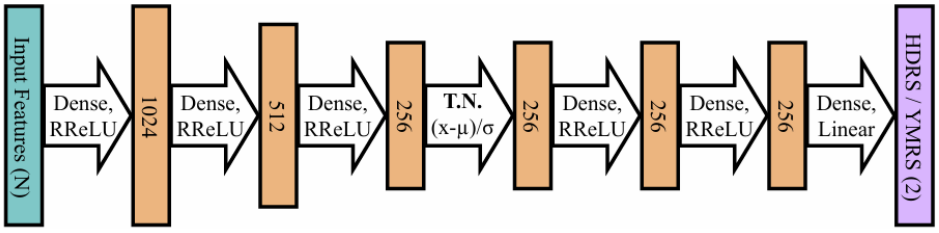



Figure 3.1.11: Proposed DNN model architecture with TempNorm layer [Gid+19].

The above methodology achieved Unweighted Average Recall (UAR) scores of 0.70 for structured clinical conversations and 0.68 for unstructured personal conversations. Transcript-based features performed best for clinical calls, likely due to the structured nature of these conversations, while emotion features worked well across both structured and natural speech. The results highlighted the importance of adapting models to individual baselines for improved anomaly detection. However, the study faced challenges, including balancing sensitivity and stability in the TempNorm algorithm through the half-life parameter, where shorter half-lives risked overfitting to recent fluctuations and longer ones delayed adaptation. The researchers also noted variability in data quality, such as transcription errors and noise in personal calls, which impacted performance. Despite these challenges, the study paved the way for personalized mood monitoring systems and emphasized the need for further work on including additional features.

The studies reviewed highlight the significant role of speech analysis in predicting relapses in mental health disorders, particularly bipolar disorder, schizophrenia, and depression. The various machine learning techniques applied to spontaneous, elicited, and natural speech conversations have demonstrated promising results in identifying mood fluctuations, and early warning signs of relapse. Key findings include the effectiveness of prosodic and spectral speech features in distinguishing between manic and euthymic states, the advantage of deep learning models such as CNNs and LSTMs in modeling emotional progression, and the ability of automatic speech recognition (ASR) tools to extract linguistic and acoustic markers relevant to mood disorders. Furthermore, personalized models such as those using Temporal Normalization (TempNorm) have shown the importance of adapting to individual speech baselines to improve relapse detection accuracy. Despite challenges such as data variability, and the need for larger datasets, speech remains a non-invasive, scalable, and objective marker for early relapse detection.

## 3.2 Modality Fusion in Mental Health

Mental health conditions can manifest across multiple aspects, such as behavior, speech, and physiological responses, making multimodal analysis a valuable tool for detecting, assessing, and monitoring these conditions. By integrating multiple data types, such as audiovisual signals, textual information and behavioral patterns, multimodal approaches can provide a greater understanding of mental health states and enable more accurate predictions of episodes or relapses.

### 3.2.1 Audiovisual and Textual Feature Fusion

The integration of audiovisual and textual features has been extensively explored in mental health research to enhance the detection of conditions such as depression. The DAIC-WOZ dataset [Rin+17] serves as the primary dataset in many studies focused on depression detection. It consists of 189 clinical interviews annotated with PHQ-8 scores [BSB96], providing multimodal data, including audio recordings, textual transcriptions and visual features such as gaze, pose, and facial Action Units (AUs). The dataset captures interactions between participants and a virtual interviewer, offering rich contextual data for analysis. The following studies demonstrate the effectiveness of multimodal fusion approaches in depression detection using the DAIC-WOZ dataset.

The study by Othmani et al. [OZ22], introduces a multimodal system for depression relapse prediction. The system processes the DAIC-WOZ dataset by extracting audio features using a modified VGGish network [SZ14] and visual features via a 1D-CNN applied to the AUs. Data fusion occurs at the feature level, where the extracted high-dimensional audio and visual embeddings are concatenated to form a single feature vector. These features are passed to a Deep Neural Network (DNN) for depression detection, achieving 78.97% accuracy. An anomaly detection distance-based approach called Model of Normality (MoN) [Abu+20], and trained on anomaly-free data, further predicts relapse by measuring similarity between a test subject's audiovisual encoding and representations of "depression" and "non-depression" (built from the training data) achieving up to 82.55% accuracy.

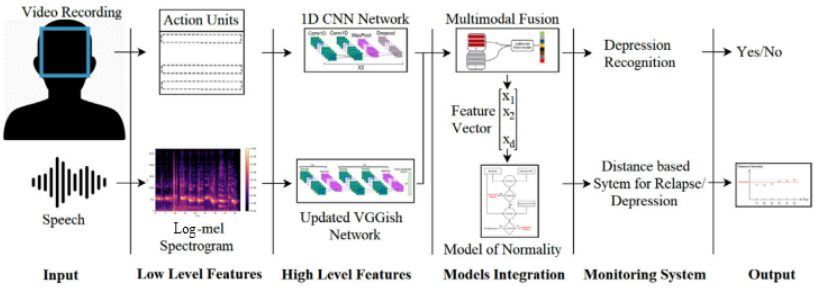



Figure 3.2.1: Proposed multimodal framework for depression recognition and depression relapse prediction. [OZM22]

Similarly, the study by Fontinele et al. [Alm+24] proposed a stacking ensemble model for automatic depression detection (ADD) by combining MFCC-based audio features, OpenFace visual descriptors, and Word2Vec text embeddings. Individual DNNs trained on each modality were stacked into a meta-classifier, achieving an F1 score of 0.857 and an AUC of 0.913. The model effectively addressed class imbalance and overfitting using SVM-SMOTE and dropout regularization.

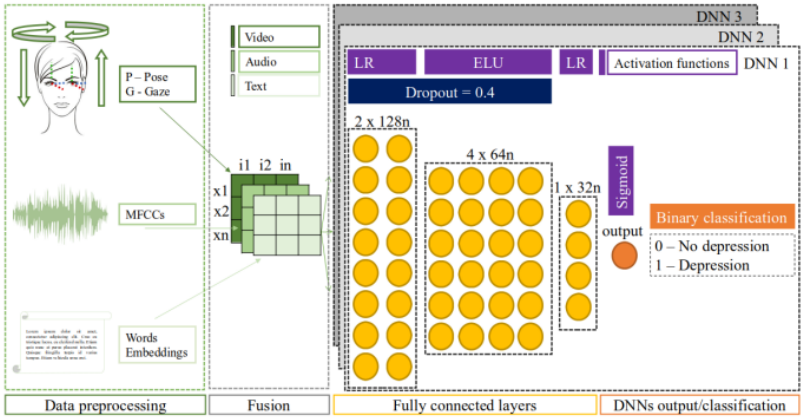



Figure 3.2.2: Proposed architecture for multimodal ADD [Alm+24].

In another similar study [**Flores2022**], the researchers proposed AudiFace, a multimodal deep learning model

designed to enhance depression screening by integrating audio, text transcripts, and temporal facial features from clinical interview videos. The model extracts features using pre-trained architectures like VGGish and BERT [Dev+19], combined with LSTMs and self-attention. These features are fused through a linear layer for classification. AudiFace outperformed the state-of-the-art AudiBERT [TTR21], having included the temporal facial features aspect, in 13 of 15 datasets of the DAIC-WOZ , achieving an average F1 score of 0.71, with eye gaze features being particularly effective.

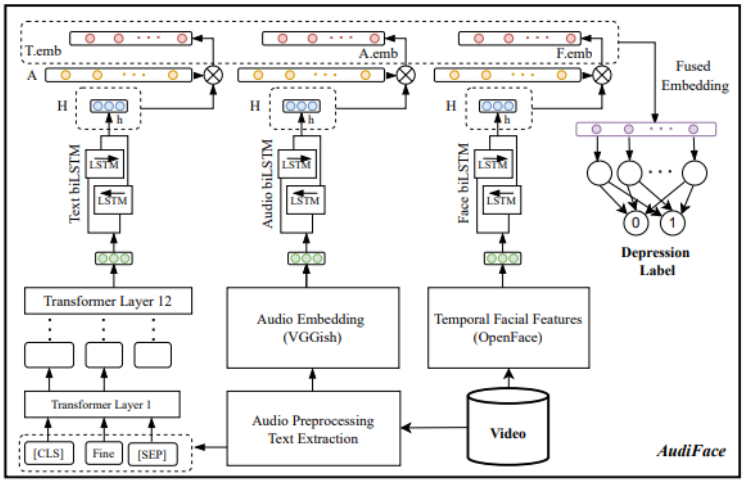



Figure 3.2.3: Proposed AudiFace framework [**Flores2022**; TTR21].

Finally, in the study by Lam et al. [LHL19], the researchers designed a multimodal fusion model for depression detection. By integrating context-aware topic modeling with deep learning, the study introduces a data augmentation framework that generates balanced training datasets from text and audio features. The models include a 1D-CNN for audio and a fine-tuned Transformer for text, with a feedforward network for multimodal fusion. The augmented models achieved significant results, with F1 scores of 0.67 for audio, 0.78 for text, and 0.87 for multimodal fusion.

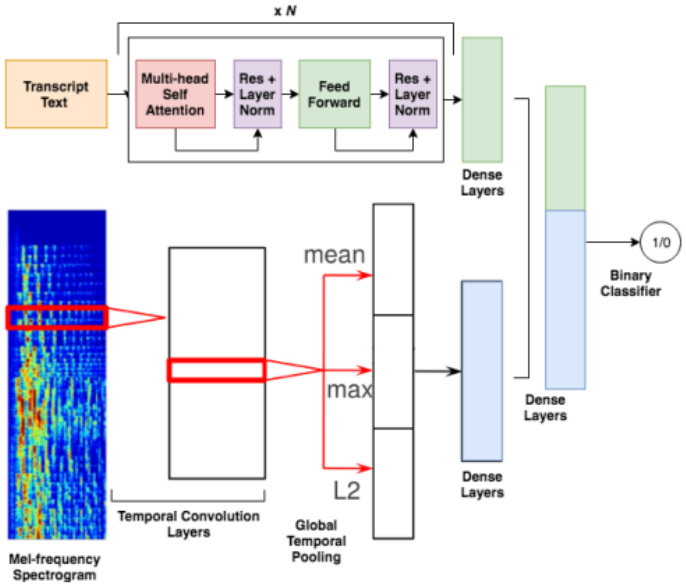



Figure 3.2.4: Proposed multimodal model consisting of Transformer and CNN models [LHL19].

The reviewed studies highlight the effectiveness of combining audiovisual and textual data in depression

detection and relapse prediction. Feature-level fusion techniques such as concatenation of embeddings have demonstrated strong performance, while deep learning models incorporating pre-trained architectures, have further enhanced detection accuracy. Despite these advances, data scarcity and class imbalance, particularly within DAIC-WOZ, often hinder model performance, necessitating the use of data augmentation and resampling techniques to improve generalizability. Additionally, feature alignment across modalities poses another major challenge, as mismatched timestamps, inconsistent data quality, and variations in data collection can disrupt effective fusion and model training.

### 3.2.2 Physiological and Behavioral Feature Fusion

In addition to audiovisual and textual data, physiological and behavioral features have been integrated to enhance mental health detection and monitoring. The study by Su et al. [Su+21] presents a smartphone-based system for assessing bipolar disorder, by predicting the score of HDRS [Ham76] and YMRS [You+78] scales. The system used heterogeneous digital phenotyping data, including passive (GPS, sleep, mood) and active (self-reports, text, speech, video) inputs. Weekly data subsets were created to account for behavioral patterns over different time frames (e.g., weekdays versus weekends) and features were extracted from each data type, such as GPS entropy, sleep regularity, emotional profiles and self-reported scale scores. Feature fusion was performed by concatenating combinations of unimodal features into multimodal feature vectors, resulting in thirty-one (31) possible configurations of data combinations. These feature vectors were used to train models like Lasso Regression, ElasticNet Regression, Polynomial Regression, and DNNs. Lasso and ElasticNet Regression achieved the best results, with MAE values of 2.73 for HDRS and 1.06 for YMRS. While feature fusion enhanced performance, challenges like data scarcity, handling missing data, and selecting the most predictive features could be areas for future improvement.

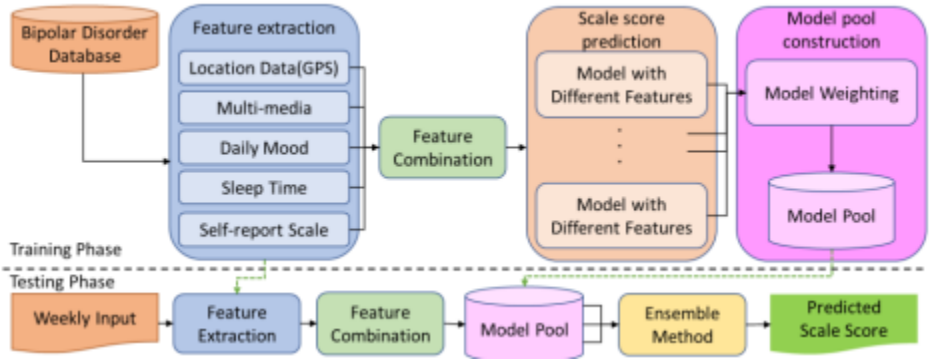



Figure 3.2.5: Proposed system framework for HDRS and YMRS prediction [Su+21].

### 3.2.3 Textual, Behavioral and Visual Feature Fusion

The study by Wang et al. [Wan+22] introduced FusionNet, a multitask learning framework designed to detect depression on online social networks (OSNs) using heterogeneous modalities, including text, social behavior (e.g. tweets and posting behavior) and image data. The proposed framework is evaluated using the developed Weibo User Depression Detection Dataset (WU3D), containing 10,325 depressed users and 22,245 normal users. The developed FusionNet model processes word embeddings generated by XLNet [Yan+19], along with manually engineered statistical features, including text-based linguistic patterns, social behavior metrics, and image-based color characteristics. The model integrates these modalities by a Bi-GRU model with attention layers followed by concatenation with statistical features for classification. A multitask loss function optimizes two objectives simultaneously: text-based feature classification and overall depression classification. The framework demonstrated superior performance, achieving an F1-score of 0.977, outperforming both traditional machine learning models (SVM, Random Forest) and deep learning models (CNN-1D, Bi-LSTM) trained on unimodal data. FusionNet also showed robustness to dataset imbalance, with the lowest intra-group F1-score variance (IFV) among tested models. Despite its strong performance, the study faces challenges such as data imbalance, feature misalignment across modalities, limited generalizability to other platforms, lack of interpretability, and reliance on static feature extraction, highlighting the need

for more diverse datasets, improved multimodal synchronization, and dynamic behavioral modeling in future research.

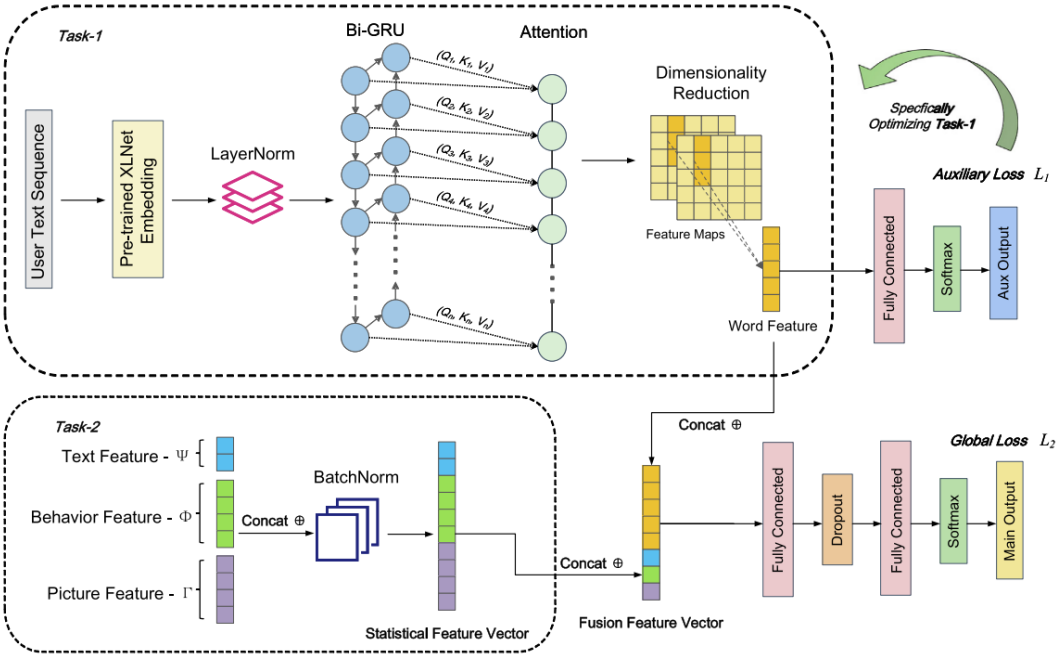



Figure 3.2.6: Proposed architecture for FusionNet [Wan+22].

In conclusion, the integration of diverse data modalities in mental health has demonstrated significant potential to enhance the accuracy and robustness of detecting and monitoring conditions such as depression and bipolar disorder. By combining audiovisual, textual, physiological and behavioral data, multimodal approaches leverage the complementary strengths of each modality to provide deeper insights into mental health states. However, challenges persist across all modalities, including data scarcity and imbalance, feature alignment and the computational complexity of integrating heterogeneous data. Addressing these issues requires larger, more diverse datasets, improved feature extraction methodologies and advanced fusion strategies. Additionally, despite advancements in integrating audio, textual, and visual data, the combined analysis of audio and biometric data remains unexplored, presenting significant opportunities for future research.

## 3.3 Anomaly Detection in Audio Data

Detecting anomalies in audio signals is a challenging task due to the complexity and variability of sound patterns, which can be influenced by environmental factors, background noise, and signal distortions. Deep learning models, particularly autoencoders, have shown promise in capturing intricate audio features and learning complex temporal patterns, making them well-suited for anomaly detection tasks.

The study by Coelho et al. [Coe+22] explored the use of deep autoencoders for unsupervised Acoustic Anomaly Detection (AAD), comparing a Dense Autoencoder (Dense AE), Convolutional Autoencoder (CNN AE), and Long Short-Term Memory Autoencoder (LSTMAE) on industrial machine monitoring and in-vehicle anomaly detection. For the industrial machine monitoring task, two public datasets were used, namely the ToyADMOS dataset [Koi+19] and the MIMII dataset [Pur+19], which contain normal and anomalous audio signals from various machines. For the in-vehicle anomaly detection task, a synthetic dataset was generated to simulate real-world conditions, including anomalous scenarios (people arguing, breaking a window, and coughing), as well as normal activities (reading a book, singing, talking, and using a smartphone). It also accounted for real-world variations such as different vehicle sizes, background noise levels (radio on/off), and environmental factors (windows open/closed, multiple passenger positions). Mel Frequency Energy Coefficients (MFECs) were extracted from the audio signals to represent the audio features, and the models were evaluated based on ROC-AUC and partial AUC (pAUC) metrics. Results demonstrated that the

proposed models achieved strong anomaly detection performance, with AUC values ranging from 72% to 91% for industrial applications and 78% to 81% for in-vehicle scenarios. The LSTMAE model exhibited the best performance in detecting in-vehicle anomalies, particularly cough detection, which was further validated in a real-world pilot demonstration, achieving up to 100% accuracy in specific conditions. Therefore, the study highlighted that deep autoencoders, particularly the LSTM AE, are effective for unsupervised AAD in industrial and in-vehicle applications; however, challenges such as computational complexity and sensitivity to real-world noise variations were noted, along with the need for further exploration of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to enhance performance.

A similar approach was taken by Bayram et al. [BDI20], who proposed a real-time AAD system for industrial processes using sequential autoencoders, specifically comparing a Convolutional Long Short-Term Memory Autoencoder (Conv-LSTMAE) with a Convolutional Autoencoder (CAE). The study utilized a custom dataset comprising industrial sounds, such as painting, cutting, welding, and robotic arm movements, along with synthetic anomaly sounds like explosions, fires, and glass breaking, mixed at various signal-to-noise ratios (SNRs). Mel-spectrograms were extracted as input features, and a sliding window approach was applied to process streaming audio in real-time. Results demonstrated that Conv-LSTMAE consistently outperformed CAE, especially in low-SNR environments, where it achieved higher anomaly detection accuracy. Glass breaking was the most distinguishable anomaly, whereas explosions and fires were more difficult to detect due to spectral overlap with industrial noise. This study validated the effectiveness of sequential autoencoders for real-time anomaly detection in manufacturing environments and suggested the integration of online learning techniques to improve adaptability to evolving acoustic patterns.

Extending the application of LSTM-based autoencoders, Mobtahej et al. [Mob+22] focused on anomaly detection in natural gas compressor systems, identifying deviations in operational signals. Their LSTM Autoencoder (LSTMAE) utilized LSTM layers to encode and decode sequential audio data, learning compact latent representations of normal operating conditions. The dataset, comprising 12,190 audio samples (8,559 normal and 3,631 anomalous), included spectral centroid and Mel-spectrogram features. The LSTM AE model achieved 100% accuracy, precision, recall, and F1 score, significantly outperforming baseline models such as GRU, LSTM, and Stacked LSTM, which were less effective at modeling high-dimensional features. Despite challenges such as class imbalance and high feature dimensionality, the model proved robust, highlighting its potential for automated anomaly detection.

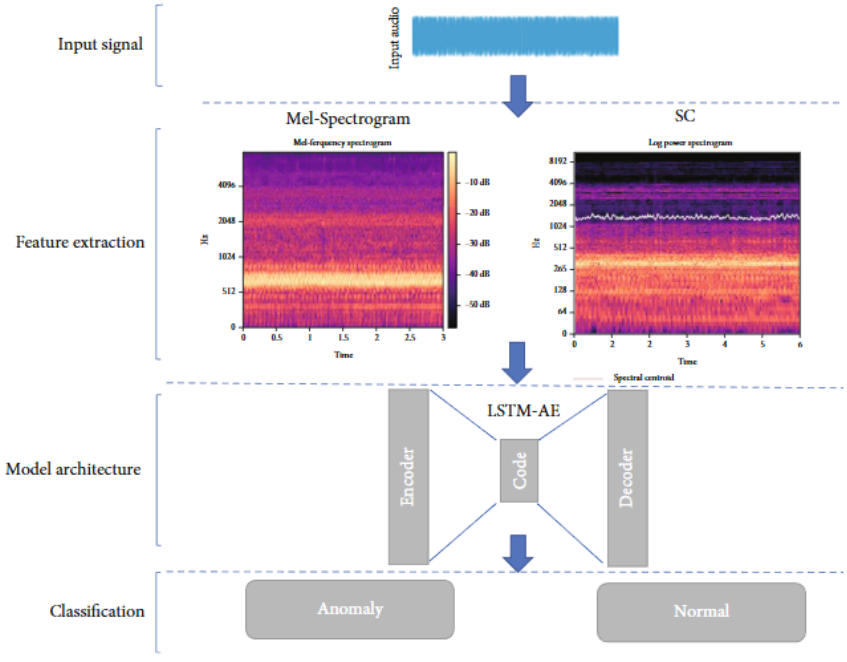



Figure 3.3.1: Proposed architecture for anomaly detection and classification using LSTM-AE [Mob+22].

In conclusion, the reviewed studies demonstrate the effectiveness of deep autoencoders, particularly LSTM-based models, for anomaly detection in audio data across various applications, highlighting their ability to capture complex temporal patterns in audio signals. While these models have been successfully applied to industrial environments, their potential could be explored in the context of detecting anomalies speech for mental health monitoring.

## 3.4 The e-Prevention Project

The e-Prevention project [Zla+22] is a long-term initiative aimed at advancing electronic health services to support the effective monitoring and relapse prevention of mental disorders, specifically bipolar disorder and schizophrenia. Spanning over three years, the project has developed an innovative integrated system to provide continuous, objective monitoring of patients' mental states. Through this system, the project aspires to identify key markers and features that correlate with mood and psychopathological changes, enabling early detection and prevention of relapses.

### The e-Prevention System

The e-Prevention system consists of three primary components. First, a non-intrusive smartwatch facilitates the long-term monitoring of biometric and behavioral indices, including heart rate variability, physical activity, and movement patterns. Second, a portable tablet installed in patients' homes records short audio-visual clips during clinician interviews, capturing social features such as speech patterns and facial expressions. Finally, all collected data is automatically stored on a secure cloud server, creating a centralized repository for analysis. This multimodal approach reflects the integration of digital phenotyping, where diverse, real-time data streams are analyzed to understand the dynamic nature of physiological and behavioral signals. The system's architecture is illustrated in Figure 3.4.1.

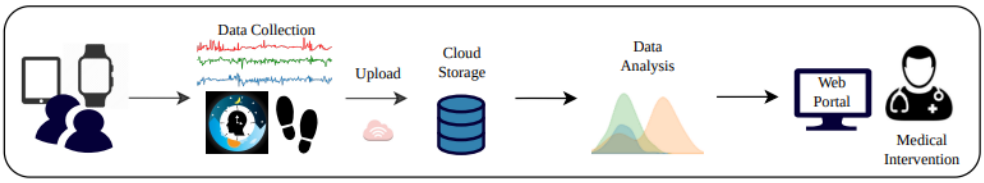



Figure 3.4.1: e-Prevention system overview [Zla+22].

### Data Collection and Statistical Analysis

The project utilized smartwatch devices to collect biometric data, including heart rate variability (HRV), accelerometer, and gyroscope readings from twenty-three (23) healthy control volunteers and twenty-four (24) patients. Statistical analyses of this data, using features mentioned in 2.2, revealed significant distinctions in physiological and behavioral markers between the patients and the healthy controls, some of which are illustrated in Figure 3.4.2. For instance, patients displayed higher variability in movement patterns during wakefulness, likely reflecting increased agitation or impulsivity. During sleep, the mean and variability of the energy (accelerometer, and gyroscope) was notably reduced in patients, which could be attributed to the sedative effects of medication. These insights formed the foundation for feature selection and machine learning models in relapse detection.

### Relapse Detection Using Autoencoder Architectures

A variety of machine learning models were implemented to evaluate their effectiveness in detecting anomalies and predicting relapse using data from both smartwatch sensors and speech recordings. For the biometric data, several autoencoder architectures were tested, including Fully Connected Neural Networks (FNNs), CNNs, GRUs, and Transformers. The results highlighted the advantages of personalized models, which were adapted to individual patients' unique data patterns. Notably, the models performed particularly well for patients experiencing moderate to severe relapses, as their physiological and behavioral markers exhibited stronger deviations compared to those with milder symptoms. Among the tested models, the CNN model

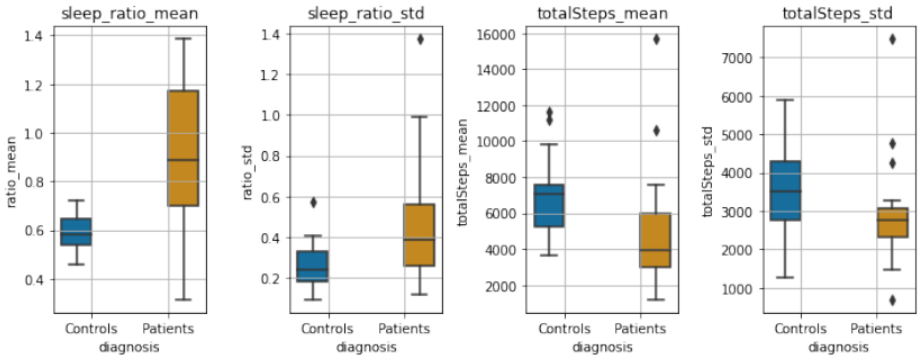



Figure 3.4.2: Example of boxplots for selected biometric features of patients and controls [Zla+22].

achieved the best performance in the personalized setting, while the FNN model performed the best in the global configuration, corresponding to a single model trained on all patients' data.

For speech data, the performance of CAE (Figure 3.4.3) and CVAE models was evaluated, focusing on their ability to extract meaningful features from mel-spectrograms using speech recordings from eight (8) patients. Similar to the biometric models, personalized approaches yielded better results overall, with CVAEs demonstrating superior performance in the global setting. This finding prompted the integration of the VAE architecture with the CNN biometric model, leading to an improvement in the global scheme and indicating decreased dependency on person-specific speech properties.

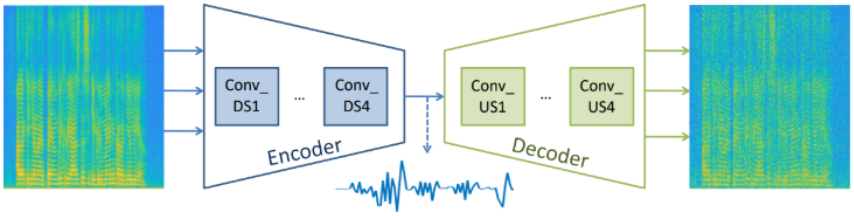



Figure 3.4.3: Overview of proposed CAE architecture for audio data [Gar+21].

Finally, the integration of results from the biometric and speech models in a single predictive framework highlighted the effectiveness of multimodal analysis. Combining physiological signals with vocal features through additive fusion improved predictive performance, demonstrating the value of leveraging complementary data sources. These results underscore the potential of multimodal approaches in accurately predicting relapses.

### Future Directions

The e-Prevention system has made significant adavances in the successful prediction of possible relapses in patients with bipolar disorder and schizophrenia. However, approaches to further enhance the system's performance could be explored. Expanding the dataset to include more patients and controls would provide a broader range of data for training and validation, potentially improving the models' generalizability. Additionally, the further refinement of multimodal fusion techniques could enhance the system's predictive capabilities by leveraging the complementary nature of biometric and speech data. As mentioned in 1.3, this thesis aims to build upon the e-Prevention project's work by implementing the aforementioned improvements and exploring other autoencoder configurations with additional data sources.

# Chapter 4

# Data and Preprocessing

## 4.1 Data Collection

### Participant Recruitment and Assessment

Participants for the e-Prevention project [Zla+22], comprising both control subjects and patients, were recruited at the University Mental Health, Neurosciences and Precision Medicine Research Institute "Costas Stefanis" (UMHRI) in Athens, Greece. The recruitment process adhered to ethical standards, with written consent obtained from all participants in accordance with the provisions of the General Data Protection Regulation (GDPR) 2016/679.

In the initial phase of the project, twenty-three (23) healthy control participants were recruited and monitored for approximately three months. In the subsequent phase, thirty-nine (39) patients diagnosed with bipolar disorder or schizophrenia spectrum disorders (SSD) were recruited.

Prior to recruitment, clinicians conducted detailed assessments of all participants, lasting approximately 180 minutes. These assessments included the collection of demographic information (age, sex, education, occupation, marital status, place of birth and residence), physical and mental health histories and neuropsychological evaluations.

### Data Collection and Monitoring

Data ware collected through wearable devices and clinical interviews. Weekly unstructured interviews, averaging 5–10 minutes each, were conducted via a dedicated web application or telephone. These interviews assessed participants' physical activity using the Greek version of the International Physical Activity Questionnaire (IPAQ-Gr) [Pap+09]. Video recordings were anonymized and securely stored on a cloud server for further analysis. For the collection of biometric data, participants wore Samsung Gear S3 Frontier smartwatches, which continuously recorded physiological and behavioral data. The mesurements included heart rate and movement activity during both wakefulness and sleep (accelerometer and gyroscope data).

### Clinical Follow-Up and Relapse Evaluation

For patients, follow-up assessments were conducted monthly to evaluate clinical status and relapse severity. These assessments included:

- **Psychopathology**: Positive and Negative Syndrome Scale (PANSS) [KFO87b].
- **Disability**: WHO Disability Assessment Schedule 2.0 (WHODAS 2.0) [Üst+10].
- **Side Effects**: Glasgow Antipsychotic Side-effect Scale (GASS) [WT08]
- **Motor Symptoms**: Abnormal Involuntary Movement Scale (AIMS) and Simpson-Angus Scale (SAS) [Guy76; SA70]
- **Additional Metrics**: Body Mass Index (BMI) and a computerized go/no-go task to assess cognitive functioning.

Relapse evaluation involved the following methods:

- Monthly assessments to identify the duration and severity of relapses (categorized as low, mid, or severe).
- Monthly administration of psychopathological scales.
- Communication with attending physicians, family members, or caregivers, as well as hospital records (if hospitalization occurred).

Relapse data, including severity and classification (e.g., psychotic or non-psychotic), were systematically documented and stored in a secure web portal for further analysis.

## 4.2 Audio Dataset

### 4.2.1 e-Prevention Audio Database

The audio database consists of audio recordings of short interviews between patients and clinicians, captured via the e-Prevention app. Initially, the database included interview data from 16 patients recorded between May 2020 and December 2021, consisting of individuals with diagnoses of schizophrenia (8), bipolar disorder I (5), schizoaffective disorder (1), schizophreniform disorder (1) and Bipolar disorder II (1). Of these participants, **8 patients** experienced a relapse during the e-Prevention study period.

Each recording was annotated based on the patient's condition at the time of the interview as follows:

- **Clean**: No relapse detected.
- **Pre-relapse**: Interviews recorded up to 28 days prior to the appearance of a relapse.
- **Relapse**: Interviews where the patient was confirmed to be relapsing.

Both relapse and pre-relapse data were considered anomalous for the purposes of anomaly detection.

### 4.2.2 e-Prevention Audio Database Expansion, Preprocessing and Feature Extraction

The initial step in our research involved expanding the e-Prevention audio database by incorporating additional interviews from both existing and new patients. This expansion included data from patients who experienced a relapse after December 2021, as well as previously excluded data from patients with isolated recordings up to that date. The same pipeline used to create the original database was applied to this expansion, encompassing the processes of audio preprocessing, speaker diarization, and feature extraction.

#### Audio Preprocessing

Since the audio recordings were coming from video files, the audio was extracted and downsampled to 16 kHz to ensure uniformity across all recordings.

#### Speaker Diarization

To separate the speech segments of the patients from those of the clinicians, we employed the the x-vector diarization pipeline provided by the Kaldi toolkit [Sny+18; Pov+11]. The x-vector diarization pipeline uses Mel-Frequency Cepstral Coefficients (MFCCs) as input to a pre-trained neural network to extract speaker-specific embeddings (x-vectors). These embeddings are grouped into separate clusters, resulting in distinct speaker segments and ensuring accurate isolation of patient speech from other sources.

#### Feature Extraction

The diarized audio segments were then processed to extract acoustic features for each speaker. For each utterance, a log mel-spectrogram (2.1.4) was computed using Librosa [McF+15], a Python library suited for audio signal processing, with the following parameters:

- Frame length: 512 samples ($\approx$ 30 ms)
- Hop length: 256 samples ($\approx$ 15 ms)
- Number of mel bands: 128

The log mel-spectrogram was then sliced into fixed-length segments of 64 frames ($\approx$ 1 second), resulting in a 128x64 feature representation for each second of speech.

After applying the pipeline mentioned above to the additional patients data, the extended database includes data from patients who experienced a relapse up to May 2022. Additionally, the expanded database containes a total of 555 interviews from 18 patients, with diagnoses of schizophrenia (9), bipolar I disorder (6), schizoaffective disorder (1), schizophreniform disorder (1) and bipolar II disorder (1), compared to the original 474

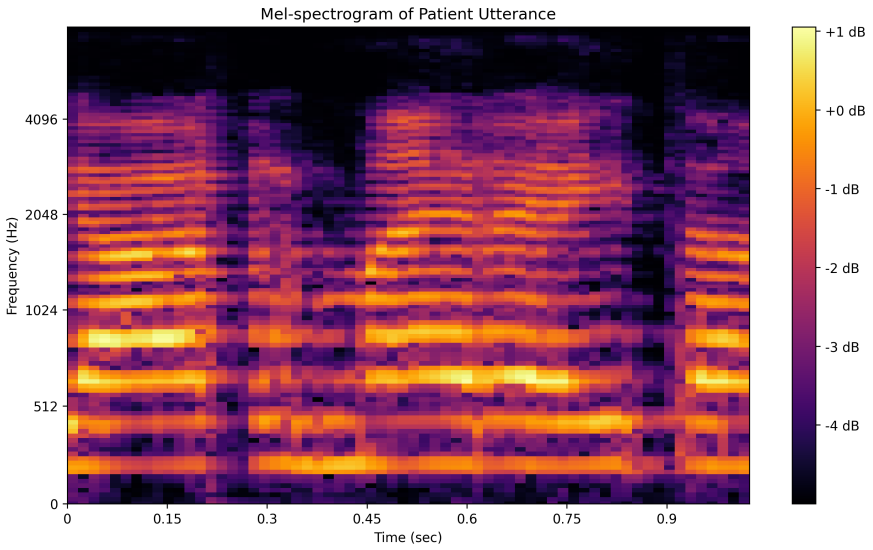



Figure 4.2.1: Example of a log mel-spectrogram computed from a patient's utterance.

interviews from 16 patients. Between the two additional patients, one experienced a relapse during the study period, while the other did not. Therefore, the extended database now includes **9 patients** who experienced a relapse during the study period, compared to the original 8 patients. This expansion also increased the total number of utterances to 16,917, with 735 minutes of diarized speech, compared to the original 14,562 utterances and 635 minutes of diarized speech. The updated database now contains 477 clean interviews, 27 pre-relapse interviews and 42 relapse interviews, in contrast to the original 396 clean, 26 pre-relapse and 36 relapse interviews.

Detailed information about patient demographics, illness details and speech data statistics for both the original and expanded datasets, involing only the relapse patients, can be found in Table 4.1.

| | **Original Database** | **Extended Database** |
|---|---|---|
| **Demographics** | | |
| Male/Female | 3/5 | 4/5 |
| Age (years) | 28.9 ± 7.7 | 28.1 ± 7.6 |
| Education (years) | 13.5 ± 1.9 | 13.3 ± 1.9 |
| Illness duration (years) | 7.0 ± 7.5 | 6.6 ± 7.2 |
| **Recorded Data** | | |
| Num. of Interviews (clean, total) | 162 | 192 |
| Num. of Interviews (clean, mean ± std) | 20.3 ± 8.1 | 21.3 ± 9.8 |
| Num. of Interviews (pre-relapse, total) | 26 | 27 |
| Num. of Interviews (pre-relapse, mean ± std) | 3.3 ± 2.1 | 3.0 ± 2.1 |
| Num. of Interviews (relapse, total) | 36 | 42 |
| Num. of Interviews (relapse, mean ± std) | 4.5 ± 2.7 | 4.7 ± 2.7 |
| Num. of Utterances (clean, total) | 5,164 | 6,459 |
| Num. of Utterances (clean, mean ± std) | 646 ± 419 | 718 ± 489 |
| Num. of Utterances (pre-relapse, total) | 909 | 934 |
| Num. of Utterances (pre-relapse, mean ± std) | 114 ± 111 | 104 ± 108 |
| Num. of Utterances (relapse, total) | 1,588 | 1,727 |
| Num. of Utterances (relapse, mean ± std) | 199 ± 204 | 192 ± 219 |

Table 4.1: Comparison of demographics, illness information, and recorded speech data statistics for relapse patients in the original and extended e-Prevention databases.

**Addressing Challenges and Inconsistencies**

During the expansion process, some challenges and inconsistencies were encountered, which required manual intervention to ensure the consistency and accuracy of the database. These included:

- Ensuring that date directories for each patient followed a consistent naming pattern to avoid misalignment during data processing.
- Addressing inconsistencies between date directories and the actual audio files, ensuring that all audio files were correctly labeled.
- Verifying and correcting annotations for pre-relapse and relapse dates to ensure they aligned with the actual relapse periods.
- Addressing cases where patient audio data did not align with the corresponding patient ID.
- While the Kaldi diarization pipeline automated much of the segmentation process, a subset of diarized utterances required manual adjustments to isolate patient speech accurately.

For all experiments involving audio data in this thesis, we use the dataset that includes the 9 patients who experienced a relapse during the study period, with the goal of detecting anomalies in their speech patterns during pre-relapse and relapse periods.

## 4.3 Biometric Dataset

### 4.3.1 e-Prevention Biometric Database

The biometric database included physiological data from twenty-four (24) patients with diagnoses of schizophrenia (12), bipolar I disorder (8), schizoaffective disorder (2), bipolar II disorder (1) and schizophreniform disorder (1), recorded with Samsung Gear S3 Frontier smartwatches due to their capability to record and send data from accelerator, gyroscope and heart rate sensors. During the study, a smartwatch application was developed to collect raw data from the smartwatches and transmit it to a cloud server for further analysis [Mag+20].

For the purposes of our experiments, the goal of using biometric data was to explore the fusion of audio and biometric signals to enhance the prediction of relapse. Therefore, a subset of the biometric dataset was used, which consisted of data from the same 9 patients included in the extended audio database collected over the same period.

### 4.3.2 Audio and Biometric Data Alignment

The first step in the biometric data processing pipeline was to accurately align the biometric data dates with those of the audio interviews, to ensure reliable results in the multimodal experiments. It is worth noting that patients for whom less than 10 interview days of biometric data were available, were excluded from the multimodal experiments.

Initially, the alignment strategy focused only on using biometric data collected on the exact day of the audio interviews. However, due to data scarcity, we resorted to different alignment strategies, where the biometric database was expanded with data within extended temporal windows surrounding the interview dates. Specifically, data collected within 3-day, 5-day, and 7-day windows centered on the interview dates were incorporated into the experiments. From now on, we will refer to these datasets as the "day-of", "3-day", "5-day", and "7-day" datasets, respectively.

The Table 4.2 provides an overview of the raw biometric data recorded the same day as the audio interviews for all 9 patients in the expanded audio relapse dataset, mentioned in 4.2.2. The data includes the number of days recorded, the number of hours recorded and the number of 5-minute intervals recorded for both the "day-of" audio interviews and the extended 7-day windows around the interviews.

| | Day-of Interview | 7-day Window |
|---|---|---|
| **Recording Statistics** | | |
| Num. of Days Recorded (total) | 275 | 475 |
| Num. of Days Recorded (mean ± std) | 15.3 ± 7.5 | 26.4 ± 10.4 |
| Num. of Hours Recorded (total) | 3,937 | 6,860 |
| Num. of Hours Recorded (mean ± std) | 218.8 ± 115 | 381.1 ± 171.8 |

Table 4.2: Comparison of raw recorded biometric data for all patients in the multimodal experiments between day-of interview and 7-day window around the interview.

### 4.3.3 Biometric Data Preprocessing and Feature Extraction

The raw data from each sensor was stored in parquet files and then converted into Pandas dataframes, with each file corresponding to a specific day. The gyroscope and accelerometer sensors sampled data at 20 Hz, while the heart rate sensor recorded at 5 Hz. The initial preprocessing procedure involved the following steps:

- **Data Availability**: Each daily dataframe was examined to ensure that it contained sufficient hours of data. Days with less than 4 hours of recorded data were dropped entirely, as they were considered too sparse for reliable analysis.
- **Data Segmentation**: To maximize the use of available data, each daily dataframe was divided into 8-hour segments. This segmentation was chosen for its balance between capturing meaningful physiological patterns and accommodating the variability in daily recording durations. Within each 8-hour segment, the recorded data was reassessed to confirm it contains at least 4 hours of valid data. Segments failing this threshold were dropped. This segmentation ensured that even on days with partial recordings, at least one segment per day would meet the minimum data standards.
- **Invalid Sample Removal**: Invalid samples were identified and removed based on predefined thresholds for each biometric sensor. For the accelerometer, limits for each axis' $(x, y, z)$ values were constrained to -19.6 to 19.6, while the gyroscope's limits were set to -573 to 573. For the heart rate sensor, negative values were considered invalid and removed.

#### Feature Extraction

The feature extraction included time-domain, frequency-domain, and non-linear features from the raw data, also used in the e-Prevention study and detailed in 2.2. These features were extracted for 5-minute intervals within each 8-hour segment (96 5-minute intervals). The choice of the interval was motivated by a previous study showing the effectiveness of 5-minute windows in capturing meaningful short-term patterns [Ret+20]. The extracted features included:

- **Gyroscope and accelerometer features**: Short-time energy (STE) of the signals, calculated as the norm of the 3-axis signals within a 5-minute window.
- **Heart Rate Variability (HRV) features**:
  - Time-domain features: Mean heart rate and mean of the RR intervals.
  - Frequency-domain features: Lomb-Scargle periodogram of the RR intervals, capturing the low-frequency (LF: 0.04-0.15 Hz) and high-frequency (HF: 0.15-0.4 Hz) components.
  - Non-linear features: Poincaré plot features, including the standard deviation of the points perpendicular to the line of identity (SD1).
  - The number of valid samples within the 5-minute interval.
- **Additional features**: Sinusoidal representation of the seconds within the 5-minute interval, capturing the chronological patterns of the data.

After feature extraction, any missing values were imputed using the median of the feature across the entire dataset.

In total, 10 features were extracted for each 5-minute interval, resulting in a feature tensor of size 96x10 for each 8-hour segment. An example of the 10 extracted features for a single 8-hour segment of a patient's biometric data is shown in Figure 4.3.1.

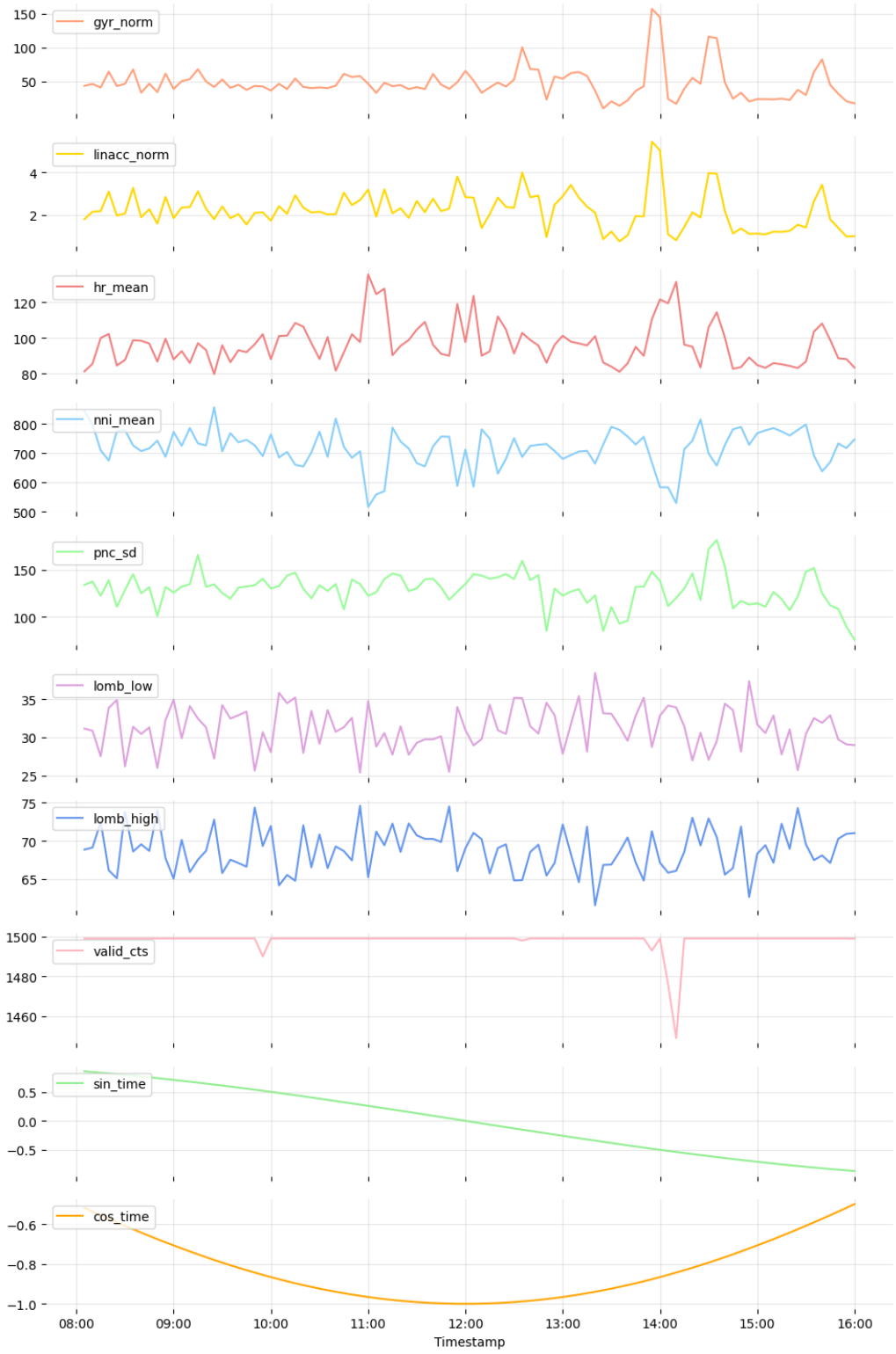



Figure 4.3.1: Example of extracted features for a single 8-hour segment of a patient's biometric data.

In Table 4.3, we provide an overview of the demographics, illness information, and recorded biometric data for the patients in the day-of, 3-day, 5-day, and 7-day datasets after preprocessing and feature extraction. All patients included in these are datasets also included in the expanded audio relapse dataset, as mentioned before. The information includes the number of days recorded, the number of hours recorded and the number of 5-minute intervals recorded for each dataset.

| **Datasets** | **Day-of** | **3-day** | **5-day** | **7-day** |
|---|---|---|---|---|
| **Demographics** | | | | |
| Male/Female | 2/2 | 2/3 | 2/3 | 3/4 |
| Age (years) | 31 ± 8.7 | 30.2 ± 8 | 30.2 ± 8 | 28 ± 7.6 |
| Education (years) | 14 ± 2 | 14.4 ± 2 | 14.4 ± 2 | 13.7 ± 2 |
| Illness duration (years) | 8.8 ± 9 | 7.2 ± 8.6 | 7.2 ± 8.6 | 6.4 ± 7.4 |
| **Recorded Data** | | | | |
| Num. of Days Recorded (total) | 66 | 102 | 124 | 158 |
| Num. of Days Recorded (mean ± std) | 16.5 ± 5 | 20.4 ± 7.5 | 24.8 ± 7.2 | 22.6 ± 10.8 |
| Num. of Hours Recorded (total) | 888 | 1,752 | 2,400 | 3,280 |
| Num. of Hours Recorded (mean ± std) | 222 ± 45.7 | 350.4 ± 89.6 | 480 ± 96.7 | 468.6 ± 185 |
| Num. of 5-min intervals (total) | 10,656 | 21,024 | 28,800 | 39,360 |
| Num. of 5-min intervals (mean ± std) | 2,664 ± 548.9 | 4,204.8 ± 1,074.9 | 5,760 ± 1171 | 5,622.9 ± 2,219.5 |

Table 4.3: Comparison of demographics, illness information, and recorded biometric data for the patients in the day-of, 3-day, 5-day, and 7-day datasets after preprocessing and feature extraction.

### Addressing Challenges and Inconsistencies

Similarily to the audio data, the biometric data presented several challenges and inconsistencies that were identified and addressed to ensure the accuracy of the dataset. These included:

- Ensuring consistency between the date names in the file names and the timestamps within the parquet files to avoid misalignment during data processing.
- Discarding dates that did not contain all three types of sensor data to ensure that all features could be extracted.
- Correcting cases where biometric data was incorrectly assigned to the wrong patient ID by cross-referencing patient metadata and reassigning the data to the appropriate patient.

# Chapter 5

# Audio Experiments

The first set of experiments in this thesis focused on a comparative analysis of autoencoder models for detecting relapse in mental health patients based on spontaneous speech. The primary objective was to evaluate the performance of Convolutional Autoencoder (CAE) and Convolutional Variational Autoencoder (CVAE) models, originally developed during the e-Prevention project, on the newly expanded audio database. As mentioned in 4.2.2, this updated dataset now includes 477 clean interviews, 27 pre-relapse interviews and 42 relapse interviews from 9 patients, compared to the original 396 clean, 26 pre-relapse and 36 relapse interviews from 8 patients, providing a broader basis for analysis.

In addition to re-assessing the CAE and CVAE models, another key objective was to develop and evaluate LSTM-based autoencoders, specifically an LSTM Autoencoder (LSTMAE) and a LSTM Variational Autoencoder (LSTMVAE). The goal was to determine whether these sequence-based architectures, which are designed to capture temporal dependencies in speech, offer improved anomaly detection capabilities compared to the convolutional models. Through this comparative study, we aim to identify the most effective autoencoder-based approach for relapse detection in speech, forming the basis for subsequent multimodal experiments.

## 5.1 Methodology

### 5.1.1 Data Normalization: Per-Patient and Global

To assess the effectiveness of autoencoder models in detecting relapse, two types of experiments were conducted: Personalized patient experiments and multi-patient global experiments. Each experimental setup was designed to evaluate the models under different generalization conditions, providing insights into both patient-specific relapse prediction and relapse detection across multiple patients.

#### Personalized Experiments

In the first experimental setup, each model was trained on the clean data of a single patient and later used to predict relapse exclusively for that patient. This approach allows the model to learn patient-specific speech patterns, capturing personalized variations that may indicate a relapse.

#### Global Experiments

The second set of experiments involved training models on the combined clean data from all 9 patients in the dataset. This approach evaluates whether a model trained on diverse patient data can generalize across different individuals and accurately identify relapses. Within this multi-patient training framework, two normalization strategies were applied:

- **Per-Patient Normalization:** Each patient's data ware normalized independently, using statistics (mean and standard deviation) computed from their respective clean training data.
- **Global Normalization:** All patient data ware normalized collectively, using statistics computed from the entire training set, which includes clean data from all patients.

By comparing these experimental setups, we aim to determine whether patient-specific models provide better relapse detection compared to models trained on multiple patients and to assess the impact of normalization strategies on model performance.

### 5.1.2 Autoencoder Architectures

#### Convolutional Autoencoder (CAE)

The CAE developed in [Gar+21; Zla+22] follows a deep convolutional neural network structure, composed of encoding layers that progressively compress the input 128×64 mel-spectrogram into a lower-dimensional latent space representation, followed by decoding layers that attempt to reconstruct the original spectrogram.

The encoder is composed of 4 sequential convolutional downsampling (DS) blocks, each consisting of a 2D-Convolution layer with a ReLU activation function, followed by a Max Pooling layer that progressively reduces the spatial dimensions of the input spectrogram. Conversely, the decoder comprises 4 convolutional upsampling (US) blocks, where each block begins with an Upsampling layer, followed by a 2D-Convolution

layer with a ReLU activation function, incrementally reconstructing the original mel-spectrogram. The final reconstructed mel-spectrogram is produced through a single-channel 2D-Convolution layer, aiming to match the input dimensions.

In Table 5.1, we present the architecture parameters of the CAE model that was used in the audio experiments. Each row in the Table corresponds to a convolutional block, downsampling (DS) or upsampling (US), and includes the number of filters, kernel size, pooling or upsampling size, and the output dimensions of the block. Additionally, the Figure 3.4.3 illustrates the architecture of the CAE model used in [Gar+21].

| Conv. Block | Filters | Kernel Size | Pooling Size | Upsampling Size | Output Dimensions |
|---|---|---|---|---|---|
| DS1 | 32 | (5,5) | (2,2) | - | 64×32×32 |
| DS2 | 64 | (5,5) | (4,2) | - | 16×16×64 |
| DS3 | 128 | (5,5) | (4,4) | - | 4×4×128 |
| DS4 | 256 | (5,5) | (4,4) | - | 1×1×256 |
| US1 | 128 | (5,5) | - | (4,4) | 4×4×128 |
| US2 | 64 | (5,5) | - | (4,4) | 16×16×64 |
| US3 | 32 | (5,5) | - | (4,2) | 64×32×32 |
| US4 | 1 | (5,5) | - | (2,2) | 128×64×1 |

Table 5.1: Architecture parameters of the Convolutional Autoencoder (CAE) used in the audio experiments.

### Convolutional Variational Autoencoder (CVAE)

The CVAE model, shown in Figure 5.1.1, extends the aforementioned CAE architecture by introducing a probabilistic latent space representation. The CVAE follows, similar to the CAE, a convolutional encoder-decoder structure, but introduces a latent space modeled as a Gaussian distribution. The encoder outputs the parameters of the latent distribution, mean $\mu$ and log variance $\log \sigma^2$, which are used to sample latent representations using the reparameterization trick, which is defined as follows:

$$z = \mu + \sigma \cdot \epsilon \tag{5.1.1}$$

Where

- $\sigma = \exp(0.5 \log \sigma^2)$ is the standard deviation of the latent distribution.
- $\epsilon \sim \mathcal{N}(0, 1)$ is random noise from a standard normal distribution.

The decoder then reconstructs the input from the sampled latent representation. The architecture parameters for the downsampling and upsampling blocks of the CVAE model are identical to the CAE model and shown in Table 5.1.

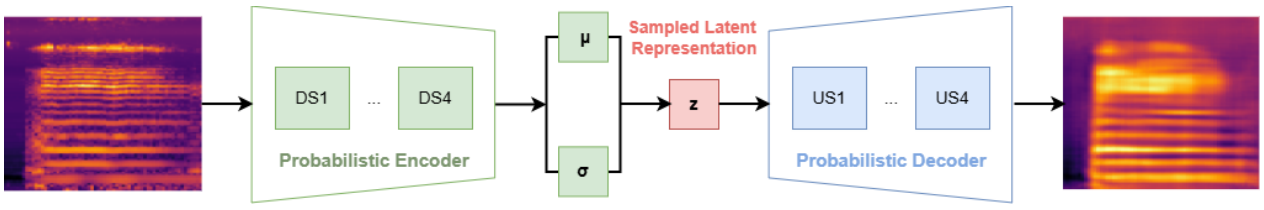



Figure 5.1.1: Overview of the proposed Convolutional Variational Autoencoder (CVAE) architecture used in the audio experiments.

### Long Short-Term Memory Autoencoder (LSTMAE)

The development of the LSTMAE was inspired from the study by Mobtahej et al. [Mob+22] that demonstrated outstanding results in anomaly detection from audio signals using mel-spectrograms, as mentioned in 3.3. Given the effectiveness of LSTM-based architectures in modeling temporal dependencies, this study highlighted their potential for capturing sequential patterns in audio samples that might be indicative of

relapse in mental health patients. Building on these findings, the next step in our research was to experiment with this architecture and assess its performance compared to the CAE.

The LSTMAE follows a sequence-to-sequence architecture consisting of an encoder and a decoder. The input is the same 128×64 mel-spectrogram as the CAE model, but for compatibility with the LSTM layers, it is reshaped to 64×128, representing 64 time steps of 128 frequency bins. The encoder processes the input mel-spectrogram sequences using an LSTM layer, which captures temporal dependencies in speech data, followed by layer normalization, Leaky ReLU activation, and dropout for regularization. The encoded sequence is then flattened and mapped to a low-dimensional latent space through a dense layer. The decoder reconstructs the original input by first expanding the latent representation using a dense layer, reshaping it into a sequence format, and then applying an LSTM layer with layer normalization and dropout to generate the final spectrogram reconstruction. The parameters of the LSTMAE architecture are presented in Table 5.2, and the model's architecture is illustrated in Figure 5.1.2.

| Layer Type | Units | Add. Parameters | Output Dimensions |
|---|---|---|---|
| LSTM | 64 | `return_sequences=True` | 64×64 |
| Flatten | - | - | 4096 |
| Dense | 64 | - | 64 |
| Dense | 4096 | - | 4096 |
| Reshape | - | - | 64×64 |
| LSTM | 64 | `return_sequences=True` | 64×64 |
| Time Distributed Dense | 128 | - | 64×128 |

Table 5.2: Architecture parameters of the Long Short-Term Memory Autoencoder (LSTMAE) used in the audio experiments.

To ensure optimal performance, the LSTMAE was fine-tuned using Python's GridSearchCV from the scikit-learn library [Ped+11]. This method systematically searches through a predefined set of hyperparameter values to identify the best-performing model configuration. The primary hyperparameters optimized include the number of LSTM units, latent dimension size, dropout rate, and the optimizer's learning rate (which will be detailed in the next section on the model's training methodology). The search was conducted over the following parameter ranges: LSTM units {32,64,128,256} and latent dimension {32,64,128,256}.

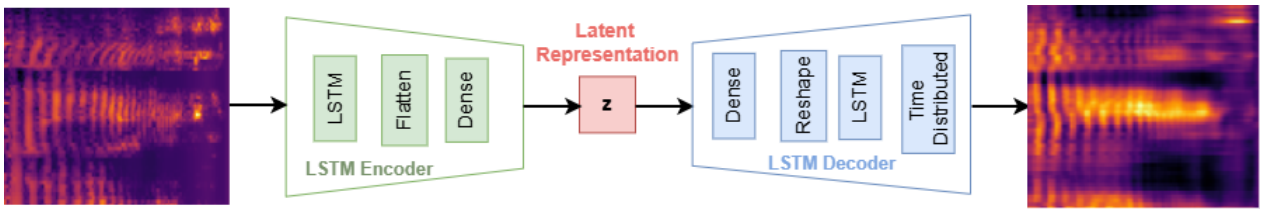



Figure 5.1.2: Overview of the proposed Long Short-Term Memory Autoencoder (LSTMAE) architecture used in the audio experiments.

### Long Short-Term Memory Variational Autoencoder (LSTMVAE)

The LSTMVAE follows the same fundamental logic as the CVAE, incorporating a probabilistic latent space. Similar to the CVAE, the LSTMVAE learns a distribution over the latent space rather than a fixed representation. The model retains the sequence-to-sequence architecture of the LSTMAE, with identical architectural parameters, including the number of LSTM units, latent dimension size and overall types of layers as shown in Table 5.2. The difference lies in the latent space, where the LSTMVAE introduces dense layers to compute the mean $\mu$ and log variance $\log \sigma^2$ of the latent distribution. The model then samples latent representations using the reparameterization trick (5.1.1) exactly as in the CVAE. The LSTMVAE architecture is illustrated in Figure 5.1.3.

All models were implemented using the TensorFlow [Aba+16] and Keras [Cho+15] libraries in Python.

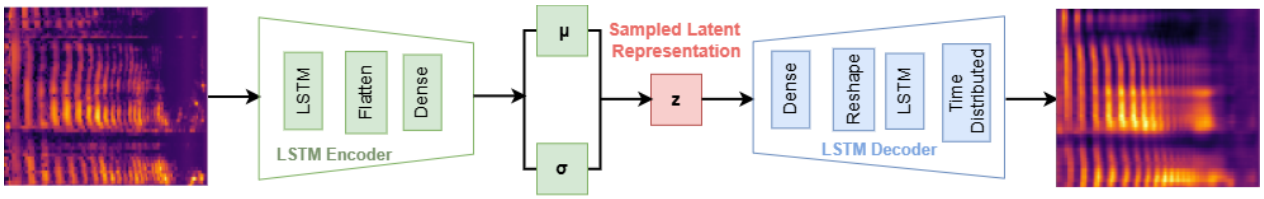



Figure 5.1.3: Overview of the proposed Long Short-Term Memory Variational Autoencoder (LSTMVAE) architecture used in the audio experiments.

### 5.1.3 Model Training

All models followed a common training pipeline, established in [Zla+22], with parameter modifications adjusted to each model's architecture to enhance performance. The training involved 5-fold cross-validation, where models were trained exclusively on clean data, while evaluation was conducted on both clean and anomalous (pre-relapse and relapse) data, allowing the models to learn a baseline representation of normal speech patterns while testing their ability to identify anomalies associated with relapse. During each cross-validation fold, the clean speech data were divided into three sets: 60% for training, 20% for validation, and 20% for testing. To prevent session-wise overfitting, the splits were structured so that all mel-spectrograms from the same interview remained within the same fold.

#### Loss Functions

The loss function used for training the CAE and LSTMAE models was the Mean Squared Error (MSE) loss, which measures the squared difference between the input mel-spectrogram and the model's estimate.

For the CVAE and LSTMVAE models, the loss function was formulated as the weighted sum of the MSE (2.3.1), applied at the output of the autoencoder, and the Kullback-Leibler (KL) divergence (2.5.4), which measures the difference between the latent distribution and a standard normal distribution. During training, the loss weights for the CVAE model were set at 0.01 for the KL divergence and 1 for the MSE, while the LSTMVAE model used a weight of 0.05 for the KL divergence and 1 for the MSE, to balance reconstruction accuracy and latent space regularization.

#### Hyperparameters and Optimization

The common hyperparameters, shown in Table 5.3, include a maximum training duration of 200 epochs, a batch size of 8, and the Adam optimizer. To prevent overfitting, early stopping was applied with a patience of 10 epochs, ensuring that training was terminated if the validation loss did not improve for 10 consecutive epochs.

| | **Epochs** | **Batch Size** | **Optimizer** | **Patience** |
|---|---|---|---|---|
| All Models | 200 | 8 | Adam | 10 |

Table 5.3: Common training hyperparameters for all audio models.

The model-specific hyperparameters, presented in Table 5.4, primarily differ in learning rate and dropout rate. The CAE and CVAE models, were trained with a learning rate of $3 \times 10^{-4}$ and did not require dropout regularization.

For the LSTMAE and LSTMVAE models, as mentioned in 5.1.2, the hyperparameters were optimized using GridSearchCV. The search was conducted over the following parameter ranges: learning rate $\{1 \times 10^{-4}$, $3 \times 10^{-4}$, $1 \times 10^{-3}\}$ and dropout rate $\{0.1,0.2,0.3,0.5\}$. The final model configuration, presented in Table 5.2, included a learning rate of $1 \times 10^{-3}$ and a dropout rate of 0.2, which was applied to the LSTM layers for regularization.

| Model | Learning Rate | Dropout |
|---|---|---|
| CAE | $3 \times 10^{-4}$ | - |
| CVAE | $3 \times 10^{-4}$ | - |
| LSTMAE | $1 \times 10^{-3}$ | 0.2 |
| LSTMVAE | $1 \times 10^{-3}$ | 0.2 |

Table 5.4: Model-specific training hyperparameters for all audio models.

### 5.1.4 Evaluation Methods and Metrics

The performance of all models was evaluated on a per-session basis, meaning that each test session was identified, and its corresponding mel-spectrograms were normalized depending on the experimental setup. As mentioned before, the testing set included clean, pre-relapse, and relapse data while the models were trained exclusively on clean data. Each mel-spectrogram within a session produced an anomaly score, which was then aggregated across time to obtain a single anomaly score for the entire session.

The method of computing the mel-spectrogram anomaly score varied based on the model architecture:

- For the CAE and LSTMAE models, the MSE was calculated between the input mel-spectrogram and the model's reconstruction.
- For the CVAE and LSTMVAE models, we experimented with both using the MSE (computed at the output of the autoencoder, i.e. the reconstructed mel-spectrograms) and the KL divergence to the $\mathcal{N}(0, 1)$ distribution at the learned latent space.

The overall model performance was evaluated using the following metrics:

- **Anomaly Score:** The anomaly score (MSE, KL divergence) was computed across all sessions for each state (clean, pre-relapse and relapse). This metric quantifies the model's ability to detect anomalies, with higher scores in pre-relapse and relapse states than in the clean state indicating successful anomaly detection.
- **Area Under the Receiver Operating Characteristic Curve (ROC-AUC) Score:** The ROC-AUC score (explained in 2.3.2) measures how well the models separate sessions that contain clean data from those associated with pre-relapse and relapse states, based on the per-session computed anomaly scores.

For the final results, the median anomaly scores and mean ROC-AUC scores were averaged across the 5 cross-validation folds, corresponding with the e-Prevention project's methodology. The evaluation was conducted separately for each model architecture and experimental setup, comparing the personalized patient experiments with the multi-patient global experiments and the per-patient normalization with the global normalization strategies.

## 5.2 Results

### 5.2.1 CAE Model Results on the Expanded Database

This section presents the evaluation of the CAE model on both the expanded relapse dataset and the original dataset for comparison. Notably, the results for the original dataset correspond to those reported in the e-Prevention paper [Zla+22]. Additionally, it is important to highlight that patient #8 is the new addition in the expanded dataset. The goals of this evaluation include the following for both experimental setups mentioned in 5.1.1:

- **Personalized Experiments:** Assess the impact of additional patient data on personalized model performance, specifically whether increasing the dataset for each patient improves the model's ability to capture individual characteristics and detect anomalies.

- **Global Experiments:** Determine whether adding more patients and utterances improves or disrupts global anomaly detection, specifically investigating whether increased data diversity helps the CAE generalize or leads to higher reconstruction errors due to patient variability.

Beyond model comparisons, a key goal of this experiment is to determine whether the CAE can effectively differentiate clean speech from anomalous states (pre-relapse and relapse) using the additional patient data.

### Personalized Experiments

Table 5.5 presents the median MSE anomaly scores across five cross-validation folds for the clean and anomalous states, with bold values indicating cases where reconstruction error increases in pre-relapse and relapse states. The bottom row summarizes median values for both datasets, highlighting performance changes after adding Patient #8. Table 5.6 presents the mean ROC-AUC scores across five cross-validation folds, with bold values marking patients for whom the CAE model performed better after dataset expansion, suggesting that additional data enhanced anomaly detection for those individuals. This annotation will be consistently applied in all subsequent tables.

| Patient ID | Original Dataset | | | Expanded Dataset | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| #1 | 0.292 ± 0.035 | **0.321** ± 0.067 | **0.313** ± 0.065 | 0.267 ± 0.020 | **0.386** ± 0.067 | **0.337** ± 0.065 |
| #2 | 0.452 ± 0.065 | **0.577** ± 0.000 | **0.464** ± 0.064 | 0.429 ± 0.068 | **0.578** ± 0.024 | **0.461** ± 0.034 |
| #3 | 0.417 ± 0.068 | **0.456** ± 0.065 | **0.504** ± 0.061 | 0.408 ± 0.074 | **0.470** ± 0.025 | **0.551** ± 0.034 |
| #4 | 0.273 ± 0.034 | **0.359** ± 0.070 | **0.295** ± 0.027 | 0.248 ± 0.025 | **0.366** ± 0.017 | **0.280** ± 0.009 |
| #5 | 0.308 ± 0.029 | **0.389** ± 0.010 | 0.286 ± 0.000 | 0.277 ± 0.037 | **0.331** ± 0.034 | 0.267 ± 0.022 |
| #6 | 0.649 ± 0.104 | 0.646 ± 0.130 | 0.599 ± 0.146 | 0.571 ± 0.080 | **0.598** ± 0.019 | **0.602** ± 0.030 |
| #7 | 0.320 ± 0.018 | **0.386** ± 0.048 | **0.380** ± 0.000 | 0.290 ± 0.010 | **0.338** ± 0.014 | **0.345** ± 0.010 |
| #8 | - | - | - | 0.322 ± 0.039 | **0.421** ± 0.010 | **0.333** ± 0.013 |
| #9 | 0.520 ± 0.051 | **0.573** ± 0.000 | **0.665** ± 0.000 | 0.458 ± 0.061 | **0.593** ± 0.034 | **0.638** ± 0.041 |
| **Median (8)** | 0.369 ± 0.123 | **0.423** ± 0.112 | **0.422** ± 0.135 | 0.349 ± 0.108 | **0.428** ± 0.102 | **0.399** ± 0.139 |
| **Median (9)** | - | - | - | 0.322 ± 0.105 | **0.421** ± 0.102 | **0.337** ± 0.139 |

Table 5.5: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the personalized CAE models on the original and expanded datasets.

| Patient ID | ROC-AUC | |
|---|---|---|
| | Original Dataset | Expanded Dataset |
| #1 | 0.549 ± 0.140 | **0.653** ± 0.048 |
| #2 | 0.465 ± 0.133 | **0.468** ± 0.159 |
| #3 | 0.718 ± 0.171 | **0.722** ± 0.165 |
| #4 | 0.665 ± 0.064 | 0.650 ± 0.093 |
| #5 | 0.780 ± 0.063 | 0.754 ± 0.082 |
| #6 | 0.489 ± 0.148 | **0.500** ± 0.159 |
| #7 | 0.883 ± 0.082 | **0.905** ± 0.055 |
| #8 | - | 0.483 ± 0.187 |
| #9 | 0.790 ± 0.245 | **0.817** ± 0.186 |
| **Mean (8)** | 0.667 ± 0.143 | **0.684** ± 0.139 |
| **Mean (9)** | - | 0.661 ± 0.146 |

Table 5.6: Comparison of (MSE) ROC-AUC values for the personalized CAE models on the original and expanded datasets.

From Table 5.5, we can observe that after the dataset expansion there is a general decrease in the MSE anomaly scores for clean states, indicating that the CAE model is better at reconstructing normal speech patterns. In contrast, the MSE anomaly scores for pre-relapse and relapse show an increase for most patients, indicating improved ability in detecting anomalous speech patterns related to relapse. This improvement in detecting anomalous states is further validated by the results in Table 5.6, which shows an overall increase in ROC-AUC scores for most patients after dataset expansion. The mean ROC-AUC for the 8 original

patients improves from 0.667 to 0.684, confirming that the expanded dataset enhances the model's ability to differentiate between normal and relapse states. Notably, some patients, such as #1, #7 and #9, exhibit a significant increase in ROC-AUC scores, with patient #7 having a ROC-AUC of 0.905 after expansion. However, for a few patients (#2 and #4), the ROC-AUC remains nearly the same, suggesting that while expansion benefits most cases, its impact may depend on individual speech patterns or the amount of added data. Overall, these results indicate that dataset expansion did improve the CAE model's ability to detect relapse-related anomalies.

#### Global Experiments

Tables 5.7 and 5.8 present the median MSE anomaly scores and mean ROC-AUC scores for the global CAE models on the original and expanded datasets. The results are provided for both per-patient and global normalization schemes, allowing for a comparison of model performance across different normalization strategies.

| Norm. | Original Dataset | | | Expanded Dataset | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **Per-Patient** | 0.341 ± 0.074 | **0.388** ± 0.142 | 0.339 ± 0.094 | 0.188 ± 0.008 | **0.236** ± 0.013 | **0.206** ± 0.007 |
| **Global** | 0.280 ± 0.055 | **0.292** ± 0.082 | 0.255 ± 0.064 | 0.199 ± 0.009 | **0.232** ± 0.007 | **0.204** ± 0.007 |

Table 5.7: Comparison of anomaly scores for per-patient and global normalization schemes for clean (C), pre-relapse (P), and relapse (R) states for the global CAE models on the original and expanded datasets.

| Norm. | ROC-AUC | |
|---|---|---|
| | Original Dataset | Expanded Dataset |
| **Per-Patient** | 0.531 ± 0.021 | **0.618** ± 0.023 |
| **Global** | 0.525 ± 0.024 | **0.633** ± 0.033 |

Table 5.8: Comparison of ROC-AUC scores for per-patient and global normalization schemes for the global CAE models on the original and expanded datasets.

From Tables 5.7 and 5.8, we can observe that the global CAE model performed significantly better after dataset expansion, with a consistent improvement in MSE anomaly scores for pre-relapse and relapse states. The ROC-AUC scores also show a significant increase from 0.531 to 0.618 and 0.525 to 0.633 for the per-patient and global models, respectively. This improvement suggests that the additional patient data helped the global model generalize better across patients and detect anomalies.

Therefore, both personalized and global experiments demonstrate that the CAE model benefits from the expanded dataset, with improved anomaly detection for most patients and normalization schemes.

### 5.2.2 CVAE Model Results on the Expanded Database

In this section, we evaluate the performance of the CVAE model using the same experimental setup as the CAE model. The evaluation metrics include both the values of the reconstruction MSE and the KL divergence as anomaly scores, and their corresponding ROC-AUC scores, aiming to assess how the inclusion of additional patient data affects the personalized and global CVAE models' performance as well their ability to detect relapse-related anomalies. By incorporating KL divergence and its associated ROC-AUC scores, we further explore the model's ability to capture meaningful latent space representations, complementing the MSE results.

#### Personalized Experiments

Tables 5.9, 5.10, 5.11, and 5.12 present the median MSE anomaly scores, mean MSE ROC-AUC scores, median KL divergence anomaly scores and mean KL ROC-AUC scores for the personalized CVAE models on the original and expanded datasets.

| Patient ID | Original Dataset | | | Expanded Dataset | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| #1 | 0.520 ± 0.040 | **0.544** ± 0.073 | **0.535** ± 0.088 | 0.523 ± 0.033 | **0.638** ± 0.022 | **0.581** ± 0.023 |
| #2 | 0.703 ± 0.074 | **0.880** ± 0.000 | 0.662 ± 0.063 | 0.653 ± 0.131 | **0.889** ± 0.036 | **0.659** ± 0.024 |
| #3 | 0.689 ± 0.101 | **0.769** ± 0.068 | **0.879** ± 0.043 | 0.685 ± 0.121 | **0.752** ± 0.065 | **0.865** ± 0.066 |
| #4 | 0.606 ± 0.065 | **0.729** ± 0.102 | **0.622** ± 0.046 | 0.599 ± 0.038 | **0.734** ± 0.024 | **0.603** ± 0.020 |
| #5 | 0.595 ± 0.044 | **0.706** ± 0.019 | 0.569 ± 0.000 | 0.569 ± 0.030 | **0.649** ± 0.024 | 0.549 ± 0.020 |
| #6 | 0.887 ± 0.091 | **0.896** ± 0.190 | 0.835 ± 0.181 | 0.846 ± 0.131 | **0.867** ± 0.040 | **0.874** ± 0.042 |
| #7 | 0.573 ± 0.033 | **0.669** ± 0.035 | **0.632** ± 0.000 | 0.504 ± 0.030 | **0.632** ± 0.031 | **0.585** ± 0.015 |
| #8 | - | - | - | 0.617 ± 0.109 | **0.631** ± 0.045 | 0.567 ± 0.047 |
| #9 | 0.776 ± 0.079 | **0.815** ± 0.000 | **0.981** ± 0.000 | 0.733 ± 0.132 | **0.809** ± 0.044 | **0.967** ± 0.065 |
| **Median (8)** | 0.648 ± 0.112 | **0.749** ± 0.108 | 0.647 ± 0.152 | 0.626 ± 0.107 | **0.743** ± 0.096 | **0.631** ± 0.154 |
| **Median (9)** | - | - | - | 0.617 ± 0.101 | **0.734** ± 0.097 | 0.603 ± 0.152 |

Table 5.9: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the personalized CVAE models on the original and expanded datasets.

| Patient ID | (MSE) ROC-AUC | |
|---|---|---|
| | Original Dataset | Expanded Dataset |
| #1 | 0.537 ± 0.011 | **0.622** ± 0.076 |
| #2 | 0.433 ± 0.127 | 0.432 ± 0.136 |
| #3 | 0.720 ± 0.162 | **0.722** ± 0.189 |
| #4 | 0.650 ± 0.092 | **0.662** ± 0.073 |
| #5 | 0.800 ± 0.063 | 0.718 ± 0.051 |
| #6 | 0.512 ± 0.115 | **0.543** ± 0.143 |
| #7 | 0.929 ± 0.090 | 0.838 ± 0.118 |
| #8 | - | 0.342 ± 0.233 |
| #9 | 0.770 ± 0.048 | **0.833** ± 0.211 |
| **Mean (8)** | 0.669 ± 0.119 | **0.671** ± 0.135 |
| **Mean (9)** | - | 0.635 ± 0.155 |

Table 5.10: Comparison of (MSE) ROC-AUC values for the personalized CVAE models on the original and expanded datasets.

| Patient ID | Original Dataset | | | Expanded Dataset | | |
|---|---|---|---|---|---|---|
| | KL-C | KL-P | KL-R | KL-C | KL-P | KL-R |
| #1 | 18.60 ± 2.26 | **21.67** ± 3.81 | 16.77 ± 3.89 | 11.49 ± 0.60 | **13.13** ± 0.76 | **12.89** ± 0.76 |
| #2 | 12.44 ± 3.23 | **18.17** ± 0.00 | 10.10 ± 1.58 | 8.57 ± 1.41 | **10.73** ± 1.40 | 6.803 ± 0.87 |
| #3 | 18.49 ± 3.62 | **22.30** ± 3.20 | **26.45** ± 8.26 | 9.96 ± 1.69 | **12.05** ± 0.66 | **10.44** ± 0.65 |
| #4 | 14.38 ± 1.51 | **24.53** ± 12.60 | **21.46** ± 2.10 | 11.09 ± 0.67 | **11.99** ± 0.52 | **11.98** ± 0.52 |
| #5 | 15.03 ± 1.65 | **27.98** ± 2.92 | 13.71 ± 0.00 | 12.03 ± 0.41 | **13.52** ± 0.31 | 11.40 ± 0.24 |
| #6 | 12.43 ± 3.03 | **16.23** ± 7.57 | **17.48** ± 4.35 | 5.86 ± 1.42 | **7.67** ± 1.16 | **7.36** ± 1.16 |
| #7 | 14.03 ± 1.95 | **20.17** ± 5.85 | **14.45** ± 0.00 | 11.21 ± 0.45 | **12.67** ± 0.61 | 10.54 ± 0.46 |
| #8 | - | - | - | 8.93 ± 1.05 | 8.82 ± 1.07 | **9.27** ± 0.76 |
| #9 | 13.32 ± 2.28 | **16.12** ± 0.00 | **20.79** ± 0.00 | 9.33 ± 1.39 | **9.44** ± 0.97 | **15.05** ± 1.74 |
| **Median (8)** | 14.21 ± 2.30 | **20.92** ± 2.85 | **17.125** ± 4.81 | 10.52 ± 1.89 | **12.02** ± 1.92 | **10.97** ± 2.56 |
| **Median (9)** | - | - | - | 9.96 ± 1.81 | **11.99** ± 1.98 | **10.55** ± 2.46 |

Table 5.11: Comparison of KL anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the personalized CVAE models on the original and expanded datasets.

From Table 5.9, we can observe that MSE-P and MSE-R values are generally higher than MSE-C, though some patients (e.g., #4 and #5) show slight drops in MSE-R, which could be possibly due to the increased variability in data post-expansion. Similarly, Table 5.10 shows a minor improvement in ROC-AUC scores,

| Patient ID | (KL) ROC-AUC | |
|---|---|---|
| | Original Dataset | Expanded Dataset |
| #1 | 0.549 ± 0.115 | **0.632** ± 0.078 |
| #2 | 0.405 ± 0.127 | **0.552** ± 0.164 |
| #3 | 0.662 ± 0.179 | **0.689** ± 0.174 |
| #4 | 0.742 ± 0.087 | 0.668 ± 0.110 |
| #5 | 0.786 ± 0.068 | 0.687 ± 0.088 |
| #6 | 0.656 ± 0.142 | **0.723** ± 0.106 |
| #7 | 0.701 ± 0.149 | 0.570 ± 0.024 |
| #8 | - | 0.483 ± 0.133 |
| #9 | 0.770 ± 0.256 | **0.833** ± 0.211 |
| **Mean (8)** | 0.659 ± 0.119 | **0.669** ± 0.091 |
| **Mean (9)** | - | 0.649 ± 0.098 |

Table 5.12: Comparison of (KL) ROC-AUC values for the personalized CVAE models on the original and expanded datasets.

with the mean increasing from 0.669 to 0.671 for the original 8 patients. This indicates a modest improvement in anomaly detection, though the effect is less pronounced compared to the CAE model.

Additionally, from Table 5.11, we can still observe a distinction between clean (KL-C), pre-relapse (KL-P), and relapse (KL-R) states in the expanded dataset. However, the overall KL scores are lower. The ROC-AUC scores in Table 5.12 show a slight improvement for some patients, with the mean ROC-AUC increasing from 0.659 to 0.669 for the original 8 patients, which is consistent with the MSE results.

Overall, the personalized CVAE model shows a slight improvement in anomaly detection after dataset expansion, with a small increase in ROC-AUC scores and a more pronounced improvement in KL divergence scores, which suggests that the expanded dataset helps the model learn more meaningful latent space representations. However, we can observe that the model for the new patient #8 shows a significant drop in performance, which affects the overall mean ROC-AUC scores.

### Global Experiments

Tables 5.13-5.16 present the median MSE anomaly scores, mean MSE ROC-AUC scores, KL divergence anomaly scores and mean KL ROC-AUC scores for the global CVAE models on the original and expanded datasets.

| Norm. | Original Dataset | | | Expanded Dataset | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.905 ± 0.176 | **1.197** ± 0.420 | **1.139** ± 0.411 | 0.519 ± 0.025 | **0.610** ± 0.019 | **0.548** ± 0.051 |
| **Global** | 0.774 ± 0.120 | **0.804** ± 0.176 | 0.722 ± 0.139 | 0.521 ± 0.024 | **0.609** ± 0.011 | **0.553** ± 0.044 |

Table 5.13: Comparison of MSE anomaly scores for per-patient and global normalization schemes for clean (C), pre-relapse (P), and relapse (R) states for the global CVAE models on the original and expanded datasets.

| Norm. | (MSE) ROC-AUC | |
|---|---|---|
| | Original Dataset | Expanded Dataset |
| **Per-patient** | 0.633 ± 0.036 | 0.621 ± 0.051 |
| **Global** | 0.519 ± 0.021 | **0.618** ± 0.044 |

Table 5.14: Comparison of (MSE) ROC-AUC scores for per-patient and global normalization schemes for the global CVAE models on the original and expanded datasets.

The results in Table 5.13 suggest that the expanded dataset enhances the model's ability to distinguish between clean, pre-relapse, and relapse states, particularly under the global normalization scheme, with

consistently higher MSE-P and MSE-R values. Similarly, the ROC-AUC scores in Table 5.14 show a significant improvement for the global normalization scheme, with the mean ROC-AUC increasing from 0.519 to 0.618. This suggests that the expanded dataset helps the global CVAE model generalize better across patients and detect anomalies more accurately. However, the per-patient normalization scheme exhibits a slight decrease in performance, with a possible explanation being the increased variability in the data post-expansion.

| Norm. | Original Dataset | | | Expanded Dataset | | |
|---|---|---|---|---|---|---|
| | KL-C | KL-P | KL-R | KL-C | KL-P | KL-R |
| **Per-patient** | 22.31 ± 4.27 | **31.64** ± 9.21 | **28.50** ± 7.43 | 13.14 ± 0.58 | **14.96** ± 0.73 | **13.34** ± 4.49 |
| **Global** | 21.58 ± 3.68 | **24.78** ± 4.94 | **21.78** ± 4.22 | 12.87 ± 0.54 | **14.32** ± 0.18 | **13.07** ± 0.17 |

Table 5.15: Comparison of KL anomaly scores for per-patient and global normalization schemes for clean (C), pre-relapse (P), and relapse (R) states for the global CVAE models on the original and expanded datasets.

| Norm. | (KL) ROC-AUC | |
|---|---|---|
| | Original Dataset | Expanded Dataset |
| **Per-patient** | 0.694 ± 0.032 | 0.604 ± 0.037 |
| **Global** | 0.576 ± 0.026 | **0.596** ± 0.035 |

Table 5.16: Comparison of (KL) ROC-AUC scores for per-patient and global normalization schemes for the global CVAE models on the original and expanded datasets.

Similar to the personalized experiments, Table 5.15 shows that the model retains its ability to distinguish between clean and anomalous states, but the KL scores are lower compared to the original dataset. The ROC-AUC scores in Table 5.16 show a slight improvement for the global normalization scheme only, with the mean ROC-AUC increasing from 0.576 to 0.596. However, the per-patient normalization scheme exhibits a decrease in performance, which is consistent with the MSE results.

Considering all experiments, the global CVAE model shows a slight improvement in performance after the dataset expansion, but the effect is less pronounced compared to the CAE model. It is possible that the CVAE model is more sensitive to the increased data variability, leading to reduced performance. Therefore, future work should explore different techniques to address this issue, such as adaptive regularization, patient-specific weights or fine-tuning for specific patients for whom the performance did not improve.

### 5.2.3 LSTMAE vs. CAE Model Comparison on the Expanded Database

In this section, we compare how the LSTMAE and CAE models perform on the expanded dataset. The main goal is to understand how adding temporal modeling with the LSTMAE influences anomaly detection compared to the CAE, which focuses on spectral/temporal features. To ensure a fair comparison, both models are evaluated using the same experimental setup and metrics.

Through these experiments, we aim to uncover the strengths and trade-offs between the LSTMAE's ability to capture sequential patterns and the CAE's focus on spectral/temporal representation. Additionally, we explore whether the expanded dataset provides a bigger advantage to the LSTMAE, given its focus on temporal dependencies, and assess how capable the model is at distinguishing between clean, pre-relapse, and relapse states.

#### Personalized Experiments

Tables 5.17 and 5.18 present the median MSE anomaly scores and mean ROC-AUC scores for the personalized CAE and LSTMAE models on the expanded dataset.

The results in Table 5.17 and 5.18 indicate that the LSTMAE model outperforms the CAE in most metrics, highlighting the advantages of incorporating temporal modeling for anomaly detection. The LSTMAE consistently achieves higher MSE-P and MSE-R scores for all patients with a better distinction between clean and anomalous states for most. This trend is further supported by higher overall ROC-AUC scores, with the

| Patient ID | CAE | | | LSTMAE | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| #1 | 0.267 ± 0.020 | **0.386** ± 0.067 | **0.337** ± 0.065 | 0.241 ± 0.040 | **0.396** ± 0.028 | **0.321** ± 0.031 |
| #2 | 0.429 ± 0.068 | **0.578** ± 0.024 | **0.461** ± 0.034 | 0.440 ± 0.047 | **0.584** ± 0.036 | **0.442** ± 0.012 |
| #3 | 0.408 ± 0.074 | **0.470** ± 0.025 | **0.551** ± 0.034 | 0.399 ± 0.115 | **0.509** ± 0.025 | **0.572** ± 0.040 |
| #4 | 0.248 ± 0.025 | **0.366** ± 0.017 | **0.280** ± 0.009 | 0.263 ± 0.024 | **0.382** ± 0.010 | **0.283** ± 0.013 |
| #5 | 0.277 ± 0.037 | **0.331** ± 0.034 | **0.267** ± 0.022 | 0.259 ± 0.017 | **0.355** ± 0.011 | **0.263** ± 0.004 |
| #6 | 0.571 ± 0.080 | **0.598** ± 0.019 | **0.602** ± 0.030 | 0.548 ± 0.130 | **0.589** ± 0.053 | **0.583** ± 0.042 |
| #7 | 0.290 ± 0.010 | **0.338** ± 0.014 | **0.345** ± 0.010 | 0.258 ± 0.026 | **0.313** ± 0.030 | **0.313** ± 0.029 |
| #8 | 0.295 ± 0.039 | **0.367** ± 0.010 | **0.284** ± 0.013 | 0.299 ± 0.067 | **0.336** ± 0.042 | **0.265** ± 0.034 |
| #9 | 0.458 ± 0.061 | **0.593** ± 0.034 | **0.638** ± 0.041 | 0.477 ± 0.088 | **0.538** ± 0.029 | **0.658** ± 0.056 |
| **Median** | 0.295 ± 0.105 | **0.386** ± 0.102 | **0.337** ± 0.139 | 0.299 ± 0.108 | **0.396** ± 0.104 | **0.321** ± 0.152 |

Table 5.17: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the personalized CAE and LSTMAE models on the expanded dataset.

| Patient ID | ROC-AUC | |
|---|---|---|
| | CAE | LSTMAE |
| #1 | 0.653 ± 0.048 | **0.668** ± 0.050 |
| #2 | 0.468 ± 0.159 | **0.500** ± 0.087 |
| #3 | 0.722 ± 0.165 | **0.733** ± 0.231 |
| #4 | 0.650 ± 0.093 | **0.653** ± 0.135 |
| #5 | 0.754 ± 0.082 | 0.727 ± 0.081 |
| #6 | 0.500 ± 0.159 | **0.510** ± 0.183 |
| #7 | 0.905 ± 0.055 | **0.958** ± 0.093 |
| #8 | 0.483 ± 0.187 | **0.492** ± 0.172 |
| #9 | 0.817 ± 0.186 | **0.850** ± 0.200 |
| **Mean** | 0.661 ± 0.146 | **0.679** ± 0.152 |

Table 5.18: Comparison of (MSE) ROC-AUC scores for the personalized CAE and LSTMAE models on the expanded dataset.

mean ROC-AUC increasing from 0.661 to 0.679 for all 9 patients, with significant improvements for individuals like Patient #7. However, for a few cases (e.g., Patient #5), the LSTMAE shows slight performance decline.

### Global Experiments

Tables 5.19 and 5.20 present the median MSE anomaly scores and mean ROC-AUC scores for the global CAE and LSTMAE models on the expanded dataset.

| Norm. | CAE | | | LSTMAE | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **Per-Patient** | 0.188 ± 0.020 | **0.386** ± 0.067 | **0.337** ± 0.065 | 0.256 ± 0.009 | **0.322** ± 0.004 | **0.267** ± 0.002 |
| **Global** | 0.429 ± 0.068 | **0.578** ± 0.024 | **0.461** ± 0.034 | 0.245 ± 0.006 | **0.326** ± 0.013 | **0.270** ± 0.004 |

Table 5.19: Comparison of MSE anomaly scores for CAE and LSTMAE under per-patient and global normalization schemes.

Similarly to the personalized experiments, the global LSTMAE model outperforms the CAE model, particularly in terms of ROC-AUC scores, as shown in Table 5.20. The mean ROC-AUC increases from 0.618 to 0.640 for the per-patient normalization scheme and from 0.633 to 0.648 for the global normalization scheme, suggesting that the LSTMAE model is more effective overall, especially under the global normalization scheme. Additionally, the MSE results in Table 5.19 demonstrate a comparable ability to distinguish between clean and anomalous states, further emphasizing the LSTMAE's effectiveness in anomaly detection.

Overall, the LSTMAE model consistently outperforms the CAE model in both personalized and global

| Norm. | ROC-AUC | |
|---|---|---|
| | CAE | LSTMAE |
| Per-Patient | 0.618 ± 0.023 | **0.640** ± 0.031 |
| Global | 0.633 ± 0.033 | **0.648** ± 0.031 |

Table 5.20: Comparison of ROC-AUC scores for CAE and LSTMAE under per-patient and global normalization schemes.

experiments, with higher ROC-AUC scores. This suggests that the LSTMAE's ability to capture sequential dependencies is beneficial for relapse prediction. However, the CAE model still shows competitive results, especially in terms of MSE scores, also highlighting the importance of spectral/temporal representation for anomaly detection.

### 5.2.4 LSTMVAE vs. CVAE Model Comparison on the Expanded Database

In this final section of the audio experiments, we compare the performance of the LSTMVAE and CVAE models to evaluate their effectiveness in anomaly detection across both personalized and global experimental setups using the expanded dataset, repeating the same analysis performed in Section 5.2.3.

#### Personalized Experiments

Tables 5.21-5.24 present the median MSE anomaly scores, mean MSE ROC-AUC scores, KL divergence anomaly scores, and mean KL ROC-AUC scores for the personalized CVAE and LSTMVAE models on the expanded dataset.

| Patient ID | CVAE | | | LSTMVAE | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| #1 | 0.523 ± 0.033 | **0.638** ± 0.022 | **0.581** ± 0.023 | 0.712 ± 0.047 | **0.831** ± 0.040 | **0.786** ± 0.044 |
| #2 | 0.653 ± 0.131 | **0.889** ± 0.036 | **0.659** ± 0.024 | 0.871 ± 0.123 | **1.066** ± 0.056 | 0.744 ± 0.048 |
| #3 | 0.685 ± 0.121 | **0.752** ± 0.065 | **0.865** ± 0.066 | 0.926 ± 0.129 | **1.014** ± 0.072 | **1.133** ± 0.070 |
| #4 | 0.599 ± 0.038 | **0.734** ± 0.024 | **0.603** ± 0.020 | 0.799 ± 0.037 | **0.977** ± 0.040 | **0.832** ± 0.038 |
| #5 | 0.569 ± 0.030 | **0.649** ± 0.024 | 0.549 ± 0.020 | 0.778 ± 0.037 | **0.920** ± 0.024 | 0.769 ± 0.026 |
| #6 | 0.846 ± 0.131 | **0.867** ± 0.040 | **0.874** ± 0.042 | 0.972 ± 0.141 | **0.991** ± 0.065 | **0.974** ± 0.062 |
| #7 | 0.504 ± 0.030 | **0.632** ± 0.031 | **0.585** ± 0.015 | 0.704 ± 0.041 | **0.867** ± 0.025 | **0.752** ± 0.014 |
| #8 | 0.617 ± 0.109 | **0.631** ± 0.045 | 0.567 ± 0.047 | 0.768 ± 0.049 | 0.739 ± 0.016 | 0.653 ± 0.021 |
| #9 | 0.733 ± 0.132 | **0.809** ± 0.044 | **0.967** ± 0.065 | 0.918 ± 0.142 | **0.974** ± 0.042 | **1.248** ± 0.077 |
| **Median** | 0.617 ± 0.101 | **0.734** ± 0.097 | 0.603 ± 0.152 | 0.778 ± 0.092 | **0.974** ± 0.096 | **0.786** ± 0.188 |

Table 5.21: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the personalized CVAE and LSTVMAE models on the expanded dataset.

| Patient ID | ROC-AUC (MSE) | |
|---|---|---|
| | CVAE | LSTMVAE |
| #1 | 0.622 ± 0.076 | **0.644** ± 0.060 |
| #2 | 0.432 ± 0.136 | **0.435** ± 0.095 |
| #3 | 0.722 ± 0.189 | **0.744** ± 0.130 |
| #4 | 0.662 ± 0.073 | **0.667** ± 0.110 |
| #5 | 0.718 ± 0.051 | **0.726** ± 0.097 |
| #6 | 0.543 ± 0.143 | **0.577** ± 0.113 |
| #7 | 0.838 ± 0.118 | 0.787 ± 0.074 |
| #8 | 0.342 ± 0.233 | **0.425** ± 0.187 |
| #9 | 0.833 ± 0.211 | **0.850** ± 0.200 |
| **Mean** | 0.635 ± 0.155 | **0.667** ± 0.188 |

Table 5.22: Comparison of (MSE) ROC-AUC scores for the personalized CVAE and LSTMVAE models on the expanded dataset.

The results from Tables 5.21 and 5.22 indicate that while the LSTMVAE offers an advantage in some aspects, the CVAE performs comparably in terms of MSE anomaly scores. For most patients, the MSE-P and MSE-R values for both models are consistently higher than MSE-C, indicating that both models can distinguish between clean, pre-relapse, and relapse states effectively. However, the LSTMVAE does not consistently outperform the CVAE in terms of MSE scores. This suggests that the added complexity of temporal modeling in the LSTMVAE does not necessarily translate into a significantly better ability to detect anomalies. Where the LSTMVAE does show an improvement is in the ROC-AUC scores. For Patients #3, #6, #8, and #9 the LSTMVAE outperforms the CVAE, with the overall mean ROC-AUC increasing from 0.635 to 0.667 for all 9 patients, as shown in Table 5.22. This suggests that the LSTMVAE is better at ranking anomalies, even if the reconstruction errors are similar.

| **Patient ID** | **CVAE** | | | **LSTMVAE** | | |
|---|---|---|---|---|---|---|
| | **KL-C** | **KL-P** | **KL-R** | **KL-C** | **KL-P** | **KL-R** |
| #1 | 11.49 ± 0.60 | **13.13** ± 0.76 | **12.89** ± 0.76 | 2.71 ± 0.29 | **3.02** ± 0.18 | **2.92** ± 0.26 |
| #2 | 8.57 ± 1.41 | **10.74** ± 1.40 | 6.803 ± 0.87 | 1.59 ± 0.17 | **1.95** ± 0.15 | 1.50 ± 0.08 |
| #3 | 9.96 ± 1.69 | **12.05** ± 0.66 | **10.44** ± 0.65 | 1.88 ± 0.26 | **2.13** ± 0.20 | 1.82 ± 0.24 |
| #4 | 11.09 ± 0.67 | **11.99** ± 0.52 | **11.98** ± 0.52 | 2.06 ± 0.12 | **2.28** ± 0.14 | **2.08** ± 0.16 |
| #5 | 12.03 ± 0.41 | **13.52** ± 0.31 | 11.40 ± 0.24 | 2.20 ± 0.25 | 2.09 ± 0.24 | 2.04 ± 0.21 |
| #6 | 5.86 ± 1.42 | **7.67** ± 1.16 | **7.36** ± 1.16 | 1.12 ± 0.19 | **1.25** ± 0.13 | **1.28** ± 0.11 |
| #7 | 11.21 ± 0.45 | **12.67** ± 0.61 | 10.54 ± 0.46 | 2.38 ± 0.27 | 2.19 ± 0.07 | 2.08 ± 0.11 |
| #8 | 8.93 ± 1.05 | 8.82 ± 1.07 | **9.27** ± 0.76 | 2.11 ± 0.18 | 1.99 ± 0.15 | **2.36** ± 0.11 |
| #9 | 9.33 ± 1.39 | 9.14 ± 0.97 | **15.05** ± 1.74 | 1.64 ± 0.25 | 1.55 ± 0.23 | **2.45** ± 0.31 |
| **Median** | 9.96 ± 1.81 | **11.99** ± 1.98 | **10.55** ± 2.46 | 2.06 ± 0.44 | **2.09** ± 0.46 | **2.08** ± 0.47 |

Table 5.23: Comparison of KL anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the personalized CVAE and LSTVMAE models on the expanded dataset.

| **Patient ID** | **ROC-AUC (KL)** | |
|---|---|---|
| | **CVAE** | **LSTMVAE** |
| #1 | 0.632 ± 0.078 | **0.675** ± 0.212 |
| #2 | 0.552 ± 0.164 | **0.582** ± 0.146 |
| #3 | 0.689 ± 0.174 | 0.633 ± 0.109 |
| #4 | 0.668 ± 0.110 | 0.611 ± 0.094 |
| #5 | 0.687 ± 0.088 | 0.361 ± 0.090 |
| #6 | 0.723 ± 0.106 | 0.490 ± 0.333 |
| #7 | 0.570 ± 0.024 | 0.268 ± 0.238 |
| #8 | 0.483 ± 0.133 | **0.525** ± 0.200 |
| #9 | 0.833 ± 0.211 | 0.767 ± 0.162 |
| **Mean** | 0.649 ± 0.098 | 0.582 ± 0.147 |

Table 5.24: Comparison of (KL) ROC-AUC scores for the personalized CVAE and LSTMVAE models on the expanded dataset.

The results presented in Tables 5.23 and 5.24 indicate that the CVAE generally outperforms the LSTMVAE in modeling latent space differences using KL divergence anomaly scores. The KL-P, and KL-R values for the CVAE are consistently higher than those for the LSTMVAE, suggesting that the CVAE captures more pronounced latent space differences between clean, pre-relapse, and relapse states. For instance, for Patients #1 and #3, the KL-P and KL-R values for the CVAE show stronger separations than those of the LSTMVAE, highlighting its superior ability to identify anomalies in these cases. In terms of ROC-AUC, the CVAE also performs better in most cases, with mean values of 0.669 for the 8-patient dataset and 0.649 for the 9-patient dataset, compared to 0.596 and 0.582, respectively for the LSTMVAE. While the LSTMVAE shows better performance for certain patients (e.g., #1 and #8), its overall lower mean scores indicate less consistent classification of anomalous states. The inferior performance of the LSTMVAE in this case can be attributed to the added complexity introduced by sequential modeling which might not translate into better anomaly detection for KL divergence.

In summary, the comparison of the CVAE and LSTMVAE personalized models across both MSE and KL divergence metrics reveals complementary strengths and weaknesses for each architecture. For MSE, the results indicate comparable performance between the two models in identifying anomalies. While the LSTMVAE demonstrated slight improvements in ROC-AUC scores for some patients, the CVAE achieved similar levels of anomaly detection through reconstruction errors. In contrast, the KL divergence results highlight the superior performance of the CVAE in modeling latent space differences between clean, pre-relapse, and relapse states. The LSTMVAE is optimized for capturing temporal dependencies, which may provide an advantage in tasks where sequential patterns are critical. However, this might lead to over-regularization, in the latent space, as reflected in its lower KL divergence performance. On the other hand, the CVAE, seems to be the more suitable model for capturing differences in clean and anomalous states.

### Global Experiments

Tables 5.25-5.28 present the median MSE anomaly scores, KL divergence anomaly scores and their corresponding ROC-AUC scores for the global CVAE and LSTMVAE models on the expanded dataset.

| Norm. | CVAE | | | LSTMVAE | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.519 ± 0.025 | **0.610** ± 0.019 | **0.548** ± 0.051 | 0.774 ± 0.011 | **0.906** ± 0.029 | **0.799** ± 0.012 |
| **Global** | 0.521 ± 0.024 | **0.609** ± 0.011 | **0.553** ± 0.044 | 0.732 ± 0.040 | **0.914** ± 0.020 | **0.797** ± 0.013 |

Table 5.25: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the personalized CVAE and LSTMVAE models on the expanded dataset.

| Norm. | ROC-AUC (MSE) | |
|---|---|---|
| | CVAE | LSTMVAE |
| **Per-patient** | 0.621 ± 0.051 | 0.619 ± 0.015 |
| **Global** | 0.618 ± 0.044 | 0.618 ± 0.032 |

Table 5.26: Comparison of (MSE) ROC-AUC scores for CVAE and LSTMVAE under per-patient and global normalization schemes.

The results in Tables 5.25 and 5.26 show that the LSTMVAE demonstrates similar performance to the CVAE in terms of MSE anomaly scores and ROC-AUC scores. This indicates that the LSTMVAE is effective in distinguishing between clean and anomalous states, but does not necessarily outperform the CVAE in terms of anomaly detection.

| Norm. | CVAE | | | LSTMVAE | | |
|---|---|---|---|---|---|---|
| | KL-C | KL-P | KL-R | KL-C | KL-P | KL-R |
| **Per-patient** | 13.14 ± 0.58 | **14.96** ± 0.73 | **13.34** ± 4.49 | 2.02 ± 0.17 | **2.26** ± 0.18 | **2.08** ± 0.15 |
| **Global** | 12.87 ± 0.54 | **14.32** ± 0.18 | **13.07** ± 0.17 | 2.03 ± 0.14 | **2.27** ± 0.16 | **2.07** ± 0.11 |

Table 5.27: Comparison of KL anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the personalized CVAE and LSTMVAE models on the expanded dataset.

| Norm. | ROC-AUC (KL) | |
|---|---|---|
| | CVAE | LSTMVAE |
| **Per-patient** | 0.604 ± 0.037 | **0.614** ± 0.030 |
| **Global** | 0.596 ± 0.035 | **0.606** ± 0.027 |

Table 5.28: Comparison of (KL) ROC-AUC scores for CVAE and LSTMVAE under per-patient and global normalization schemes.

The KL anomaly scores in Table 5.27 indicate that the CVAE captures more pronounced differences between clean, pre-relase, and relapse states compared to the LSTMVAE. Despite the CVAE's higher KL scores, the

ROC-AUC results in Table 5.28 show that the LSTMVAE slightly outperforms the CVAE. The LSTMVAE achieves better mean ROC-AUC values under both normalization schemes, with 0.614 for per-patient and 0.606 for global normalization, compared to 0.604 and 0.596 for the CVAE, respectively. This suggests that the LSTMVAE is more effective at detecting anomalies, even if the KL divergence scores are lower.

In conclusion, the results from both experimental setups highlight that while the LSTMVAE is an effective and comparable alternative, the CVAE emerges as better suited for this specific task.

## 5.3 Discussion

Expanding the dataset to include additional patients and more utterances per patient has notably improved anomaly detection performance across all models, although the degree of improvement depends on both the model architecture and experimental setup.

A principal observation from the CAE experiments is that increasing data diversity and volume generally leads to higher reconstruction errors for anomalous speech, especially during relapse states. In personalized experiments, most patients exhibit stronger separation between clean and anomalous states, with consistent increases in ROC-AUC. In global experiments, the CAE demonstrated significant improvement in ROC-AUC, suggesting that the model benefits from increased data diversity, enabling better generalization across different patients.

The CVAE exhibits strong distinction ability in personalized experiments, effectively separating clean and anomalous states, with slight improvements in ROC-AUC scores. However, in global experiments, the per-patient normalization scheme shows a slight decrease in ROC-AUC, indicating sensitivity to the added individual variability introduced by the expanded dataset. This sensitivity likely could possibly lead to over-regularization in the latent space, emphasizing the need for refinements such as adaptive regularization techniques, patient-specific weights, or fine-tuning. These adjustments could also enhance the CVAE's performance in personalized experiments. Conversely, the global normalization scheme shows a significant increase in ROC-AUC, indicating that a common normalization approach enhances the CVAE's ability to generalize across patients by prioritizing shared patterns over individual differences.

The inclusion of temporal modeling with the LSTMAE proves to be a promising approach for this task, showcasing clear advantages over the CAE across both experimental setups and normalization schemes. By leveraging sequential dependencies, the LSTMAE effectively captures temporal patterns associated with pre-relapse and relapse states. This capability translates into consistently higher ROC-AUC values for most patients, underscoring its potential as a robust method for relapse detection in speech data.

Despite the LSTMAE's strengths in temporal modeling, the CVAE emerges as the better-suited variational autoencoder model for this task. In personalized experiments, the CVAE performed comparably to the LSTMVAE, with the first model showing clearer separations across states in terms of KL-divergence scores, while the latter demonstrated better distinction in terms of MSE scores. Furthermore, the CVAE's ability to generalize across patients in global experiments suggests it is more resilient to increased data variability compared to the LSTMVAE. However, the LSTMAE remains a strong candidate for tasks where sequential and temporal cues are critical, and its performance could potentially be further enhanced through regularization, weighting, and fine-tuning strategies.

# Chapter 6

# Multimodal Experiments

This chapter presents the multimodal experiments, which form the central focus of this thesis, aiming to enhance relapse detection in mental health patients by integrating diverse data modalities. While the earlier chapter focused on audio-based anomaly detection, this chapter explores the potential of combining speech features with complementary biometric data collected from smartwatch sensors. These biometric signals, derived from accelerometer, gyroscope and heart rate measurements, capture key aspects of physical activity and physiological state. This fusion is based on the assumption that relapse-related anomalies may manifest across multiple physiological and behavioral domains and aims to provide a more holisitc and accurate relapse detection.

To explore the potential of multimodal fusion, two joint autoencoder models were developed, each featuring distinct autoencoder branches for audio and biometric data. The audio branch incorporates the CAE and LSTMAE models introduced in Chapter 5, which demonstrated strong performance in audio anomaly detection. For the biometric signals, a CAE model was developed, based on the best performing model in the e-Prevention experiments mentioned in 3.4. The two branches are fused in the latent space, allowing the model to learn a unified representation of multimodal data. Additionally, to validate this approach, ablation studies were conducted following the main experiments. These included testing the effect of disabling one of the branches during training to assess the contribution of each modality independently.

The ultimate objective of these experiments is to showcase the potential of multimodal fusion in enhancing relapse detection, compared to unimodal methods, and contribute meaningfully to relapse detection research in general.

## 6.1 Methodology

### 6.1.1 Audio and Biometric Data Alignment

As mentioned in Section 4.3.2, accurate alignment between audio and biometric data is necessary for training the multimodal joint autoencoder models. Initially, alignment was performed by matching the biometric data to the exact dates of the audio interviews. After preprocessing the bio data, as described in Section 4.3.3, the resulting "day-of" dataset consisted of 4 patients who had sufficient and valid data However, the small size of the dataset was insufficient for reliably training the joint autoencoder models. To address this limitation, we expanded the dataset by including biometric data within defined time windows around the interview dates. This approach resulted in the creation of three additional subsets: 3-day, 5-day, and 7-day datasets. The biometric data in these subsets were preprocessed similarly to the day-of dataset, with all dataset sizes summarized in Table 6.1.

| **Datasets** | **Day-of** | **3-day** | **5-day** | **7-day** |
|---|---|---|---|---|
| **Demographics** | | | | |
| Male/Female | 2/2 | 2/3 | 2/3 | 3/4 |
| Age (years) | 31 ± 8.7 | 30.2 ± 8 | 30.2 ± 8 | 28 ± 7.6 |
| Education (years) | 14 ± 2 | 14.4 ± 2 | 14.4 ± 2 | 13.7 ± 2 |
| Illness duration (years) | 8.8 ± 9 | 7.2 ± 8.6 | 7.2 ± 8.6 | 6.4 ± 7.4 |
| **Recorded Data** | | | | |
| Num. of Days Recorded (total) | 66 | 102 | 124 | 158 |
| Num. of Days Recorded (mean ± std) | 16.5 ± 5 | 20.4 ± 7.5 | 24.8 ± 7.2 | 22.6 ± 10.8 |
| Num. of Hours Recorded (total) | 888 | 1,752 | 2,400 | 3,280 |
| Num. of Hours Recorded (mean ± std) | 222 ± 45.7 | 350.4 ± 89.6 | 480 ± 96.7 | 468.6 ± 185 |
| Num. of 5-min intervals (total) | 10,656 | 21,024 | 28,800 | 39,360 |
| Num. of 5-min intervals (mean ± std) | 2,664 ± 548.9 | 4,204.8 ± 1,074.9 | 5,760 ± 1171 | 5,622.9 ± 2,219.5 |

Table 6.1: Comparison of demographics, illness information, and recorded biometric data for the patients in the day-of, 3-day, 5-day, and 7-day datasets after preprocessing and feature extraction.

Aligning audio and biometric data posed several challenges, primarily due to fundamental differences in their temporal structure and sampling characteristics. While mel-spectrograms are extracted from short

speech segments, biometric feature tensors capture continuous physiological signals over extended periods, often covering an entire day. This discrepancy complicated direct alignment, as there was no natural 1-to-1 correspondence between the two modalities. Additionally, as shown in Figure 6.1.1, each interview date in the audio dataset included a significantly larger number of mel-spectrograms compared to the number of biometric feature tensors available for the same date, further highlighting the challenge of direct alignment.

Figure 6.1.1 illustrates the disparity in the available mel-spectrograms and biometric feature tensors across the day-of, 3-day, 5-day, and 7-day datasets for the same patient and a specific interview date.

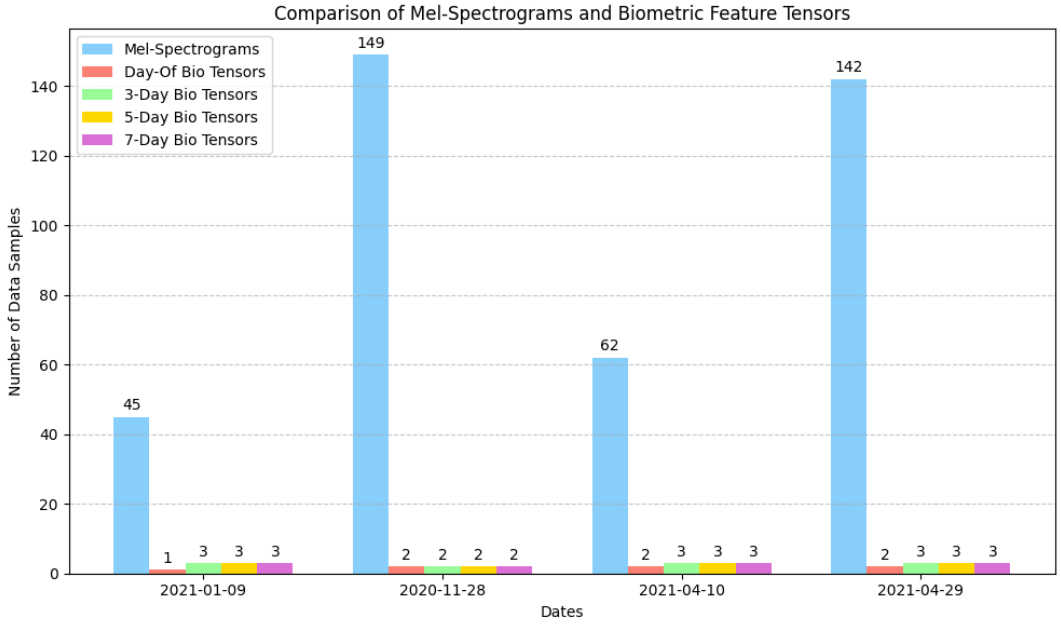



Figure 6.1.1: Comparison of the number of mel-spectrograms and biometric feature tensors across the day-of, 3-day, 5-day, and 7-day datasets for the same patient and sessions.

To address this challenge and ensure reliable and diverse training data, each mel-spectrogram was matched to multiple instances of the corresponding biometric data from the same date. This was done by duplicating the available biometric data to match the quantity of audio data, followed by a randomized assignment of biometric tensors to mel-spectrograms, before training. This approach ensured sufficient diversity in the aligned data while maintaining a temporal relationship between the audio and biometric data, avoiding inconsistencies during training.

### 6.1.2 Joint Autoencoder Architectures

The joint autoencoder architectures developed during this study combine two distinct branches: an audio branch and a biometric data branch, using different autoencoders for each data modality.

#### Audio Branch

For the audio branch of the joint autoencoder, we experimented with the CAE and the LSTMAE architectures introduced in Section 5.1.2. The specific architectural details of the CAE are presented in Table 5.1 and Figure 3.4.3, while the architecture of the LSTMAE is presented in Table 5.2 and Figure 5.1.2. For reference, the input, latent representation and output dimensions of the audio branches are summarized in Table 6.2.

| | CAE | LSTMAE |
|---|---|---|
| **Encoder Input** | 128×64×1 | 64×128 |
| **Latent Representation** | 1×1×256 | 64 |
| **Decoder Output** | 128×64×1 | 64×128 |

Table 6.2: Details of the input, latent representation and output dimensions of the CAE and LSTMAE audio branch architectures.

### Bio Branch

For the biometric data, a CAE model was implemented and adapted based on the best performing architecture from the experiments on the e-Prevention study [Zla+22]. The model follows a deep convolutional neural network structure, composed of encoding layers that progressively compress the 96×10 biometric feature input tensor, where 96 represents the number or 5-minute intervals in a 8 hour window and 10 is the number of features extracted from the accelerometer, gyroscopeand heart rate sensors, as described in Section 4.3.3, into a lower-dimensional latent representation. The encoding layers are followed by decoding layers that reconstruct the original input tensor.

The encoder is composed of 4 sequential convolutional downsampling (DS) blocks. Each block consists of a 1D-Convolution layer, followed by Batch Normalization, a LeakyReLU activation function, and a MaxPooling layer. The final encoder layer uses a 1D-Convolution layer to compress the input into the latent space representation, which has a shape of 1x16. The decoder mirrors the encoder in reverse, with 4 sequential convolutional upsampling (US) blocks. Each block starts with an Upsampling layer, progressively increasing the dimension of the latent space representation. The Upsampling layers are followed by a 1D-Convolution layer, Batch Normalization and a LeakyReLU activation function. The final decoder layer consists of a Dense layer with an output size of 10 and a linear activation function, reconstructing the original input tensor. The full architecture parameters are presented in Table 6.3, where the downsampling (DS) and upsampling (US) blocks are detailed, along with the number of filters, kernel size, pooling or upsampling size and output dimensions of each block.

| Conv. Block | Filters | Kernel Size | Pooling Size | Upsampling Size | Output Dimensions |
|---|---|---|---|---|---|
| DS1 | 4 | 5 | 2 | - | 48×4 |
| DS2 | 8 | 5 | 2 | - | 24×8 |
| DS3 | 16 | 5 | 2 | - | 12×16 |
| DS4 | 32 | 5 | 2 | - | 6×32 |
| Latent | 16 | - | - | - | 1×16 |
| US1 | 32 | 4 | - | 4 | 4×32 |
| US2 | 16 | 4 | - | 4 | 16×16 |
| US3 | 8 | 5 | - | 3 | 48×8 |
| US4 | 4 | 5 | - | 2 | 96×4 |
| Dense | 10 | - | - | - | 96×10 |

Table 6.3: Architecture parameters of the Convolutional Autoencoder (CAE) used in the biometric data branch.

The latent representation size was set to 16 after conducting a hyperparameter search using GridSearchCV over the range {8, 16, 32, 64}. The architecture parameters were then selected to achieve this latent space size while maintaining a relatively simple architecture and ensuring compatibility with the audio CAE implementation.

### CAE-CAE and LSTMAE-CAE Joint Autoencoders

The architecture of the joint autoencoder models was implemented as follows:

- **Input and Encoding**: Each input is first passed through its respective encoder branch. For mel spectrograms, the audio branch encoder compresses the input into a latent representation of size 256

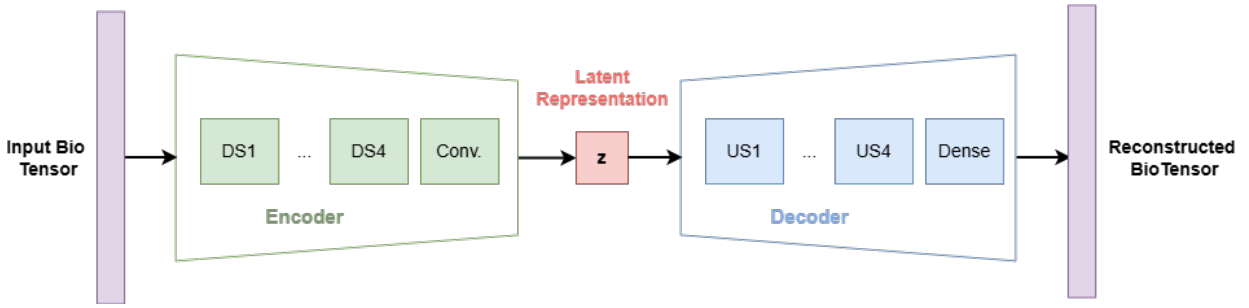



Figure 6.1.2: Overview of the proposed Convolutional Autoencoder (CAE) architecture for the biometric data branch.

in the case of the CAE or size 64 for the LSTMAE. Biometric feature tensors, on the other hand, are passed through the CAE bio branch encoder, which generates a latent representation of size 16.

- **Latent Space Fusion**: After encoding, the latent representations from each branch are concatenated to form a joint latent representation. For the CAE-CAE joint autoencoder, the concatenated latent representation has a size of 272, combining the 256-dimensional audio latent space and the 16-dimensional biometric latent space. For the LSTMAE-CAE joint autoencoder, the concatenated latent representation has a size of 80, combining the 64-dimensional LSTMAE latent space and the 16-dimensional biometric latent space. The joint latent representation is not reduced further. Experimental observations showed that reducing the dimensionality at this stage did not improve model performance.
- **Decoding and Output**: Finally, the joint latent representation is processed through Dense and Reshape layers, transforming it back into the respective encoded latent sizes of each branch. The decoders then reconstruct the original mel spectrograms and biometric feature tensors, completing the autoencoder's pipeline.

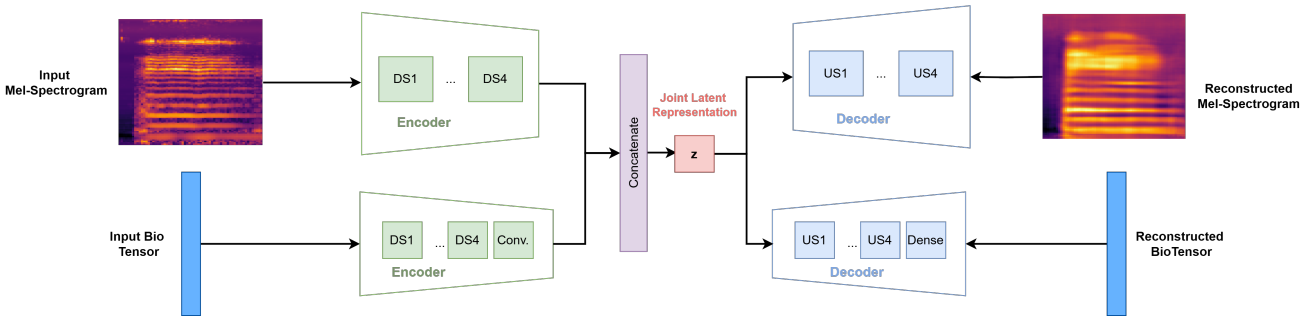



Figure 6.1.3: Overview of the proposed CAE-CAE joint autoencoder architecture.

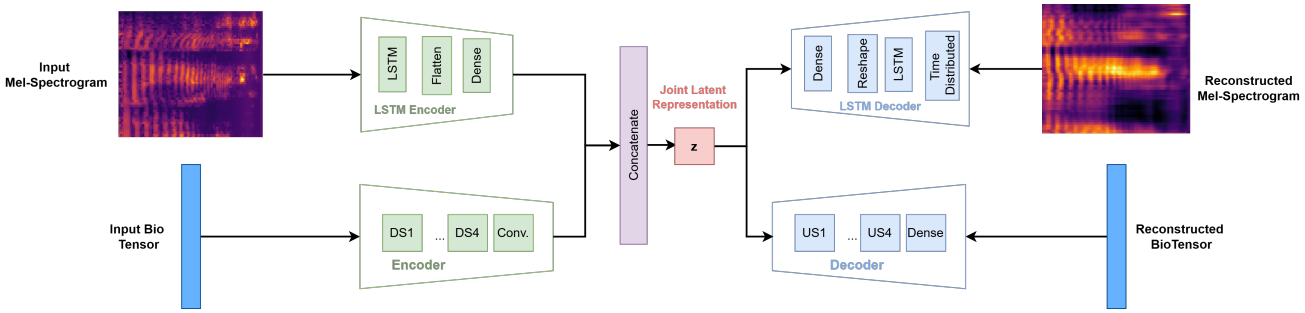



Figure 6.1.4: Overview of the proposed LSTMAE-CAE joint autoencoder architecture.

### 6.1.3 Model Training

The training pipeline for the joint autoencoder models followed the same methodology as in the case of the unimodal audio model, described in Section 5.1.3, with specific adjustments tailored to the multimodal data. Both the audio and biometric data were split into training, validation, and test datasets in a 60-20-20 ratio based on the interview sessions. As mentioned before, pre-relapse and relapse data were excluded from the training and validation datasets and was used exclusively during testing to evaluate the model's ability to identify anomalies. Additionally, as in the unimodal experiments, to prevent session-wise overfitting, the splits were structured so that all mel-spectrograms and biometric feature tensors corresponding to the same interview session remained within the same fold.

#### Loss Function

The loss function used for both branches was the Mean Squared Error (MSE) loss, focusing on minimizing the reconstruction error for both the mel spectrograms and the biometric features. The loss weights for each autoencoder branch during training were initialized based on extensive experimentation to balance the reconstruction quality of both the audio and bio branches. As shown in Table 6.4, the audio branch was assigned a higher weight in both joint models, reflecting the intention to prioritize audio reconstruction.

| Model | Audio Branch | Bio Branch |
|---|---|---|
| CAE-CAE | 0.6 | 0.4 |
| LSTMAE-CAE | 0.8 | 0.2 |

Table 6.4: Loss weights of the audio and bio branches during training of the CAE-CAE and LSTMAE-CAE joint autoencoder models.

#### Hyperparameters and Optimization

Both joint autoencoder models were trained using a maximum training duration of 200 epochs, a batch size of 8, and the Adam optimizer. To prevent overfitting, early stopping was applied with a patience of 10 epochs. Additionally, for the learning rate, a value of $3 \times 10^{-4}$ was selected for the CAE-CAE model, while a value of $1 \times 10^{-3}$ was chosen for the LSTMAE-CAE model. A dropout rate of 0.2 was applied to the audio branch of the LSTMAE-CAE model.

### 6.1.4 Evaluation Metrics

The performance of the joint autoencoder models was evaluated on both MSE and ROC-AUC, as mentioned in Section 5.1.4. Specifically, the MSE and ROC-AUC scores of each branch (audio and biometric) were compared to the metrics obtained from the unimodal models. This comparison allowed for a direct assessment of how the joint models leveraged both modalities relative to the independent approaches.

Additionally, a combined MSE score was calculated for the joint models using a weighted sum of the audio and bio branch MSE scores, while the corresponding ROC-AUC scores were computed from these combined anomaly scores. The combined MSE was calculated as follows:

$$\mathrm{MSE}_{\mathrm{joint}} = \mathrm{w}_{\mathrm{audio}} \cdot \mathrm{MSE}_{\mathrm{audio}} + \mathrm{w}_{\mathrm{bio}} \cdot \mathrm{MSE}_{\mathrm{bio}} \tag{6.1.1}$$

Where $\mathrm{w}_{\mathrm{audio}}$ and $\mathrm{w}_{\mathrm{bio}}$ are the weights assigned to the audio and bio branches during training, as shown in Table 6.4.

The combined metrics were then evaluated against both the unimodal baselines to determine the overall effectiveness of the joint models.

### 6.1.5 Unimodal Model Baselines

As mentioned previously, in order to evaluate the performance of the joint autoencoder models, unimodal baselines were established for both the audio and biometric data. The unimodal models were trained and

evaluated to establish baseline performance metrics for each dataset, including the day of, 3-day, 5-day and 7-day datasets. The CAE and LSTMAE models for the audio branch were trained exactly as described in Section 5.1.3. Similarly, the bio CAE model followed the same training methodology, using a learning rate of $3 \times 10^{-4}$, a batch size of 8, 200 epochs, and the Adam optimizer. The evaluation methodology for all unimodal models was consistent with that described earlier and in Section 5.1.4.

## 6.2 Joint CAE-CAE Model

This section presents the results of the CAE-CAE joint autoencoder experiments conducted on the day-of, 3-day, 5-day and 7-day datasets, with the overall format of the results following the one used in Section 5.2.

For the day-of dataset's personalized experiments, results are presented with MSE and ROC-AUC values for all individual patients, in order to highlight the imbalance in performance between the audio and bio branches, underscoring the necessity for expanded datasets. For the remaining temporal window datasets, only the median MSE and mean ROC-AUC scores across patients are reported, since they effectively summarize the model's performance across all patients and are sufficient to highlight the trends and insights. The results for the global experiments also follow the same format as in Section 5.2.

The performance of the audio and bio branches is compared against their respective unimodal baselines, while the combined MSE and ROC-AUC metrics are computed to evaluate the overall effectiveness of the joint model. These experiments aim to evaluate the CAE-CAE model's ability to leverage audio and bio features for relapse detection, assess its performance against unimodal models and determine the influence of the temporal windows on its overall effectiveness.

### 6.2.1 Day of Interview

The day-of dataset includes the audio and biometric data of patients #1, #2, #4, and #6, with a total of 66 days of recorded data.

#### Personalized Experiments

Tables 6.5, 6.6, and 6.7 present the MSE anomaly scores of the audio and bio unimodal models and branches of the joint CAE-CAE model and its combined MSE scores, respectively.

| Patient ID | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| #1 | 0.366±0.016 | **0.673**±0.041 | **0.522**±0.034 | 0.459±0.044 | **0.761**±0.027 | **0.640**±0.036 |
| #2 | 0.553±0.061 | **0.648**±0.031 | **0.589**±0.035 | 0.604±0.077 | **0.790**±0.031 | **0.697**±0.023 |
| #4 | 0.310±0.079 | **0.413**±0.027 | **0.316**±0.015 | 0.400±0.075 | **0.504**±0.035 | **0.401**±0.021 |
| #6 | 0.675±0.116 | **0.908**±0.056 | **0.796**±0.060 | 0.854±0.138 | **1.124**±0.081 | **0.897**±0.024 |
| **Median** | 0.460±0.146 | **0.661**±0.175 | **0.556**±0.171 | 0.532±0.175 | **0.776**±0.220 | **0.669**±0.177 |

Table 6.5: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized unimodal audio model and audio branch of the joint CAE-CAE model for the day-of dataset.

| Patient ID | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| #1 | 0.959±0.265 | 0.798±0.055 | 0.793±0.076 | 1.223±0.367 | 1.021±0.090 | 0.839±0.212 |
| #2 | 0.698±0.087 | **0.726**±0.059 | **0.810**±0.050 | 0.847±0.082 | 0.834±0.034 | **0.874**±0.029 |
| #4 | 1.006±0.583 | 0.941±0.115 | **1.049**±0.161 | 1.252±0.772 | 1.117±0.193 | **1.289**±0.121 |
| #6 | 0.936±0.131 | **1.238**±0.258 | 0.644±0.231 | 1.216±0.150 | **1.144**±0.178 | 0.681±0.158 |
| **Median** | 0.948±0.119 | 0.870±0.196 | 0.802±0.145 | 1.220±0.167 | 1.069±0.122 | 0.857±0.225 |

Table 6.6: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized unimodal bio model and bio branch of the joint CAE-CAE model for the day-of dataset.

| Patient ID | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| #1 | 0.773±0.148 | **0.866**±0.034 | **0.859**±0.034 |
| #2 | 0.735±0.024 | **0.817**±0.026 | **0.765**±0.022 |
| #4 | 0.771±0.285 | 0.731±0.081 | 0.754±0.046 |
| #6 | 0.999±0.411 | **1.097**±0.090 | 0.921±0.069 |
| **Median** | 0.772±0.105 | **0.842**±0.135 | **0.812**±0.069 |

Table 6.7: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized joint CAE-CAE model for the day-of dataset.

Based on the results presented in Tables 6.5 and 6.6, we can clearly observe that the unimodal audio model has superior performance compared to the unimodal bio model in terms of distinguishing between the clean, pre-relapse, and relapse states. One contributing factor to this is the small size of the dataset (in terms of available biometric feature tensors) which may have limited the bio model's ability to learn sufficient representations of the data.

As it was expected, this limitation affected the joint autoencoder as well, since the audio branch of the joint CAE-CAE model also outperforms the bio branch in terms of MSE anomaly scores, indicating that the biometric data did not provide significant additional information to complement the audio modality. The combined MSE scores in Table 6.7 show that the joint CAE-CAE model's performance did not significantly improve the audio model's performance, but it did improve the bio model's performance, which is expected given the imbalance.

Table 6.8 presents the ROC-AUC scores for the personalized unimodal audio and bio models, the audio and bio branches of the joint CAE-CAE and the combined ROC-AUC scores for the day-of dataset. The bolded ROC-AUC scores represent the cases where the joint autoencoder branch outperformed its unimodal counterpart and the cases where the combined ROC-AUC score is higher than both the ROC-AUC scores of the audio and bio unimodal models. This notation is going to be used in the all following Tables.

| Patient ID | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| #1 | 0.860±0.049 | 0.860±0.049 | 0.240±0.206 | **0.360**±0.344 | 0.640±0.273 |
| #2 | 0.700±0.187 | **0.750**±0.224 | 0.650±0.200 | 0.500±0.158 | **0.850**±0.200 |
| #4 | 0.629±0.178 | 0.629±0.203 | 0.496±0.206 | **0.535**±0.196 | 0.594±0.181 |
| #6 | 0.675±0.100 | 0.650±0.094 | 0.325±0.218 | **0.450**±0.281 | 0.475±0.278 |
| **Mean** | 0.716±0.087 | **0.722**±0.092 | 0.428±0.158 | **0.461**±0.066 | 0.640±0.136 |

Table 6.8: Comparison of ROC-AUC scores of the personalized unimodal models and branches of the joint CAE-CAE model, as well as the combined ROC-AUC scores for the day-of dataset.

The ROC-AUC scores presented in Table 6.8 further validate the imbalance in performance between the audio and bio models, suggesting that the bio branch did not provide sufficient additional information to significantly enhance the joint model's performance. Despite this, the branches and combined ROC-AUC scores are comparable to, and in some cases slightly better than the unimodal models, which suggests that the added modality could be beneficial for relapse detection, given a larger and more diverse dataset to enable the joint model to effectively leverage complementary information from both modalities.

### Global Experiments

Tables 6.9-6.11 present the MSE anomaly scores of the global audio and bio unimodal models and branches of the joint CAE-CAE model and its combined MSE scores, respectively.

| Norm. | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.275±0.019 | **0.397**±0.022 | **0.309**±0.016 | 0.340±0.028 | **0.474**±0.031 | **0.375**±0.019 |
| **Global** | 0.250±0.043 | **0.354**±0.020 | **0.272**±0.011 | 0.348±0.029 | **0.465**±0.022 | **0.385**±0.010 |

Table 6.9: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global unimodal audio model and audio branch of the joint CAE-CAE model for the day-of dataset.

| Norm. | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.746±0.131 | 0.743±0.045 | **0.761**±0.042 | 0.861±0.175 | 0.858±0.081 | **0.868**±0.038 |
| **Global** | 0.828±0.171 | 0.704±0.082 | 0.721±0.114 | 0.961±0.147 | 0.931±0.110 | 0.862±0.026 |

Table 6.10: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the global unimodal bio CAE model and the audio branch of the joint CAE-CAE model for the day-of dataset.

| Patient ID | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.569±0.057 | **0.618**±0.041 | **0.592**±0.021 |
| **Global** | 0.597±0.036 | **0.657**±0.045 | 0.571±0.016 |

Table 6.11: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global joint CAE-CAE model for the day-of dataset.

The results in Tables 6.9 and 6.10 also highlight that the imbalance in performance between the audio and bio models is present in the global experiments as well. The audio unimodal model achieved clear distinction between clean, pre-relapse and relapse states, while the bio model's performance is significantly lower. The audio branch follows the same trend, with its performance being slightly improved compared to the unimodal model. This suggest that the biometric modality did provide some additional information to the joint model, but it was not sufficient to significantly improve the model's overall performance, as also shown by the combined MSE scores in Table 6.11.

Table 6.12 presents the ROC-AUC scores for the global unimodal models and branches of the joint CAE-CAE model and its combined ROC-AUC scores. The results show that the audio and bio branches' ROC-AUC scores are again higher than their unimodal counterparts, but with the imbalance between them persisting. The combined ROC-AUC scores indicate that the joint model's performance is improved only compared to the bio unimodal model, which is consistent with the MSE results.

| Norm. | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| **Per-patient** | 0.684±0.125 | **0.698**±0.162 | 0.452±0.125 | **0.474**±0.162 | 0.541±0.140 |
| **Global** | 0.674±0.127 | **0.691**±0.118 | 0.437±0.127 | **0.483**±0.118 | 0.557±0.119 |

Table 6.12: Comparison of ROC-AUC scores of the global unimodal models and branches of the joint CAE-CAE model, as well as the combined ROC-AUC scores for the day-of dataset.

Overall, the results of the day-of dataset experiments show that there is potential for the joint CAE-CAE model to improve relapse detection, but the imbalance in performance between the audio and bio models highlights the need for a larger and more diverse dataset to fully leverage the benefits of the multimodal approach.

### 6.2.2 Temporal Windows Around Interview

The 3-day dataset consists of audio and biometric data from patients #1, #2, #3, #4, and #6, totaling 102 days of recorded data. The 5-day dataset includes the same patients, with an expanded dataset of 124 days. The 7-day dataset extends the patients to #1, #2, #3, #4, #6, #8, and #9, including 158 recorded days in total.

#### Personalized Experiments

Tables 6.13-6.14 present the median MSE anomaly scores of the audio and bio unimodal models and branches of the joint CAE-CAE model and its combined MSE scores, respectively, for each temporal window dataset.

| Dataset | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.526±0.180 | **0.588**±0.202 | 0.480±0.197 | 0.553±0.162 | **0.741**±0.202 | **0.696**±0.211 |
| **5-day** | 0.464±0.121 | **0.548**±0.122 | 0.463±0.122 | 0.559±0.176 | **0.653**±0.209 | **0.594**±0.191 |
| **7-day** | 0.469±0.175 | **0.525**±0.106 | **0.503**±0.159 | 0.527±0.151 | **0.615**±0.124 | **0.647**±0.188 |

Table 6.13: Comparison of median MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized unimodal audio model and audio branch of the joint CAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.679±0.108 | 0.643±0.451 | **0.684**±0.273 | 0.768±0.204 | **0.821**±0.456 | **0.776**±0.351 |
| **5-day** | 0.541±0.073 | **0.600**±2.849 | **0.648**±0.255 | 0.662±0.225 | **0.783**±2.359 | **0.775**±0.382 |
| **7-day** | 0.615±0.196 | **0.832**±0.166 | **0.624**±0.355 | 0.675±0.259 | **1.056**±0.259 | **0.706**±0.390 |

Table 6.14: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized unimodal bio model and bio branch of the joint CAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.639±0.115 | **0.779**±0.642 | **0.734**±0.133 |
| **5-day** | 0.627±0.164 | **0.682**±0.996 | **0.724**±0.311 |
| **7-day** | 0.651±0.133 | **0.710**±0.135 | **0.807**±0.311 |

Table 6.15: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global joint CAE-CAE model for the 3, 5, and 7-day datasets.

The results in Tables 6.13-6.15 consistently demonstrate the superior performance of the joint CAE-CAE model over its respective unimodal models in detecting pre-relapse and relapse states across all datasets. In the 3-day dataset, both the audio and bio branches achieve higher MSE anomaly scores for the pre-relapse and relapse states, which is also reflected in the combined MSE scores. The similar trend is observed in the 5-day dataset, with higher MSE scores in anomalous states for both branches, and the combined MSE scores as well. Finally, the 7-day dataset exhibits the strongest improvements in detecting pre-relapse and relapse states, with the MSE anomaly scores consistently highlighting significant gains for both branches and the joint model overall.

Table 6.16 presents the ROC-AUC scores for the personalized unimodal audio and bio models, the audio and bio branches of the joint CAE-CAE and the combined ROC-AUC scores for the 3, 5, and 7-day datasets. Additionally, Figure 6.2.1 illustrates the mean ROC-AUC scores for the personalized experiments across all datasets, including the day-of dataset.

| Dataset | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | **Audio Unimodal** | **Audio Branch** | **Bio Unimodal** | **Bio Branch** | **Combined** |
| **3-day** | 0.618±0.062 | **0.674**±0.099 | 0.547±0.198 | **0.564**±0.188 | **0.656**±0.186 |
| **5-day** | 0.625±0.064 | **0.654**±0.079 | 0.603±0.126 | **0.626**±0.129 | **0.654**±0.143 |
| **7-day** | 0.598±0.064 | **0.643**±0.079 | 0.557±0.136 | **0.629**±0.120 | **0.652**±0.170 |

Table 6.16: Comparison of ROC-AUC scores of the personalized unimodal models and branches of the joint CAE-CAE model, as well as the combined ROC-AUC scores for the 3, 5, and 7-day datasets.

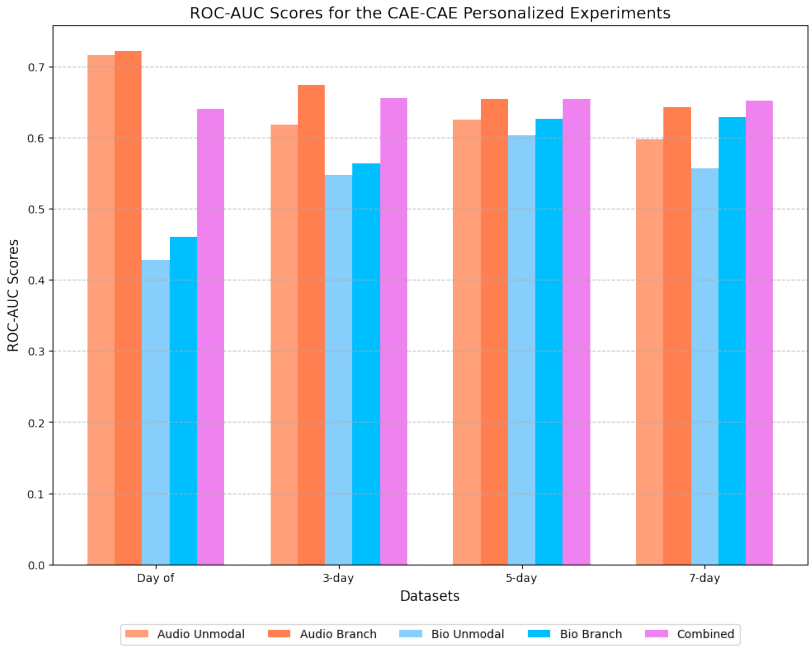



Figure 6.2.1: Mean ROC-AUC scores for the personalized experiments for all datasets.

The ROC-AUC results in Table 6.16 and Figure 6.2.1 further demonstrate the advantages of integrating audio and biometric data for relapse detection. The joint CAE-CAE model consistently surpasses the unimodal models in ROC-AUC scores across all datasets, with the audio branch exhibiting the most substantial improvement over its unimodal counterpart, particularly in the 3-day dataset (0.618 to 0.674), suggesting that biometric data enhances speech-based anomaly detection. Similarly, the bio branch also shows improvements in performance, although not as pronounced, confirming that audio data contributes to more effective predictions. Finally, the combined ROC-AUC scores also show improvements compared to both unimodal models, with the 3-day dataset achieving the highest score of 0.656. It should be noted that the comparisons are not between datasets, but rather between the models' performance within each dataset compared to their respective unimodal models.

Overall, in the personalized experiments, the multimodal approach proves to be effective in detecting pre-relapse and relapse anomalies. The findings emphasize that biometric data significantly improve speech-based models, while audio data reinforces biometric predictions, making the combined approach more robust than either modality alone. Additionally, the results suggest that the additional biometric data in the larger datasets had a positive impact on the joint model's performance, since the imbalance between the audio and bio branches was less pronounced compared to the day-of dataset.

## Global Experiments

### Per-Patient Normalization

Tables 6.17-6.19 present the MSE anomaly scores of the global unimodal audio and bio models and branches of the joint CAE-CAE model and its combined MSE scores, respectively, for each temporal window dataset with per-patient normalization. Table 6.20 and Figure 6.2.2 present the ROC-AUC scores.

| Dataset | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.286±0.037 | **0.337**±0.009 | 0.281±0.011 | 0.301±0.044 | **0.395**±0.024 | **0.312**±0.012 |
| **5-day** | 0.237±0.018 | **0.326**±0.010 | **0.258**±0.011 | 0.272±0.030 | **0.348**±0.018 | **0.285**±0.015 |
| **7-day** | 0.233±0.017 | **0.274**±0.015 | **0.245**±0.008 | 0.254±0.021 | **0.319**±0.011 | **0.283**±0.009 |

Table 6.17: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global (per-patient normalization) unimodal audio model and audio branch of the joint CAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.704±0.063 | 0.640±0.045 | 0.653±0.058 | 0.761±0.100 | 0.688±0.035 | **0.768**±0.081 |
| **5-day** | 0.558±0.062 | 0.531±0.027 | **0.620**±0.030 | 0.594±0.044 | **0.631**±0.029 | **0.655**±0.072 |
| **7-day** | 0.549±0.063 | **0.630**±0.033 | **0.639**±0.057 | 0.578±0.023 | **0.658**±0.041 | **0.597**±0.024 |

Table 6.18: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the global (per-patient normalization) unimodal bio CAE model and the audio branch of the joint CAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.479±0.040 | **0.508**±0.037 | **0.518**±0.015 |
| **5-day** | 0.396±0.034 | **0.437**±0.026 | **0.473**±0.029 |
| **7-day** | 0.401±0.008 | **0.466**±0.013 | **0.429**±0.008 |

Table 6.19: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global (per-patient normalization) joint CAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| **3-day** | 0.615±0.069 | **0.639**±0.112 | 0.490±0.049 | **0.539**±0.042 | 0.602±0.054 |
| **5-day** | 0.605±0.032 | **0.626**±0.075 | 0.519±0.077 | **0.556**±0.040 | **0.612**±0.054 |
| **7-day** | 0.603±0.059 | **0.614**±0.048 | 0.553±0.029 | **0.555**±0.032 | 0.582±0.028 |

Table 6.20: Comparison of ROC-AUC scores of the global (per-patient normalization) unimodal models and branches of the joint CAE-CAE model, as well as the combined ROC-AUC scores for the 3, 5, and 7-day datasets.

The results in Tables 6.17-6.19 reveal a moderate enhancement in the joint CAE-CAE model's performance compared to the unimodal models. The improvement is more evident in the ROC-AUC scores in Table 6.20 and Figure fig:cae-cae-roc-auc-global-per-patient, where both the audio and bio branches consistently outperform their unimodal counterparts. However, the combined ROC-AUC scores show a more varied trend; while the 5-day dataset demonstrates a slight overall advantage over both unimodal models, in the 3-day and 7-day datasets, the combined model primarily surpasses the bio unimodal model.

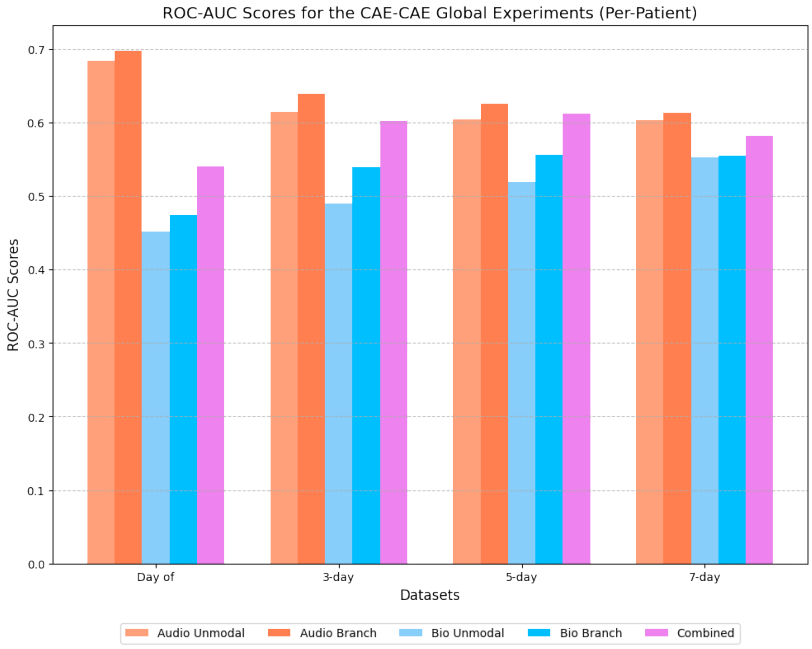



Figure 6.2.2: Mean ROC-AUC scores for the global experiments (per-patient normalization) for all datasets.

### Global Normalization

Tables 6.21-6.23 present the MSE anomaly scores of the global unimodal audio and bio models and branches of the joint CAE-CAE model and its combined MSE scores, respectively, for each temporal window dataset with global normalization. Table 6.24 and Figure 6.2.3 present the ROC-AUC scores.

| Dataset | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | **MSE-C** | **MSE-P** | **MSE-R** | **MSE-C** | **MSE-P** | **MSE-R** |
| **3-day** | 0.252±0.034 | **0.340**±0.008 | **0.279**±0.007 | 0.299±0.050 | **0.372**±0.031 | **0.313**±0.022 |
| **5-day** | 0.233±0.016 | **0.326**±0.019 | **0.257**±0.011 | 0.291±0.028 | **0.368**±0.013 | **0.292**±0.018 |
| **7-day** | 0.233±0.019 | **0.277**±0.010 | **0.250**±0.006 | 0.257±0.021 | **0.321**±0.012 | **0.276**±0.011 |

Table 6.21: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global (global normalization) unimodal audio model and audio branch of the joint CAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | **MSE-C** | **MSE-P** | **MSE-R** | **MSE-C** | **MSE-P** | **MSE-R** |
| **3-day** | 0.640±0.036 | 0.598±0.022 | 0.626±0.028 | 0.697±0.023 | 0.675±0.054 | **0.698**±0.009 |
| **5-day** | 0.555±0.043 | 0.537±0.028 | **0.601**±0.040 | 0.579±0.089 | **0.598**±0.043 | **0.680**±0.064 |
| **7-day** | 0.589±0.055 | **0.629**±0.032 | **0.663**±0.056 | 0.559±0.077 | **0.607**±0.066 | **0.589**±0.076 |

Table 6.22: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the global (global normalization) unimodal bio CAE model and the audio branch of the joint CAE-CAE model for the 3-day dataset.

| Datasets | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.453±0.022 | **0.495**±0.016 | **0.510**±0.013 |
| **5-day** | 0.435±0.047 | **0.455**±0.018 | **0.488**±0.035 |
| **7-day** | 0.375±0.043 | **0.449**±0.032 | **0.428**±0.033 |

Table 6.23: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global (global normalization) joint CAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| **3-day** | 0.619±0.086 | **0.622**±0.101 | 0.497±0.044 | **0.519**±0.028 | 0.598±0.033 |
| **5-day** | 0.602±0.061 | **0.610**±0.071 | 0.536±0.054 | **0.551**±0.062 | 0.588±0.048 |
| **7-day** | 0.600±0.060 | **0.607**±0.053 | 0.543±0.050 | **0.572**±0.039 | 0.598±0.046 |

Table 6.24: Comparison of ROC-AUC scores of the global (global normalization) unimodal models and branches of the joint CAE-CAE model, as well as the combined ROC-AUC scores for the 3, 5, and 7-day datasets.

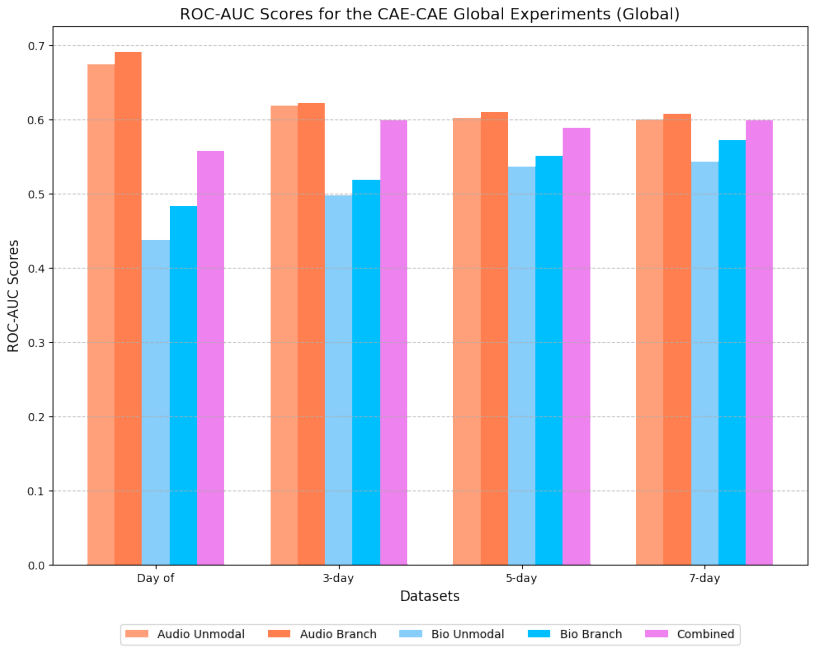



Figure 6.2.3: Mean ROC-AUC scores for the global experiments (global normalization) for all datasets.

Under global normalization, the performance gains are more moderate compared to the personalized results. While both branches of the joint CAE-CAE model still outperform unimodal approaches, the improvements are less pronounced and the combined model's performance is more variable, surpassing only the bio unimodal models. Therefore, the global normalization scheme may not be as effective as the per-person normalization, but the benefits of the multimodal approach are still present.

Therefore, the joint CAE-CAE model demonstrated its strongest performance in the personalized experiments across all larger datasets, where it effectively captured individual patient characteristics, leading to enhanced relapse detection. In the global experiments, the model consistently showed slight improvements over unimodal approaches, suggesting the potential for generalization in global setups with further research. Mixed results may stem from patient variability and differences in relapse severity, highlighting the challenges of generalizing across diverse populations.

Overall, the integration of multimodal data proved beneficial for relapse detection, with personalized models showing the most significant gains. These findings highlight the value of multimodal modeling for personalized mental health monitoring and emphasize the need for further refinement in global settings to enhance generalization across diverse patient datasets.

## 6.3 Joint LSTMAE-CAE Model

This section presents the results of the LSTMAE-CAE joint autoencoder experiments conducted on the day-of, 3-day, 5-day and 7-day datasets. The results are presented in the same format and structure as the joint CAE-CAE model results, with the same metrics and evaluation methods used to assess the model's performance.

### 6.3.1 Day of Interview

#### Personalized Experiments

Tables 6.25-6.27 present the MSE anomaly scores of the audio and bio unimodal models and branches of the joint LSTMSE-CAE model and its combined MSE scores, respectively.

| Patient ID | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| #1 | 0.391±0.024 | **0.758**±0.021 | **0.587**±0.018 | 0.359±0.034 | **0.703**±0.046 | **0.551**±0.025 |
| #2 | 0.555±0.078 | **0.662**±0.021 | **0.631**±0.028 | 0.554±0.063 | **0.676**±0.045 | **0.614**±0.044 |
| #4 | 0.309±0.084 | **0.441**±0.015 | **0.321**±0.014 | 0.354±0.094 | **0.486**±0.056 | **0.373**±0.030 |
| #6 | 0.717±0.131 | **1.006**±0.099 | **0.883**±0.084 | 0.666±0.155 | **0.906**±0.154 | **0.785**±0.100 |
| **Median** | 0.473±0.157 | **0.710**±0.203 | **0.609**±0.199 | 0.457±0.133 | **0.690**±0.149 | **0.583**±0.147 |

Table 6.25: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized unimodal audio model and audio branch of the joint LSTMAE-CAE model for the day-of dataset.

| Patient ID | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| #1 | 0.959±0.256 | 0.798±0.055 | 0.793±0.076 | 1.089±0.217 | 0.997±0.097 | 0.962±0.257 |
| #2 | 0.698±0.087 | **0.726**±0.059 | **0.810**±0.050 | 0.855±0.186 | **0.887**±0.070 | 0.820±0.234 |
| #4 | 1.006±0.583 | 0.941±0.115 | **1.049**±0.161 | 0.958±0.619 | **1.130**±0.202 | **1.200**±0.163 |
| #6 | 0.936±0.668 | **1.238**±0.258 | 0.644±0.231 | 1.004±0.631 | 0.946±0.311 | 0.626±0.327 |
| **Median** | 0.948±0.119 | 0.870±0.196 | 0.802±0.145 | 0.981±0.084 | 0.972±0.090 | 0.891±0.209 |

Table 6.26: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized unimodal bio model and bio branch of the joint LSTMAE-CAE model for the day-of dataset.

| Patient ID | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| #1 | 0.517±0.051 | **0.789**±0.045 | **0.716**±0.029 |
| #2 | 0.615±0.048 | **0.709**±0.024 | **0.648**±0.053 |
| #4 | 0.617±0.089 | 0.578±0.050 | 0.555±0.022 |
| #6 | 0.882±0.347 | **1.018**±0.145 | 0.829±0.142 |
| **Median** | 0.616±0.133 | **0.749**±0.149 | **0.682**±0.147 |

Table 6.27: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized joint LSTMAE-CAE model for the day-of dataset.

As it was observed in the CAE-CAE personalized experiments, the results presented in Tables 6.25 and 6.26 show that the unimodal audio model significantly outperforms the unimodal bio model in terms of distinguishing between the clean, pre-relapse, and relapse states.

This is also observed in the audio branch which outperforms the bio branch in the same way, with the performance of each branch separately being comparable to the unimodal models. The combined MSE scores in Table 6.27 also show that the joint LSTMAE-CAE model's performance is comparable to the audio branch's performance, which is expected given the imbalance and it suggests that the biometric data did not provide significant additional information to complement the audio modality in this case as well.

Table 6.28 presents the ROC-AUC scores for the personalized unimodal audio and bio models, the audio and bio branches of the joint CAE-CAE and the combined ROC-AUC scores for the day-of dataset.

| Patient ID | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| #1 | 0.840±0.049 | **0.880**±0.040 | 0.240±0.206 | **0.300**±0.200 | **0.860**±0.080 |
| #2 | 0.800±0.245 | 0.800±0.187 | 0.650±0.200 | 0.600±0.122 | 0.800±0.187 |
| #4 | 0.654±0.122 | **0.669**±0.150 | 0.496±0.206 | **0.500**±0.181 | 0.525±0.151 |
| #6 | 0.700±0.100 | 0.675±0.127 | 0.325±0.218 | **0.350**±0.242 | 0.450±0.257 |
| **Mean** | 0.700±0.100 | **0.756**±0.127 | 0.428±0.158 | **0.438**±0.158 | 0.659±0.089 |

Table 6.28: Comparison of ROC-AUC scores of the personalized unimodal models and branches of the joint LSTMAE-CAE model and its combined ROC-AUC scores for the day-of dataset.

The ROC-AUC scores presented in Table 6.28 also hightlight the imbalance in performance between the audio and bio models in a similar way to the CAE-CAE model. The mean ROC-AUC scores show that the audio and bio branches performed slightly better than the unimodal models, which suggests that the additional modalities could be beneficial for relapse detection, given a larger and more diverse dataset that balances the contribution of each modality.

### Global Experiments

Tables 6.29-6.31 present the MSE anomaly scores of the global audio and bio unimodal models and branches of the joint LSTMAE-CAE model and its combined MSE scores, respectively.

| Norm. | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.269±0.048 | **0.391**±0.012 | **0.305**±0.012 | 0.310±0.039 | **0.479**±0.037 | **0.396**±0.018 |
| **Global** | 0.277±0.021 | **0.382**±0.018 | **0.305**±0.012 | 0.354±0.028 | **0.511**±0.042 | **0.400**±0.012 |

Table 6.29: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global unimodal audio model and audio branch of the joint LSTMAE-CAE model for the day-of dataset.

| Norm. | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.746±0.131 | 0.743±0.045 | **0.761**±0.042 | 0.788±0.121 | **0.863**±0.062 | **0.866**±0.044 |
| **Global** | 0.828±0.171 | 0.704±0.082 | 0.721±0.114 | 0.802±0.136 | **0.825**±0.051 | **0.807**±0.050 |

Table 6.30: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the global unimodal bio CAE model and the audio branch of the joint LSTMAE-CAE model for the day-of dataset.

| Patient ID | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.450±0.022 | **0.514**±0.032 | **0.527**±0.038 |
| **Global** | 0.452±0.022 | **0.540**±0.031 | **0.541**±0.019 |

Table 6.31: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global joint LSTME-CAE model for the day-of dataset.

The results in Tables 6.29 and 6.30 indicate that both branches outperform their respective unimodal models

in distinguishing between clean and anomalous states. Furthermore, the combined MSE scores in Table 6.31 also demonstrate an improvement in the joint LSTMAE-CAE model's performance compared to the unimodal models, however the imbalance is still evident.

Table 6.32 presents the ROC-AUC scores for the global unimodal models and branches of the joint LSTMAE-CAE model and its combined ROC-AUC scores.

| Norm. | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| **Per-patient** | 0.694±0.111 | **0.718**±0.040 | 0.452±0.125 | **0.462**±0.112 | 0.643±0.048 |
| **Global** | 0.691±0.046 | **0.704**±0.043 | 0.437±0.127 | **0.476**±0.154 | 0.617±0.091 |

Table 6.32: Comparison of ROC-AUC scores of the global unimodal models and branches of the joint LSTMAE-CAE model and its combined ROC-AUC scores for the day-of dataset.

The ROC-AUC scores in Table 6.32 also show that the audio and bio branches' performance is increased compared to their respective unimodal models, with the combined ROC-AUC score showing an improvement over the bio model only, as it was observed in the personalized setup.

Overall, the performance of the LSTMAE-CAE model on the day-of dataset is comparable and slightly better than the CAE-CAE model and also it demonstrates the same behavior in terms of the imbalance between the audio and bio modalities. However, the results suggest that the joint LSTMAE-CAE model has the potential to outperform the unimodal models on a larger dataset and improve the overall relapse detection performance.

### 6.3.2 Temporal Windows Around Interview

### Personalized Experiments

Tables 6.33-6.35 present the median MSE anomaly scores of the audio and bio unimodal models and branches of the joint CAE-CAE model and its combined MSE scores, respectively, for each temporal window dataset. Table 6.36 and Figure 6.3.1 present the mean ROC-AUC scores for the personalized experiments across all datasets, including the day-of dataset in the Figure.

| Dataset | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.491±0.156 | **0.593**±0.206 | 0.475±0.178 | 0.452±0.159 | **0.584**±0.168 | **0.566**±0.144 |
| **5-day** | 0.440±0.117 | **0.587**±0.132 | **0.481**±0.126 | 0.443±0.106 | **0.574**±0.109 | **0.499**±0.108 |
| **7-day** | 0.482±0.145 | **0.552**±0.112 | 0.457±0.193 | 0.459±0.115 | **0.541**±0.097 | **0.550**±0.173 |

Table 6.33: Comparison of median MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized unimodal audio model and audio branch of the joint LSTMAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.679±0.108 | 0.643±0.451 | **0.684**±0.273 | 0.796±0.101 | **0.885**±0.956 | **0.832**±0.267 |
| **5-day** | 0.541±0.073 | **0.600**±2.849 | **0.648**±0.255 | 0.571±0.237 | **0.779**±2.707 | **0.771**±0.395 |
| **7-day** | 0.615±0.196 | **0.832**±0.166 | **0.624**±0.355 | 0.639±0.237 | **0.845**±0.223 | **0.699**±0.380 |

Table 6.34: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized unimodal bio model and bio branch of the joint LSTMAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.516±0.125 | **0.638**±0.620 | **0.618**±0.580 |
| **5-day** | 0.529±0.086 | **0.571**±0.552 | **0.557**±0.439 |
| **7-day** | 0.520±0.137 | **0.634**±0.111 | **0.677**±0.287 |

Table 6.35: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global joint LSTMAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| **3-day** | 0.642±0.087 | **0.669**±0.061 | 0.547±0.198 | **0.583**±0.173 | **0.662**±0.169 |
| **5-day** | 0.633±0.082 | **0.650**±0.094 | 0.603±0.126 | **0.630**±0.142 | **0.646**±0.165 |
| **7-day** | 0.624±0.143 | **0.650**±0.143 | 0.557±0.136 | **0.597**±0.151 | **0.627**±0.163 |

Table 6.36: Comparison of ROC-AUC scores of the personalized unimodal models and branches of the joint LSTMAE-CAE model and its combined ROC-AUC scores for the 3, 5, and 7-day datasets.

The results in Tables 6.33-6.35 show an advantage of the joint LSTMAE-CAE model over its unimodal counterparts, yielding higher MSE anomaly scores for pre-relapse and relapse states across all datasets. ROC-AUC scores in Table 6.36 and Figure 6.3.1 further confirm the observed trend, with the combined ROC-AUC scores exceeding those of the unimodal models, which suggests that the LSTMAE-CAE model is also effective in detecting pre-relapse and relapse states for individual patients.

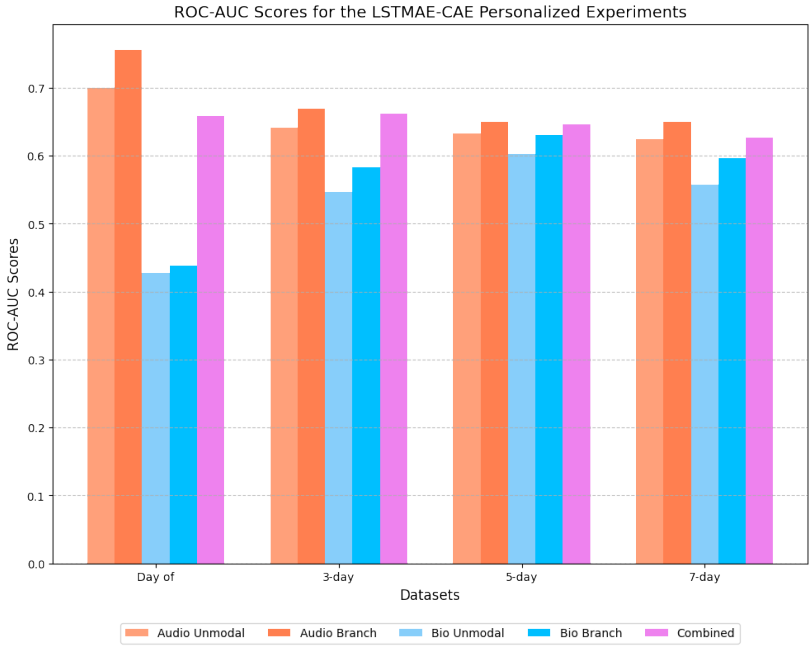



Figure 6.3.1: Mean ROC-AUC scores for the personalized experiments for all datasets.

Overall, the results are comparable to the CAE-CAE model, although the CAE-CAE shows a slightly better performance in terms of the ROC-AUC scores, especially in the combined results.

## Global Experiments

### Per-Patient Normalization

Tables 6.37-6.39 present the MSE anomaly scores of the global unimodal audio and bio models and branches of the joint CAE-CAE model and its combined MSE scores, respectively, for each temporal window dataset with per-patient normalization. Table 6.40 and Figure 6.3.2 present the ROC-AUC scores.

| Dataset | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.298±0.054 | **0.397**±0.022 | **0.321**±0.014 | 0.334±0.046 | **0.425**±0.029 | **0.363**±0.013 |
| **5-day** | 0.293±0.031 | **0.381**±0.021 | **0.303**±0.009 | 0.323±0.035 | **0.426**±0.015 | **0.336**±0.012 |
| **7-day** | 0.272±0.020 | **0.352**±0.016 | **0.297**±0.003 | 0.321±0.021 | **0.378**±0.004 | **0.335**±0.008 |

Table 6.37: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global (per-patient normalization) unimodal audio model and audio branch of the joint LSTMAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.704±0.063 | 0.640±0.045 | 0.653±0.058 | 0.679±0.033 | **0.706**±0.037 | **0.695**±0.037 |
| **5-day** | 0.558±0.062 | 0.531±0.027 | **0.620**±0.030 | 0.563±0.080 | **0.603**±0.050 | **0.632**±0.080 |
| **7-day** | 0.549±0.063 | **0.630**±0.033 | **0.639**±0.057 | 0.511±0.015 | **0.610**±0.018 | **0.577**±0.016 |

Table 6.38: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the global (per-patient normalization) unimodal bio CAE model and the audio branch of the joint LSTMAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.414±0.040 | **0.491**±0.011 | **0.487**±0.015 |
| **5-day** | 0.388±0.041 | **0.464**±0.020 | **0.440**±0.022 |
| **7-day** | 0.376±0.033 | **0.428**±0.006 | **0.413**±0.010 |

Table 6.39: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global (per-patient normalization) joint LSTMAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| **3-day** | 0.639±0.090 | **0.643**±0.091 | 0.490±0.049 | **0.540**±0.038 | **0.641**±0.058 |
| **5-day** | 0.622±0.052 | **0.637**±0.065 | 0.519±0.077 | **0.562**±0.080 | 0.613±0.057 |
| **7-day** | 0.632±0.032 | 0.612±0.042 | 0.553±0.029 | **0.576**±0.033 | 0.603±0.043 |

Table 6.40: Comparison of ROC-AUC scores of the global (per-patient normalization) unimodal models and branches of the joint LSTMAE-CAE model and its combined ROC-AUC scores for the 3, 5, and 7-day datasets.

In the global experiments using per-patient normalization, the joint model achieved moderate improvements over unimodal models. The MSE scores reflected a slight enhancement in identifying relapse states, while the ROC-AUC scores provided more noticeable improvements, particularly within the bio branch. The combined ROC-AUC scores showed mixed results, outperforming both unimodal models only for the 3-day dataset and for the others demonstrating only a marginal advantage. These findings suggest that per-patient normalization preserved some of the individualized benefits seen in the personalized experiments, enabling a better distinction of relapse states while still being constrained by the challenges of generalization.

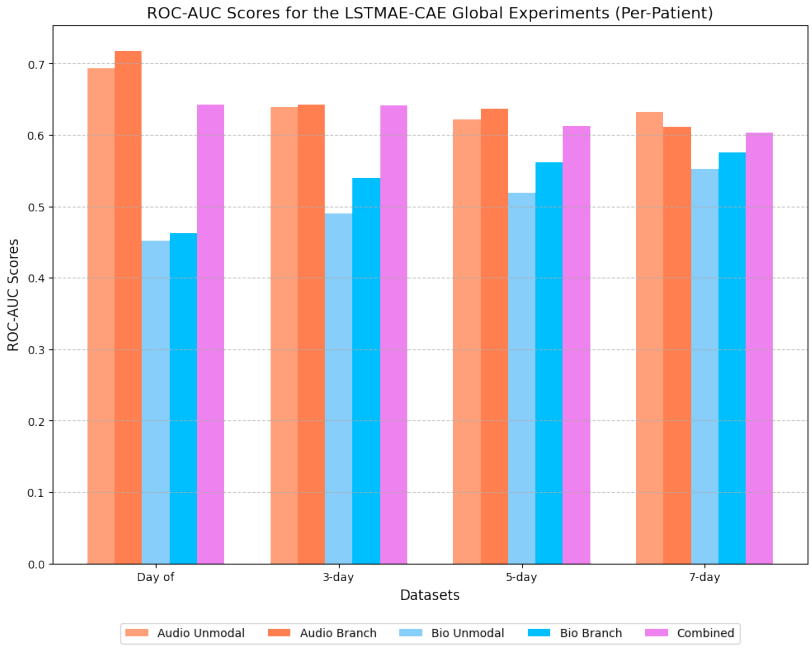



Figure 6.3.2: Mean ROC-AUC scores for the global experiments (per-patient normalization) for all datasets.

### Global Normalization

Tables 6.41-6.43 present the MSE anomaly scores of the global unimodal audio and bio models and branches of the joint CAE-CAE model and its combined MSE scores, respectively, for each temporal window dataset with per-patient normalization. Table 6.44 and Figure 6.3.3 present the ROC-AUC scores.

| Dataset | Audio Unimodal | | | Audio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.293±0.017 | **0.392**±0.025 | **0.317**±0.011 | 0.344±0.050 | **0.424**±0.022 | **0.361**±0.015 |
| **5-day** | 0.270±0.011 | **0.386**±0.011 | **0.300**±0.003 | 0.339±0.036 | **0.425**±0.019 | **0.342**±0.008 |
| **7-day** | 0.271±0.010 | **0.353**±0.015 | 0.296±0.012 | 0.284±0.029 | **0.376**±0.007 | **0.330**±0.008 |

Table 6.41: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global (global normalization) unimodal audio model and audio branch of the joint LSTMAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | Bio Unimodal | | | Bio Branch | | |
|---|---|---|---|---|---|---|
| | MSE-C | MSE-P | MSE-R | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.640±0.036 | 0.598±0.022 | 0.626±0.028 | 0.665±0.014 | **0.695**±0.059 | **0.702**±0.051 |
| **5-day** | 0.555±0.043 | 0.537±0.028 | **0.601**±0.040 | 0.579±0.089 | **0.598**±0.043 | **0.680**±0.064 |
| **7-day** | 0.589±0.055 | **0.629**±0.032 | **0.663**±0.056 | 0.546±0.035 | **0.605**±0.047 | **0.595**±0.037 |

Table 6.42: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states for the global (global normalization) unimodal bio CAE model and the audio branch of the joint LSTMAE-CAE model for the 3-day dataset.

Under global normalization, the joint model's performance closely aligned with that of the unimodal models, offering only minimal gains in relapse detection. The MSE scores in Tables 6.41-6.43 reflected a slight increase in anomaly differentiation, but the overall improvements were less pronounced than in the personalized and per-patient normalization setups. The ROC-AUC scores in Table 6.44 and Figure 6.3.3 showed that, while the

| Datasets | Combined | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **3-day** | 0.410±0.051 | **0.469**±0.015 | **0.480**±0.032 |
| **5-day** | 0.397±0.037 | **0.458**±0.017 | **0.452**±0.015 |
| **7-day** | 0.384±0.035 | **0.430**±0.010 | 0.426±0.046 |

Table 6.43: Combined MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global (global normalization) joint LSTMAE-CAE model for the 3, 5, and 7-day datasets.

| Dataset | ROC-AUC | | | | |
|---|---|---|---|---|---|
| | Audio Unimodal | Audio Branch | Bio Unimodal | Bio Branch | Combined |
| **3-day** | 0.636±0.027 | **0.643**±0.113 | 0.497±0.044 | **0.530**±0.045 | **0.644**±0.073 |
| **5-day** | 0.626±0.028 | **0.638**±0.069 | 0.536±0.054 | **0.548**±0.066 | 0.614±0.060 |
| **7-day** | 0.617±0.022 | **0.629**±0.051 | 0.543±0.050 | **0.556**±0.037 | 0.612±0.046 |

Table 6.44: Comparison of ROC-AUC scores of the global (global normalization) unimodal models and branches of the joint LSTMAE-CAE model and its combined ROC-AUC scores for the 3, 5, and 7-day datasets.

bio and audio branches generally outperformed their unimodal counterparts, the combined score displayed only slight gains, especially for the 3-day dataset.

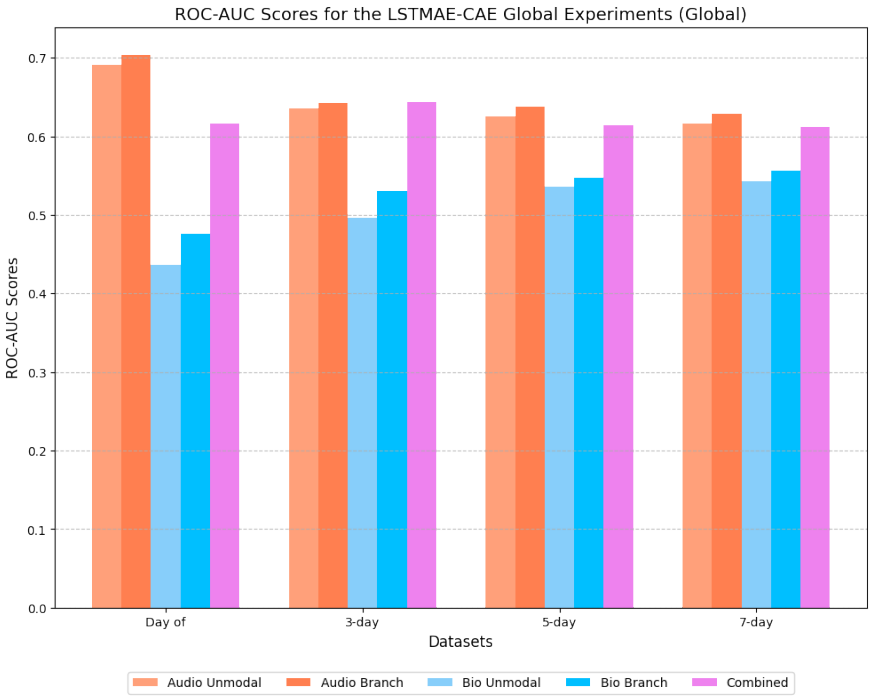



Figure 6.3.3: Mean ROC-AUC scores for the global experiments (global normalization) for all datasets.

Overall, the joint LSTMAE-CAE model demonstrated its strongest performance in personalized experiments, where it effectively captured individual patient patterns, leading to improved relapse detection compared to unimodal models. Per-patient normalization in global experiments provided a moderate advantage, preserving some of the benefits of personalization while enabling generalization across patients. However, global normalization resulted in the least pronounced improvements, highlighting the challenges of applying a this approach to anomaly detection in diverse patient populations. Despite these limitations, the multimodal integration of audio and biometric data consistently outperformed unimodal models in all setups, reinforcing the value of this approach for relapse prediction.

Additionally, in comparison to the CAE-CAE model, the performance of the LSTMAE-CAE model was comparable, with the CAE-CAE model showing a slight advantage in the personalized experiments. However, the LSTMAE-CAE model demonstrated a better performance in the global experiments, especially in the 3-day dataset, where the combined ROC-AUC score was higher than both unimodal models.

## 6.4 Evaluating Modality Contributions Through Branch Disabling

Ablation experiments were conducted to further evaluate the contributions of each modality to the performance of the joint model. By selectively disabling one branch at a time during inference, we aimed to asses the performance of the remaining branch and determine whether the joint model was effectively leveraging both modalities.

To disable one branch, we zeroed out its input during inference while keeping the other branch intact. This removes the contribution of the disabled branch on the model's output, allowing us to evaluate the performance of the remaining branch in isolation. The experiments were conducted on the 7-day dataset using the CAE-CAE joint model since it was the best performing model in the multimodal experiments.

We compare the MSE anomaly scores and ROC-AUC scores of the enabled branch, when the other branch is disabled, with its unimodal counterpart and the branch itself when both branches are enabled for both personalized and global experiments.

### Personalized Experiments

Tables 6.45 and 6.46 show the comparison of MSE anomaly scores and ROC-AUC scores for the personalized audio unimodal model, the audio branch of the CAE-CAE joint model with the bio branch disabled and enabled, respectively.

Similarily, the Tables 6.47 and 6.48 show the comparison of MSE anomaly scores and ROC-AUC scores for the personalized bio unimodal model, the bio branch of the CAE-CAE joint model with the audio branch disabled and enabled, respectively.

| **All Patients** | **Audio Unimodal** | | |
|---|---|---|---|
| | **MSE-C** | **MSE-P** | **MSE-R** |
| **Median** | 0.469±0.175 | **0.525**±0.106 | **0.503**±0.159 |

(a) Unimodal audio model.

| **All Patients** | **Audio Branch (Bio Disabled)** | | |
|---|---|---|---|
| | **MSE-C** | **MSE-P** | **MSE-R** |
| **Median** | 0.502±0.153 | **0.559**±0.119 | **0.539**±0.195 |

(b) Audio branch of the CAE-CAE joint model with the bio branch disabled.

| **All Patients** | **Audio Branch** | | |
|---|---|---|---|
| | **MSE-C** | **MSE-P** | **MSE-R** |
| **Median** | 0.527±0.151 | **0.615**±0.124 | **0.647**±0.188 |

(c) Audio branch of the CAE-CAE joint model with the bio branch enabled.

Table 6.45: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized audio unimodal model, the audio branch of the CAE-CAE joint model with the bio branch disabled and enabled respectively.

The MSE anomaly scores for the audio branch, as seen in Table 6.45, show a progressive improvement from the unimodal audio model to the audio branch with the bio branch disabled, and finally to the fully enabled joint model. This suggests that the addition of biometric data improves the performance of the audio branch, which is further supported by the ROC-AUC scores in Table 6.46.

| All Patients | ROC-AUC | | |
|---|---|---|---|
| | Audio Unimodal | Audio Branch (Bio Disabled) | Audio Branch |
| **Mean** | 0.598±0.064 | 0.620±0.108 | 0.643±0.079 |

Table 6.46: Comparison of ROC-AUC scores of the personalized audio unimodal model, the audio branch of the CAE-CAE joint model with the bio branch disabled and enabled respectively.

| All Patients | Bio Unimodal | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Median** | 0.615±0.196 | **0.832**±0.166 | **0.624**±0.355 |

(a) Unimodal bio model.

| All Patients | Bio Branch (Audio Disabled) | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Median** | 0.750±0.252 | **1.079**±0.208 | **0.787**±0.364 |

(b) Bio branch of the CAE-CAE joint model with the audio branch disabled.

| All Patients | Bio Branch | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Median** | 0.675±0.259 | **1.056**±0.259 | **0.706**±0.390 |

(c) Bio branch of the CAE-CAE joint model with the audio branch enabled.

Table 6.47: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the personalized bio unimodal model, the bio branch of the CAE-CAE joint model with the audio branch disabled and enabled respectively.

| All Patients | ROC-AUC | | |
|---|---|---|---|
| | Bio Unimodal | Bio Branch (Audio Disabled) | Bio Branch |
| **Mean** | 0.557±0.136 | 0.602±0.108 | 0.629±0.120 |

Table 6.48: Comparison of ROC-AUC scores of the personalized bio unimodal model, the bio branch of the CAE-CAE joint model with the audio branch disabled and enabled respectively.

The results in Table 6.47 and Table 6.48 show a similar trend to the audio modality, with the bio branch of the joint model, while the audio branch is disabled, outperforming the unimodal bio model, but still underperforming the fully enabled joint model.

Overall, the results from both the audio and bio branch disabling experiments confirm that the joint model is effectively leveraging both modalities to improve relapse detection over their unimodal counterparts.

### Global Experiments

Tables 6.49-6.52 show the comparison of MSE anomaly scores and ROC-AUC scores for the global unimodal models, the branch of the CAE-CAE joint model with the other branch disabled and enabled, respectively.

For both the audio and bio branches, the MSE anomaly scores and ROC-AUC scores follow a similar pattern to the personalized experiments, albeit with less pronounced improvements. This is consistent with the findings from the experiments on the 7-day dataset, where the global models exhibited more moderate results compared to the personalized models. Although the improvements are smaller, the results from the global branch disabling experiments still indicate that the joint model effectively integrates both modalities to enhance relapse detection beyond what unimodal models achieve.

| Norm. | Audio Unimodal | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.233±0.017 | **0.274**±0.015 | **0.245**±0.008 |
| **Global** | 0.233±0.019 | **0.277**±0.010 | **0.250**±0.006 |

(a) Unimodal audio model.

| Norm. | Audio Branch (Bio Disabled) | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Per-patient** | 0.232±0.021 | **0.287**±0.013 | **0.253**±0.011 |
| **Global** | 0.224±0.020 | **0.288**±0.009 | **0.251**±0.005 |

(b) Audio branch of the CAE-CAE joint model with the bio branch disabled.

| Norm. | Audio Branch | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Per-Patient** | 0.254±0.021 | **0.319**±0.011 | **0.283**±0.009 |
| **Global** | 0.257±0.021 | **0.321**±0.012 | **0.276**±0.011 |

(c) Audio branch of the CAE-CAE joint model with the bio branch enabled.

Table 6.49: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the globalaudio unimodal model, the audio branch of the CAE-CAE joint model with the bio branch disabled and enabled respectively.

| Norm. | ROC-AUC | | |
|---|---|---|---|
| | Audio Unimodal | Audio Branch (Bio Disabled) | Audio Branch |
| **Per-Patient** | 0.603±0.059 | 0.607±0.093 | 0.614±0.048 |
| **Global** | 0.600±0.060 | 0.602±0.056 | 0.607±0.053 |

Table 6.50: Comparison of ROC-AUC scores of the global audio unimodal model, the audio branch of the CAE-CAE joint model with the bio branch disabled and enabled respectively.

| Norm. | Bio Unimodal | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Per-Patient** | 0.549±0.063 | **0.630**±0.033 | **0.639**±0.057 |
| **Global** | 0.589±0.055 | **0.629**±0.032 | **0.663**±0.056 |

(a) Unimodal bio model.

| Norm. | Bio Branch (Audio Disabled) | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Per-Patient** | 0.635±0.066 | **0.684**±0.044 | **0.655**±0.073 |
| **Global** | 0.620±0.055 | **0.675**±0.025 | **0.653**±0.032 |

(b) Bio branch of the CAE-CAE joint model with the audio branch disabled.

| Norm. | Bio Branch | | |
|---|---|---|---|
| | MSE-C | MSE-P | MSE-R |
| **Per-Patient** | 0.578±0.023 | **0.658**±0.041 | **0.597**±0.024 |
| **Global** | 0.559±0.077 | **0.607**±0.066 | **0.589**±0.076 |

(c) Bio branch of the CAE-CAE joint model with the audio branch enabled.

Table 6.51: Comparison of MSE anomaly scores for clean (C), pre-relapse (P), and relapse (R) states of the global bio unimodal model, the bio branch of the CAE-CAE joint model with the audio branch disabled and enabled respectively.

| Norm. | ROC-AUC | | |
|---|---|---|---|
| | Bio Unimodal | Bio Branch (Audio Disabled) | Bio Branch |
| **Per-Patient** | 0.553±0.029 | 0.555±0.040 | 0.555±0.032 |
| **Global** | 0.543±0.050 | 0.550±0.057 | 0.572±0.039 |

Table 6.52: Comparison of ROC-AUC scores of the global bio unimodal model, the bio branch of the CAE-CAE joint model with the audio branch disabled and enabled respectively.

## 6.5 Discussion

Incorporating additional data around the interview dates led to a consistent and notable improvement in detecting pre-relapse and relapse states for both the personalized CAE-CAE and LSTMAE-CAE models across all datasets. This improvement was reflected in higher ROC-AUC scores for each branch as well as the combined scores, highlighting a clear advantage over the unimodal baselines.

A key observation across all experiments was the consistent superior performance of personalized setups compared to global setups. Personalized models were particularly beneficial and well suited for individual patients, as they could more effectively capture and leverage patient-specific patterns from both audio and biometric modalities. This advantage translated into a more significant improvement in detecting pre-relapse and relapse states, underscoring the potential of personalized multimodal models for early relapse detection in mental health.

In contrast, the global setups produced more variable results. While the 3-day and 5-day datasets showed potential—especially under per-patient normalization—the 7-day dataset yielded more mixed outcomes, likely due to increased variability introduced by additional patients. Interestingly, the global models demonstrated a more significant improvement in the bio branch compared to the audio branch, suggesting that the integration of audio data was particularly beneficial for enhancing the bio branch's performance.

Comparing the CAE-CAE and LSTMAE-CAE models, their performance was relatively similar across the datasets, with the CAE-CAE model slightly outperforming the LSTMAE-CAE model in the personalized setup, while the LSTMAE-CAE model showed a slight advantage in the global setup. Therefore, the choice between the two models may depend on the specific requirements of the application, with the CAE-CAE model potentially being more suitable for personalized setups, while the LSTMAE-CAE model may be more appropriate for global setups.

Finally, the branch disabling experiments provided further confirmation that the joint models effectively utilized both modalities to enhance relapse detection, though the degree of improvement varied between the personalized and global setups. In the personalized experiments, both the audio and bio branches showed clear performance gains when the complementary modality was enabled, demonstrating that the joint model successfully leveraged both modalities. The global experiments followed a similar trend, but with more subtle improvements, consistent with earlier observations of the global models.

# Chapter 7

# Conclusion

In this thesis, we conducted an in-depth analysis of audio signals and biometric markers to enhance relapse prediction in individuals with bipolar disorder and schizophrenia spectrum disorders. Building upon the e-Prevention project, we first expanded the existing audio database by incorporating additional patient data, including more relapse interviews. This expanded dataset allowed us to re-evaluate the CAE and CVAE models developed during the e-Prevention project, demonstrating that increased data volume improved model performance, particularly for the CAE, which showed significant improvements in both personalized and global setups. The CVAE remained effective in personalized experiments but exhibited more variability in global setups, possibly due to the added complexity of additional data. To further enhance anomaly detection, we introduced temporal modeling using LSTM-based autoencoders (LSTMAE) and variational autoencoders (LSTMVAE). The LSTMAE effectively captured sequential dependencies, outperforming the CAE across all experimental settings. The LSTMVAE produced results comparable to the CVAE, although the latter proved to be the most reliable variational autoencoder approach, while the LSTMVAE showed potential for further exploration. The most significant contribution of this research was the integration of multimodal data, combining speech with biometric signals (accelerometer, gyroscope, and heart rate) to improve relapse detection. We developed joint autoencoder frameworks incorporating CAE and LSTMAE models for the audio branch and a CAE model for the biometric branch, training and evaluating them on datasets that included biometric data from the exact interview day and extended temporal windows around the interview date (3-day, 5-day, and 7-day). Both multimodal models consistently outperformed unimodal models in detecting pre-relapse and relapse states, with personalized models proving particularly effective at capturing patient-specific patterns. Comparing CAE-CAE and LSTMAE-CAE models, we found their performance to be similar, with CAE-CAE excelling in personalized setups and LSTMAE-CAE performing slightly better in global setups, suggesting that model choice should be tailored to the application's needs. Lastly, branch disabling experiments confirmed that the joint models effectively leveraged both modalities.

The contributions of our work can be summarized as follows:

- **Dataset Expansion and Evaluation of Autoencoder Architectures:** We expanded the existing e-Prevention audio database by incorporating additional patient data and relapse interviews, which provided a more diverse dataset for evaluating the performance of the already developed Convolutional Autoencoder (CAE) and Convolutional Variational Autoencoder (CVAE) architectures. Our findings confirmed that a larger dataset enhances the detection of relapse states, across both personalized and global models.
- **Development of LSTM-based Autoencoders:** We explored the potential of sequential models to capture temporal dependencies in speech data by developing Long Short-Term Memory (LSTM)-based autoencoder architectures, namely LSTMAE and LSTMVAE models. Our experiments demonstrated that the LSTMAE model was effective in capturing temporal patterns in audio data, leading to improved performance in relapse detection compared to the CAE model.
- **Multimodal Fusion nd Joint Autoencoder Frameworks:** We explored the benefits of multimodal fusion by combining audio and biometric data. By designing joint autoencoder frameworks that inte-

grate the learned representations from both modalities into a common latent space, we demonstrated that multimodal approaches can significantly improve the accuracy of relapse detection. Our findings suggest that the fusion of diverse data sources offers complementary information that strengthens predictive performance.

- **Personalized and Global Setups:** Overall, our experiments indicate that personalized models—tailored to the specific characteristics of individual patients—consistently outperform global models, which are trained on the entire dataset.

In summary, this research demonstrates that integrating audio and biometric data through advanced autoencoder architectures can enhance the early detection of relapse in bipolar disorder and schizophrenia spectrum disorders. By leveraging multimodal data, we contribute to the efforts aimed at more timely clinical interventions and personalized care for patients with mental health conditions.

## Future Work

The work presented in this thesis can be extended in several directions to further improve the performance of relapse detection models and multimodal fusion approaches. Future research could focus on the following:

- **Dataset Expansion:** Further expanding the dataset to include more patients additional relapse cases could enhance even more the generalization of the models.
- **Latent Space Regularization:** For the variational autoencoder models, who generally performed less effectively than the traditional autoencoder models, additional regularization techniques could be explored, such as adaptive regularization, patient-specific weights or fine-tuning for specific patients for whom the performance did not improve.
- **Temporal Modeling:** Investigating more advanced temporal modeling techniques, such as attention mechanisms, could further improve the performance of the LSTM-based autoencoders.
- **Multimodal Data Alignment and Augmentation:** Exploring advanced alignment techniques, and data augmentation strategies could enhance the robustness of the multimodal fusion models by reducing the imbalance between the audio and biometric data.

# Appendix A

# Bibliography


[Aba+16] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from. 2016.

[Abu+20] Aburakhia, S., Tayeh, T., Myers, R., and Shami, A. "A Transfer Learning Framework for Anomaly Detection Using Model of Normality". In: *IEEE* (2020).

[Adl+20] Adler, D. A., Ben-Zeev, D., Tseng, V. W.-S., Kane, J. M., Brian, R., Campbell, A. T., Hauser, M., Scherer, E. A., and Choudhury, T. "Predicting Early Warning Signs of Psychotic Relapse From Passive Sensing Data: An Approach Using Encoder-Decoder Neural Networks". In: *JMIR mHealth and uHealth* 8.8 (2020), e19962.

[Alm+24] Almeida, F. F. de, Aires, K. R. T., Soares, A. C. B., Sousa Britto Neto, L. de, and Melo Souza Veras, R. de. "Multimodal Fusion for Depression Detection Assisted by Stacking Deep Neural Networks". In: *IEEE Journal* (2024).

[Ame13] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th. American Psychiatric Publishing, 2013.

[AMC17] Aung, M. H., Matthews, M., and Choudhury, T. "Sensing Behavioral Symptoms of Mental Health and Delivering Personalized Interventions Using Mobile Technologies". In: *Depression and Anxiety* 34 (2017), pp. 603–609. DOI: 10.1002/da.22646.

[BKH16] Ba, J. L., Kiros, J. R., and Hinton, G. E. "Layer Normalization". In: (2016). arXiv: 1607.06450.

[Bar+18] Barnett, I., Torous, J., Staples, P., Sandoval, L., Keshavan, M., and Onnela, J. P. "Relapse Prediction in Schizophrenia through Digital Phenotyping: A Pilot Study". In: *Neuropsychopharmacology* (2018).

[BDI20] Bayram, B., Duman, T. B., and Ince, G. "Real time detection of acoustic anomalies in industrial processes using sequential autoencoders". In: *Expert Systems* e12564 (2020). DOI: 10.1111/exsy.12564.

[BSB96] Beck, A. T., Steer, R. A., and Brown, G. K. *Beck Depression Inventory–II*. 1996.

[Ber+24] Berahmand, K., Daneshfar, F., Salehi, E. S., and Safabakhsh, R. "Autoencoders and their applications in machine learning: a survey". In: *Artificial Intelligence Review* 57 (2024), p. 28. DOI: 10.1007/s10462-023-10662-6.

[Bis94] Bishop, C. M. "Neural networks and their applications". In: *Review of Scientific Instruments* 65.6 (1994), pp. 1803–1832.

[BI21] Biswas, A. and Islam, M. "An Efficient CNN Model for Automated Digital Handwritten Digit Classification". In: *Journal of Information Systems Engineering and Business Intelligence* 7 (Apr. 2021), pp. 42–55.

[BMV12] Bowersox, N. W., McCarthy, D. E., and Valenstein, M. "Predictors of Relapse in the Year After Hospital Discharge Among State Hospital Patients With Schizophrenia". In: *Psychiatric Services* 63.1 (2012), pp. 89–90. DOI: 10.1176/appi.ps.201100084.

[BPK01] Brennan, M., Palaniswami, M., and Kamen, P. "Do existing measures of Poincaré plot geometry reflect nonlinear features of heart rate variability?" In: *IEEE Transactions on Biomedical Engineering* 48.11 (2001), pp. 1342–1347.

[Bre+00] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. "LOF: Identifying Density-Based Local Outliers". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Dallas, TX, USA: ACM, 2000, pp. 93–104.

[Cap+19] Cappon, G., Vettoretti, M., Sparacino, G., and Facchinetti, A. "Continuous Glucose Monitoring Sensors for Diabetes Management: A Review of Technologies and Applications". In: *Diabetes & Metabolism Journal* 43 (July 2019), pp. 383–397.

[CCM24] Carbonera, M., Ciavotta, M., and Messina, E. "Variational Autoencoders and Generative Adversarial Networks for Multivariate Scenario Generation". In: *Data Science for Transportation* 6 (2024), p. 23.

[Cha+17] Chang, Y.-H. S., Liao, Y.-f., Wang, S.-M., Wang, J.-H., Wang, S.-y., Chen, J.-w., and Chen, Y.-d. "Development of a Large-Scale Mandarin Radio Speech Corpus". In: *Proceedings of the IEEE International Conference on Consumer Electronics-Taiwan* (June 2017).

[CG16] Chen, T. and Guestrin, C. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. San Francisco, CA, USA: Association for Computing Machinery, 2016, pp. 785–794.

[Cho+15] Chollet, F. et al. *Keras: Deep Learning for Humans*. Software available from. 2015.

[Cho+09] Chouchou, F., Pichot, V., Garet, M., Barthelemy, J.-C., and Roche, F. "Dominance in cardiac parasympathetic activity during real recreational SCUBA diving". In: *European journal of applied physiology* 106 (Mar. 2009), pp. 345–52.

[Coe+22] Coelho, G., Matos, L. M., Pereira, P. J., Ferreira, A., Pilastri, A., and Cortez, P. "Deep autoencoders for acoustic anomaly detection: experiments with working machine and in-vehicle audio". In: *Neural Computing and Applications* 34 (2022), pp. 19485–19499. DOI: 10.1007/s00521-022-07375-2.

[DG21] Deary, M. and Griffiths, S. "A novel approach to the development of 1-hour threshold concentrations for exposure to particulate matter during episodic air pollution events". In: *Journal of Hazardous Materials* 418 (June 2021), p. 126334.

[DAS22] Dese, K., Ayana, G., and Simegn, G. L. "Low Cost, Non-Invasive, and Continuous Vital Signs Monitoring Device for Pregnant Women in Low Resource Settings (Lvital Device)". In: *HardwareX* 11 (Feb. 2022), e00276.

[Dev+19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2019).

[Fau+16] Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., Christensen, E. M., Winther, O., Bardram, J. E., and Kessing, L. V. "Voice Analysis as an Objective State Marker in Bipolar Disorder". In: *Translational Psychiatry* 6 (July 2016), e856.

[Fay16] Fayek, H. *Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs), and What's In-Between*. 2016.

[Fil+20] Filntisis, P. P., Zlatintsi, A., Efthymiou, N., Kalisperakis, E., Karantinos, T., Lazaridi, M., Smyrnis, N., and Maragos, P. "Identifying Differences in Physical Activity and Autonomic Function Patterns Between Psychotic Patients and Controls Over a Long Period of Continuous Monitoring Using Wearable Sensors". In: *Proceedings of the International Conference on Digital Phenotyping and Mental Health*. IEEE, 2020.

[Gar+21] Garoufis, C., Zlatintsi, A., Filntisis, P. P., Efthymiou, N., Kalisperakis, E., Garyfalli, V., Karantinos, T., Mantonakis, L., Smyrnis, N., and Maragos, P. "An Unsupervised Learning Approach for Detecting Relapses from Spontaneous Speech in Patients with Psychosis". In: *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2021.

[Gid+19] Gideon, J., Matton, K., Anderau, S., McInnis, M. G., and Mower Provost, E. "When to Intervene: Detecting Abnormal Mood Using Everyday Smartphone Conversations". In: *IEEE Transactions on Affective Computing (Preprint)* (2019).

[GMP19] Gideon, J., McInnis, M., and Provost, E. M. "Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDOG)". In: *IEEE Transactions on Affective Computing* (2019).

[Gou+21] Gouverneur, P., Li, F., Adamczyk, W., Szikszay, T., Luedtke, K., and Grzegorzek, M. "Comparison of Feature Extraction Methods for Physiological Signals for Heat-Based Pain Recognition". In: *Sensors* 21 (July 2021), p. 4838. DOI: 10.3390/s21144838.

[Gra+16] Grande, I., Berk, M., Birmaher, B., and Vieta, E. "Bipolar Disorder". In: *The Lancet* 387.10027 (Apr. 2016), pp. 1561–1572.

[Guy76] Guy, W. *ECDEU Assessment Manual for Psychopharmacology*. Rockville, MD, USA: US Department of Health, Education, and Welfare, Public Health Service, 1976.

[Ham76] Hamilton, M. "Hamilton Depression Scale". In: *ECDEU Assessment Manual for Psychopharmacology*. Revised Edition. National Institute of Mental Health, 1976, pp. 179–192.

[HGD17] Hatami, N., Gavet, Y., and Debayle, J. "Classification of Time-Series Images Using Deep Convolutional Neural Networks". In: *arXiv preprint* 1710.00886v2 (2017).

[Hen+10] Henry, B. L., Minassian, A., Paulus, M. P., Geyer, M. A., and Perry, W. "Heart Rate Variability in Bipolar Mania and Schizophrenia". In: *Journal of Psychiatric Research* 44 (Feb. 2010), pp. 168–176.

[Het+23] Hett, D., Morales-Muñoz, I., Durdurak, B. B., Carlish, M., and Marwaha, S. "Rates and associations of relapse over 5 years of 2649 people with bipolar disorder: a retrospective UK cohort study". In: *International Journal of Bipolar Disorders* 11.1 (2023), p. 23. DOI: 10.1186/s40345-023-00302-x.

[Hig+17] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. "beta–VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: (2017).

[Hig88] Higuchi, T. "Approach to an irregular time series on the basis of the fractal theory". In: *Physica D: Nonlinear Phenomena* 31.2 (1988), pp. 277–283.

[HS97] Hochreiter, S. and Schmidhuber, J. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[HWS18] Huang, K.-Y., Wu, C.-H., and Su, M.-H. "Attention-based Convolutional Neural Network and Long Short-term Memory for Short-term Detection of Mood Disorders Based on Elicited Speech Responses". In: *Pattern Recognition* (2018).

[Ikä+24] Ikäheimonen, A., Luong, N., Baryshnikov, I., Darst, R., Heikkilä, R., Holmen, J., Martikkala, A., Riihimäki, K., Saleva, O., Isometsä, E., and Aledavood, T. "Predicting and Monitoring Symptoms in Diagnosed Depression Using Mobile Phone Data: An Observational Study". In: *medRxiv Preprint* (2024).

[İşb+23] İşbitirici, A., Falcone, P., Giarre, L., and Xu, W. "LSTM-based Virtual Load Sensor for Heavy-Duty Vehicles". In: (Nov. 2023).

[JMM18] Johansson, D., Malmgren, K., and Murphy, M. A. "Wearable Sensors for Clinical Applications in Epilepsy, Parkinson's Disease, and Stroke: A Mixed–Methods Systematic Review". In: *Journal of Neurology* 265 (Feb. 2018), pp. 1740–1752.

[JA03] Judd, L. L. and Akiskal, H. S. "The Prevalence and Disability of Bipolar Spectrum Disorders in the US Population: Re-analysis of the ECA Database Taking into Account Subthreshold Cases". In: *Journal of Affective Disorders* 73.1-2 (2003), pp. 123–131.

[KFO87a] Kay, S. R., Fiszbein, A., and Opler, L. A. "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia". In: *Schizophrenia Bulletin* 13 (June 1987), pp. 261–276.

[KFO87b] Kay, S. R., Fiszbein, A., and Opler, L. A. "The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia". In: *Schizophrenia Bulletin* 13.2 (1987), pp. 261–276.

[Kho+18] Khorram, S., Jaiswal, M., Gideon, J., McInnis, M., and Provost, E. M. "The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-The-Wild". In: *Proceedings of Interspeech*. ISCA. 2018, pp. 1903–1907.

[Koi+19] Koizumi, Y., Saito, S., Uematsu, H., Harada, N., and Imoto, K. "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection". In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 313–317. DOI: 10.1109/WASPAA.2019.8937164.

[LHL19] Lam, G., Huang, D., and Lin, W. "Context-Aware Deep Learning for Multi-Modal Depression Detection". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3946–3950.

[LNM14] Laursen, T. M., Nordentoft, M., and Mortensen, P. B. "Excess Early Mortality in Schizophrenia". In: *Annual Review of Clinical Psychology* 10 (Feb. 2014), pp. 425–448.

[Lee+24] Lee, Y., Park, C., Kim, N., Ahn, J., and Jeong, J. "LSTM-Autoencoder Based Anomaly Detection Using Vibration Data of Wind Turbines". In: *Sensors* 24 (2024), p. 2833.

[Li+23] Li, Y., Wang, Y., Yang, X., and Im, S.-K. "Speech emotion recognition based on Graph-LSTM neural network". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2023 (2023), p. 40. DOI: 10.1186/s13636-023-00303-9.

[LDG13] Lieberman, J. A., Dixon, L. B., and Goldman, H. H. "Early Detection and Intervention in Schizophrenia: A New Therapeutic Model". In: *JAMA* 310.7 (Aug. 2013), pp. 689–690.

[Lin+21] Lindemann, B., Müller, T., Vietz, H., Jazdi, N., and Weyrich, M. "A survey on long short-term memory networks for time series prediction". In: *Procedia CIRP* 99 (2021), pp. 650–655. DOI: 10.1016/j.procir.2021.03.088.

[Liu+13] Liu, L., Chen, X., Luo, D., Lu, Y., Peng, Y., and Du, J. "HSC: A Spectral Clustering Algorithm Combined with Hierarchical Method". In: *Neural Network World* 23.6 (2013), pp. 499–521.

[LBG20] Low, D. M., Bentley, K. H., and Ghosh, S. S. "Automated Assessment of Psychiatric Disorders Using Speech: A Systematic Review". In: *Laryngoscope Investigative Otolaryngology* 5 (Jan. 2020), pp. 96–116.

[Maa13] Maas, A. L. "Rectifier Nonlinearities Improve Neural Network Acoustic Models". In: (2013). URL:

[Mag+20] Maglogiannis, I., Zlatintsi, A., Menychtas, A., Papadimatos, D., Filntisis, P. P., Efthymiou, N., Retsinas, G., Tsanakas, P., and Maragos, P. "An Intelligent Cloud-Based Platform for Effective Monitoring of Patients with Psychotic Disorders". In: *IFIP Advances in Information and Communication Technology (AIAI)*. Vol. 584. Springer, Cham, 2020, pp. 293–307.

[Man82] Mandelbrot, B. B. *The Fractal Geometry of Nature*. New York: W. H. Freeman and Company, 1982. ISBN: 978-0716711865.

[Mar94] Maragos, P. "Fractal Signal Analysis Using Mathematical Morphology". In: *Advances in Electronics and Electron Physics* 88 (1994), pp. 199–246.

[Max+16] Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., and Morales, E. F. "Classification of Bipolar Disorder Episodes Based on Analysis of Voice and Motor Activity of Patients". In: *Pervasive and Mobile Computing* 31 (Feb. 2016), pp. 50–65.

[McF+15] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O. *librosa: Audio and Music Signal Processing in Python*. Proceedings of the 14th Python in Science Conference. 2015. DOI: 10.25080/Majora-7b98e3ed-003.

[MIB20] Mihajlovic, S., Ivetic, D., and Berković, I. "Applications of Convolutional Neural Networks". In: (Oct. 2020).

[Mob+22] Mobtahej, P., Zhang, X., Hamidi, M., and Zhang, J. "An LSTM-Autoencoder Architecture for Anomaly Detection Applied on Compressors Audio Data". In: *Computational and Mathematical Methods* 2022 (2022), pp. 1–22.

[MZS17] Mohr, D. C., Zhang, M., and Schueller, S. M. "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning". In: *Annual Review of Clinical Psychology* 13 (May 2017), pp. 23–47.

[Mou+21] Mouchabac, S., Conejero, I., Lakhlifi, C., Msellek, I., Malandain, L., Adrien, V., Ferreri, F., Millet, B., Bonnot, O., Bourla, A., and Maatoug, R. "Improving clinical decision-making in psychiatry: implementation of digital phenotyping could mitigate the influence of patient's and practitioner's individual cognitive biases". In: *Dialogues in Clinical Neuroscience* 23 (Jan. 2021), pp. 52–61.

[Nie+23] Nierenberg, A. A., Agustini, B., Köhler-Forsberg, O., Cusin, C., Katz, D., Sylvia, L. G., Peters, A., and Berk, M. "Diagnosis and Treatment of Bipolar Disorder: A Review". In: *JAMA* 330.14 (Oct. 2023), pp. 1370–1380.

[Ola15] Olah, C. "Understanding LSTMs". In: (2015).

[OR16] Onnela, J.-P. and Rauch, S. L. "Harnessing Smartphone–Based Digital Phenotyping to Enhance Behavioral and Mental Health". In: *Neuropsychopharmacology* 41 (Feb. 2016), pp. 1691–1700.

[Osm+15] Osmani, V., Gruenerbl, A., Bahle, G., Haring, C., Lukowicz, P., and Mayora, O. "Smartphones in Mental Health: Detecting Depressive and Manic Episodes". In: *IEEE Pervasive Computing* 14.3 (2015), pp. 10–13.

[OZ22] Othmani, A. and Zeghina, A. O. "A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept". In: *Healthcare Analytics* 2 (2022), p. 100090.

[OZM22] Othmani, A., Zeghina, A.-O., and Muzammel, M. "A Model of Normality Inspired Deep Learning Framework for Depression Relapse Prediction Using Audiovisual Data". In: *Computer Methods and Programs in Biomedicine* 226 (2022), p. 107132.

[Pan+18] Pan, Z., Gui, C., Zhang, J., Zhu, J., and Cui, D. "Detecting Manic State of Bipolar Disorder Based on Support Vector Machine and Gaussian Mixture Model Using Spontaneous Speech". In: *Psychiatry Investigation* 15.7 (July 2018), pp. 695–700.

[Pap+09] Papathanasiou, G., Georgoudis, G., Papandreou, M., Spyropoulos, P., Georgakopoulos, D., Kalfakakou, V., and Evangelou, A. "Reliability Measures of the Short International Physical Activity Questionnaire (IPAQ) in Greek Young Adults". In: *Hellenic Journal of Cardiology* 50 (2009), pp. 283–294.

[Ped+11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL:

[PG07] Piskorski, J. and Guzik, P. "Geometry of the Poincaré plot of RR intervals and its asymmetry in healthy adults". In: *Physiological Measurement* 28.3 (2007), pp. 287–300.

[Pov+11] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. "The Kaldi Speech Recognition Toolkit". In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Hawaii, USA, Dec. 2011.

[Pur+19] Purohit, H., Tanabe, R., Ichige, T., Endo, T., Nikaido, Y., Suefusa, K., and Kawaguchi, Y. "MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection". In: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. 2019, pp. 209–213.

[Raj23] Raj, V. "Exploring the Power of Neural Networks". In: (2023). URL:

[Ret+20] Retsinas, G., Filntisis, P. P., Efthymiou, N., Theodosis, E., Zlatintsi, A., and Maragos, P. "Person Identification Using Deep CNNs on Short-Term Signals from Wearable Sensors". In: *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, pp. 1–5.

[Rin+17] Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., and Pantic, M. "AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge". In: AVEC '17. Association for Computing Machinery, 2017, pp. 3–9.

[Sca82] Scargle, J. D. "Studies in Astronomical Time Series Analysis. II-Statistical Aspects of Spectral Analysis of Unevenly Spaced Data". In: *The Astrophysical Journal* 263 (1982), pp. 835–853.

[SG17a] Shaffer, F. and Ginsberg, J. P. "An Overview of Heart Rate Variability Metrics and Norms". In: *Frontiers in Public Health* 5 (Sept. 2017), p. 258.

[SG17b] Shaffer, F. and Ginsberg, J. P. "An Overview of Heart Rate Variability Metrics and Norms". In: *Frontiers in Public Health* 5 (Sept. 2017), p. 258.

[SZ14] Simonyan, K. and Zisserman, A. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[SA70] Simpson, G. and Angus, J. "A Rating Scale for Extrapyramidal Side Effects". In: *Acta Psychiatrica Scandinavica* 45 (1970), pp. 11–19. DOI: 10.1111/j.1600-0447.1970.tb02066.x.

[Sny+18] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. "X-Vectors: Robust DNN Embeddings for Speaker Recognition". In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Calgary, AB, Canada, Apr. 2018, pp. 5329–5333.

[Su+21] Su, H.-Y., Wu, C.-H., Liou, C.-R., Lin, E. C.-L., and Chen, P. S. "Assessment of Bipolar Disorder Using Heterogeneous Data of Smartphone-Based Digital Phenotyping". In: *Proceedings of the*

*IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4260–4264.

[TTR21] Toto, E., Tlachac, M., and Rundensteiner, E. "Audibert: A Deep Transfer Learning Multimodal Classification Framework for Depression Screening". In: *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM) Applied Research Track*. Association for Computing Machinery, 2021, pp. 4145–4154.

[Üst+10] Üstün, T. B., Chatterji, S., Kostanjsek, N., Rehm, J., Kennedy, C., Jordan, E.-J., Saxena, S., Korff, M. von, and Pull, C. "Developing the World Health Organization Disability Assessment Schedule 2.0". In: *Bulletin of the World Health Organization* 88 (2010), pp. 815–823.

[VK09] Van Os, J. and Kapur, S. "Schizophrenia". In: *The Lancet* 374.9690 (Aug. 2009), pp. 635–645.

[WT08] Waddell, L. and Taylor, M. "A New Self-Rating Scale for Detecting Atypical or Second-Generation Antipsychotic Side Effects". In: *Journal of Psychopharmacology* 22.3 (May 2008), pp. 238–243. DOI: 10.1177/0269881107087976.

[WM17] Walther, S. and Mittal, V. A. "Motor System Pathology in Psychosis". In: *Current Psychiatry Reports* 19 (Oct. 2017), p. 97.

[Wan+22] Wang, Y., Wang, Z., Li, C., Zhang, Y., and Wang, H. "Online Social Network Individual Depression Detection Using a Multitask Heterogeneous Modality Fusion Approach". In: *Information Sciences* 609 (2022), pp. 727–749.

[Wen18] Weng, L. "From Autoencoder to Beta-VAE". In: (2018).

[Wor22] World Health Organization. *Schizophrenia*. 2022.

[Wu+23] Wu, C.-H., Hsu, J.-H., Liou, C.-R., Su, H.-Y., Lin, E., and Chen, P. "Automatic Bipolar Disorder Assessment Using Machine Learning With Smartphone-Based Digital Phenotyping". In: *IEEE Access* PP (Jan. 2023), pp. 1–1.

[Yan+19] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: (Dec. 2019), pp. 5753–5763.

[YIS19] Yani, M., Irawan, S., and Setianingsih, C. "Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail". In: *Journal of Physics: Conference Series* 1201 (May 2019), p. 012052.

[You+78] Young, R. C., Biggs, J. T., Ziegler, V. E., and Meyer, D. A. "A Rating Scale for Mania: Reliability, Validity and Sensitivity". In: *British Journal of Psychiatry* 133 (Oct. 1978), pp. 429–435.

[Zho+24] Zhong, Y., Chen, Y., Su, X., Wang, M., Li, Q., Shao, Z., and Sun, L. "Global, regional and national burdens of bipolar disorders in adolescents and young adults: a trend analysis from 1990 to 2019". In: *General Psychiatry* 37 (2024). DOI: 10.1136/gpsych-2023-101255.

[Zla+22] Zlatintsi, A., Filntisis, P. P., Garoufis, C., Efthymiou, N., Maragos, P., Menychtas, A., Maglogiannis, I., Tsanakas, P., Sounapoglou, T., Kalisperakis, E., et al. "E–Prevention: Advanced Support System for Monitoring and Relapse Prevention in Patients with Psychotic Disorders Analyzing Long–Term Multimodal Data from Wearables and Video Captures". In: *Sensors* 22.19 (Oct. 2022), p. 7544.